

Proceedings of ISSI 2007

**11th International Conference of the International Society for
Scientometrics and Informetrics**

Edited by Daniel Torres-Salinas and Henk F. Moed

Proceedings of ISSI 2007

**11th International Conference of the International Society for
Scientometrics and Informetrics**

CSIC, Madrid, Spain
25-27 June 2007

VOLUME I

Edited by

» Daniel Torres-Salinas

- EC³ Research Group (Evaluación de la Ciencia y la Comunicación Científica)
Universidad de Granada, Granada, Spain
- Centro de Investigación Médica Aplicada (CIMA)
Universidad de Navarra, Pamplona, Spain

» Henk F. Moed

- Centre for Science & Technology Studies (CWTS)
Leiden University, Leiden, the Netherlands

Published for the International Society for Informetrics and Scientometrics (ISSI) by the
Centre for Scientific Information and Documentation (CINDOC) of the Spanish Research
Council (CSIC), Madrid, Spain

All rights reserved. No part of this book may be reproduced in any form without the written permission of the authors.

» ORGANISATION AND COMMITTEES

• **Conference Chairs:**

Prof. Isabel Gómez- Caridad
CINDOC-CSIC, Madrid, Spain

Dr. María Bordons
CINDOC-CSIC, Madrid, Spain

Isidro F. Agullo
CINDOC-CSIC, Madrid, Spain

• **Programme Chair:**

Dr. Henk F. Moed, Centre for Science & Technology Studies (CWTS), Leiden University, Leiden, The Netherlands

• **Poster Chair:**

Dr. Ed Noyons
Centre for Science & Technology Studies (CWTS), Leiden University, Leiden, The Netherlands

• **Satellite Workshops Chair:**

Prof. Peter Ingwersen
Royal School of Library & Information Science, Copenhagen, Denmark

• **Doctoral Forum Chairs:**

Dr. Rickard Danell, Umeå University, Umeå, Sweden.
Dr. Birger Larsen, Royal School of Library & Information Science, Copenhagen, Denmark.

• **Regional Programme Chairs:**

North America: Dr. Kate McCain, Drexel University, Philadelphia, USA.
Latin America: Dr. Jane Russell, National Autonomous University of Mexico (UNAM), Mexico City. Mexico.
Australia - Pacific: Dr. Linda Butler. The Australian National University. Canberra ACT , Australia.
China - Far East: Dr. Liang Liming. Henan Normal University, Xinxiang, China
Europe - Africa: Prof. Peter Ingwersen, Royal School of Library & Information Science, Copenhagen, Denmark
India - Middle East: Prof. Ravichandra Rao, Indian Statistical Institute, India.

• **Local Committee:**

Dr. M. Teresa Fernández, CINDOC-CSIC, Madrid, Spain
Dr. Evaristo Jiménez Contreras, Universidad de Granada, Granada, Spain
Dr. Fernanda Morillo, CINDOC-CSIC, Madrid, Spain
Dr. Luis Plaza, CINDOC-CSIC, Madrid, Spain
Dr. Adelaida Román, CINDOC-CSIC, Madrid, Spain
Dr. Rosa Sancho, CINDOC-CSIC, Madrid, Spain
Dr. Elías Sanz, Universidad Carlos III de Madrid, Madrid, Spain
Dr. M. Angeles Zulueta, Universidad de Alcalá, Alcalá de Henares, Spain

• **International Scientific Committee:**

Isidro Agullo
Subbiah Arunachalam
Judit Bar-Ilan
Elise Bassecoulard
Aparna Basu
Sujit Bhattacharya
Lennart Björneborn
Manfred Bonitz
Maria Bordons
Kevin Boyack
Tibor Braun
Quentin Burrell
Linda Butler
Chaomei Chen
Heting Chu
Blaise Cronin
Leo Egghe
Wolfgang Glänzel
Isabel Gomez

Sybille Hinze
Peter Ingwersen
Evaristo Jiménez-Contreras
Hildrun Kretschmer
Birger Larsen
Grit Laudel
Grant Lewison
Liang Liming
Terttu Luukonen
Cesar Macias-Chapula
Valentina Markusova
Katherine McCain
Martin Meyer
Henk Moed
Steven Morris
Ylle Must
Michael Nelson
Ed Noyons
Dennis Ocholla

Olle Persson
Ravichandra Rao
Ed Rinia
Ronald Rousseau
Ian Rowlands
Jane Russell
Elias Sanz
Henry Small
Alastair Smith
Mike Thelwall
Robert Tijssen
Thed Van Leeuwen
Liwen Vaughan
Peter Vinkler
Mei-Mei Wu
Wu Yishan
Michel Zitt
Alesia Zuccala

» LIST OF SPONSORS OF THE CONFERENCE¹

- International Society for Scientometrics and Informetrics (ISSI)
- Consejo Superior de Investigaciones Científicas (CSIC)
- Ministerio de Educación y Ciencia (MEC)
- Fundación Española para la Ciencia y la Tecnología (FECYT)
- Comunidad de Madrid
- Eugene Garfield Foundation
- Thomson Scientific
- Journal of Informetrics/Elsevier
- Scopus
- Ayuntamiento de Madrid



¹ Reflects the situation at April 1, 2007

The 11th International Conference of the International Society for Scientometrics and Informetrics is held on June 25-27, 2007 at the Serrano Central Campus of the Spanish Research Council (CSIC) in Madrid. This is a major event with participants from all over the world.

About 240 submissions were made to the conference. All submissions were thoroughly reviewed by members of the International Scientific Committee. Some 60 contributions were not yet sufficiently ripe for presentation and had to be rejected, and another 10 were retracted. Those that conform to the quality standards in the field were accepted for presentation at the conference. In the final program 2 keynotes are presented, 57 full papers, 35 research-in-progress papers, and 77 posters.

The Proceedings of the conference are published in two volumes:

Volume I (pp. 1-530):

- Keynotes
- Full papers and research-in-progress papers with first authors A-L

Volume II (pp. 542-975):

- Full papers and research-in-progress papers with first authors M-Z
- Poster papers

In each category of contributions, – keynotes, full papers and research-in-progress papers, and posters –, the papers are ordered alphabetically by first author. Each volume contains a complete Table of Contents covering the papers from both volumes. An author index is included at the beginning of Volume I and at the end of Volume II.

On behalf of the conference organizers, Isabel Gómez, María Bordons and Isidro Agullo, I want to thank all authors for their submissions, and the members of the Scientific Committee for their efforts in the review process. I am most grateful to my co-editor Daniel Torres-Salinas for his technical assistance in the preparation of the final manuscript, and to Evaristo Jiménez-Contreras for hosting me at the University of Granada during the time period these proceedings were prepared.

Henk F. Moed
Program Chair ISSI 2007

» INDEX

VOLUME I

KEYNOTES..... 19

- Eugene Garfield. *From The Science of Science to Scientometrics. Visualizing the History of Science with HistCite Software* 21
Stevan Harnad. *Open Access Scientometrics and the UK Research Assessment Exercise* 27

FULL PAPERS AND RESEARCH-IN-PROGRESS PAPERS..... 35

- Rafael Aleixandre-Benavent, Gregorio González-Alcaide, Alberto Miguel-Dasit, Carolina Navarro-Molina and Juan Carlos Valderrama-Zurián. *Full-Text Publications in Peer-Reviewed Journals Derived from Presentations at Two ISSI Conferences* 37
Eric Archambault and Vincent Larivière. *Origins of Measures of Journal Impact: Historical Contingencies and their Consequences on Current Use* 45
Judit Bar-Ilan and Bluma C. Peritz. *The Lifespan of "Informetrics" on the Web: an Eight Year Study (1998-2006)* 52
Franz Barjak and Simon Robinson. *International Collaboration, Mobility and Team Diversity in the Life Sciences: Impact on Research Performance* 63
Elise Bassecoulard, Alain Lelu, and Michel Zitt. *A Modular Sequence of Retrieval Procedures to Delineate a Scientific Field: from Vocabulary to Citations and Back* 74
Peter van den Besselaar, Gaston Heimeriks, Koen Frenken. *Variety in Web Spheres between Research Fields: Content and Function* 85
Sujit Bhattacharya. *Impact of Indian Patents: Assessment through Citation Analysis* 95
Omwoyo Bosire Onyancha and Dennis N. Ocholla. *Is HIV/AIDS in Africa Distinct? What Can we Learn from an Analysis of the Literature?* 100
Kevin W. Boyack, Katy Börner and Richard Klavan. *Mapping the Structure and Evolution of Chemistry Research* 112
Kevin W. Boyack. *Using Detailed Maps of Science to Identify Potential Collaborations* 124
Robert Braam. *Everything about Genes: some Results on the Dynamics of Genomics Research* 136
Lutz Bornmann and Hans-Dieter Daniel. *Functional Use of Frequently and Infrequently Cited Articles in Citing Publications. A Content Analysis of Citations to Articles with Low and High Citation Counts* 149
Jenny-Ann Brodin Danell and Rickard Danell. *Spiritualised Medicine? A Bibliometric Study of Complementary and Alternative Medicine* 154
Quentin L. Burrell. *Hirsch's h-index and Egghe's g-index* 162
Linda Butler and Kumara Henadeera. *Is there a Role for Novel Citation Measures for the Social Sciences and Humanities in a National Research Assessment Exercise?* 170
Clara Calero Medina and Ed C.M. Noyons. *Combining Mapping and Citation Network Analysis for a Better Understanding of the Scientific Development: The case of the Absorptive Capacity Field* 179
Mônica G. Campiteli, Pablo D. Batista and Alexandre S. Martinez. *A Research Productivity Index to Account for Different Scientific Disciplines* 184
Chaomei Chen, Il-Yeol Song and Weizhong Zhu. *Trends in Conceptual Modeling: Citation Analysis of the ER Conference Papers (1979-2005)* 189
Chen Li, Pan Yuntao, Ma Zheng, Su Cheng and Wu Yishan. *A Comparative Study between International and Domestic Interdisciplinary Journals and Specialty Journals: A Trial Analysis of Medical Journals, Philosophy Journals and Journals in Philosophy of Medicine* 201
Heting Chu and Thomas Krichel. *Downloads vs. Citations in Economics: Relationships, Contributing Factors and Beyond* 207
Mario Coccia. *Does Bureaucracy Affect Research Performance of Public Research Organizations?* 216
Rodrigo Costas and Maria Bordons. *A Classificatory Scheme for the Analysis of Bibliometric Profiles at the Micro Level* 226
Viv Cothey. *Applying Egghe's General Theory of the Evolution of Information Production Processes to the World Wide Web* 231
S.M. Dhawan and B.M. Gupta. *High Productivity Physics Institutions in India: A Study of their Performance in terms of Quantitative and Qualitative Indicators* 241

Leo Egghe. <i>Distributions of the h-index and the g-index</i>	245
Shu Fang, Xian Zhang and Guo-hua Xiao. <i>Research and Application of Patent Map Analysis</i>	254
Francisco Fernández-Izquierdo, Adelaida Román-Román, Cruz Rubio-Liniers, Francisco-Javier Moreno- Díaz-del-Campo, Carmen Martín-Moreno, Carlos García-Zorita, María Luisa Lascurain-Sánchez, Preiddy-Efraín García, Elisa Povedano and Elías Sanz-Casado. <i>Bibliometric Study of Early Modern History in Spain Based on Bibliographic References in National Scientific Journals and Conference Proceedings</i>	266
Enrico Forti, Chiara Franzoni and Maurizio Sobrero. <i>The Effect of Patenting on the Networks and Connections of Academic Scientists</i>	272
Chiara Franzoni, Christopher L. Simpkins, Baoli Li and Ashwin Ram. <i>Using Content Analysis to Investigate the Research Paths Chosen by Scientists Over Time</i>	285
Rainer Frietsch, Sybille Hinze and Pari Patel. <i>Using Patent Data for Monitoring the Globalisation of R&D</i>	295
Antonio García Romero, José Navarrete Cortés, Cristina Escudero, Juan Antonio Fernández López, and Juan Antonio Chaichio Moreno. <i>Measuring the Contribution of Clinical Trials to Bibliometric Indicators: Citations and Journal Impact Factor®</i>	300
Monica Gaughan, Branco Ponomariov and Barry Bozeman. <i>Using Quasi-Experimental Design and the Curriculum Vitae to Evaluate Impacts of Earmarked Center Funding on Faculty Productivity, Collaboration, and Grant Activity</i>	305
Yves Gingras. <i>Mapping the Changing Centrality of Physicists (1900-1944)</i>	314
Wolfgang Glänzel,, Frizo Janssens and Bart Thijs. <i>A Comparative Analysis of Publication Activity and Citation Impact Based on the Core Literature in Bioinformatics</i>	321
Isabel Gómez, María Bordons, M.Teresa Fernández, Fernanda Morillo. <i>Structure and Research Performance of Spanish Universities</i>	335
Yuri Jack Gómez. <i>Revisiting the "Heroic" Age: From Externalism to Internalism in Serial History of Science</i>	346
Iina Hellsten, Renaud Lambiotte, Andrea Scharnhorst and Marcel Ausloss. <i>Self-citation Networks as Traces of Scientific Careers</i> '	365
Jean-Pierre V. M. Hérubel. <i>Pre 1990 French Doctoral Dissertations in Philosophy: A Bibliometric Profile of a Canonical Discipline</i>	368
Kim Holmberg and Mike Thelwall. <i>Local Government Web Sites in Finland: A Geographic and Webometric Analysis</i>	378
Peter A. Hook. <i>Visualizing the Topic Space of the United States Supreme Court</i>	387
Stefan Hornbostel and Susan Böhmer. <i>Determinants for Young Researcher Careers in Germany. Comparative Evaluation of Postdoctoral Programmes</i>	397
Isabel Iribarren-Maestro, María Luisa Lascurain-Sánchez and Elías Sanz-Casado. <i>Are Multi-Authorship and Visibility Related? Study of Ten Research Areas at Carlos III University of Madrid</i>	401
Frizo Janssens, Wolfgang Glänzel and Bart De Moor. <i>A Hybrid Mapping of Information Science</i>	408
Evaristo Jiménez-Contreras, Rafael Bailón Moreno, Daniel Torres-Salinas, Rosario Ruiz Baños, Rafael Ruiz-Pérez, Mercedes Moneda Corrochano and Emilio Delgado López-Cózar. <i>Response Surface Methodology and its Application in Evaluating Scientific Activity</i>	421
Bihui Jin, Ronald Rousseau, Richard P. Suttmeier and Cong Cao. <i>The Role of Ethnic Ties in International Collaboration: The Overseas Chinese Phenomenon</i>	427
Richard Klavans and Kevin W. Boyack. <i>Is there a Convergent Structure of Science? A Comparison of Maps using the ISI and Scopus Databases</i>	437
Vincent Larivière, Éric Archambault and Yves Gingras. <i>Long-Term Patterns in the Aging of the Scientific Literature, 1900–2004</i>	449
Gavin LaRowe, Sumeet Ambre, John Burgoon, Weimao Ke and Katy Börner. <i>The Scholarly Database and Its Utility for Scientometrics Research</i>	457
Katarina Larsen. <i>Interdisciplinarity in Environmental Technology Applications – Examining Knowledge Interaction between Physics and Chemistry Research Teams</i>	463
Thed N. van Leeuwen and Robert J.W. Tijssen. <i>Strength and Weakness of National Science Systems. A Bibliometric Analysis through Cooperation Patterns</i>	469
Jacqueline Leta and Flávio Martins Teixeira. <i>Science in Brazil: Contribution of Male and Female Scientists</i>	480
Jonathan Levitt and Mike Thelwall. <i>Atypical Citation Patterns in the Twenty Most Highly Cited Documents in Library and Information Science</i>	485
Grant Lewison. <i>The References on UK Cancer Clinical Guidelines</i>	489

Loet Leydesdorff and Caroline Wagner. <i>Is the United States Losing Ground in Science? A Global Perspective on the World Science System in 2005</i>	499
Liming Liang. <i>Revealed Similarities between the Journals Nature and Science: Using a New Cluster of Rhythm Indicators</i>	508
Yuxian Liu and Ronald Rousseau. <i>Hirsch-Type Indices and Library Management: The Case of Tongji University Library</i>	514
Ma. Elena Luna-Morales, Francisco Collazo-Reyes and Jane M. Russell. <i>A Quantitative Hstoriography of Mexican Integration into the International Standards of Scientific Research.</i>	523

VOLUME II

V. Markusova, M. Jansz, I. Libkind and A.Varshavsky. <i>Trends in Russian Research Output in Post-soviet Era</i>	542
Elba Mauleón and María Bordons. <i>Women Involvement in Editorial Boards of Mathematics Journals</i>	552
Katherine W. McCain. <i>Analysing Influence Over Time: An Historiographic Mapping of the Research of Conrad Hal Waddington (1905-1975)</i>	558
Lokman Meho and Kiduk Yang. <i>Fusion Approach to Citation-Based Quality Assessment</i>	568
Félix Moya-Anegón, Zaida Chinchilla-Rodríguez, Benjamín Vargas-Quesada, Elena Corera-Álvarez, Antonio Muñoz-Molina, Francisco José Muñoz-Fernández and Rocío Gómez-Crisóstomo. <i>Scientific Output by Gender in Spain (Web of Science, 2004)</i>	582
Ahat A. Nabiullin. <i>Emergence of a New Discipline in the Earth Sciences: Bibliometric Analysis of Photogrammetry and Remote Sensing Literature</i>	594
Ed C.M. Noyons and Clara Calero-Medina. <i>Applying Bibliometric Mapping in a High Level Science Policy Context. Mapping the Research Areas of Three Dutch Universities of Technology</i>	599
José Luis Ortega and Isidro Aguillo. <i>Linear Analysis of Cybermetric Data: Quantifying The European University Web Space</i>	608
Xavier Polanco and Eric Sanjuan. <i>Hypergraph Modelling and Graph Clustering Process Applied to Co-word Analysis</i>	613
Anastassios Pouris and Anthipi Pouris. <i>The State of Science and Technology in Africa (2000-2004): A Scientometric Assessment</i>	619
Ismael Rafols and Martin Meyer. <i>Diversity Measures and Network Centralities as Indicators of Interdisciplinarity: Case Studies in BionanoScience</i>	631
Suzy Ramanana-Rahary, Michel Zitt and Ronald Rousseau. <i>Aggregation Properties of Relative Impact and other Classical Indicators: Convexity Issues and the Yule-Simpson Paradox</i>	643
I.K. Ravichandra Rao. <i>Distributions of Hirsch-Index and G-index: An Empirical Study</i>	655
Victor Rodriguez, Frizo Janssens, Koenraad Debackere and Bart De Moor. <i>On Material Transfer Agreements and Visibility of Researchers in Biotechnology</i>	659
Adelaida Román, Mª-Dolores Alcain and Elea Giménez. <i>Evaluation of Scientific Publications in the Humanities</i>	672
Ana Romero-de-Pablos and Joaquín M. Azagra-Caro. <i>Internationalisation of Patents by Public Research Organisations from a Historical and an Economic Perspective</i>	677
Ulf Sandström and Martin Hällsten. <i>Gender, Funding Diversity and Quality of Research</i>	685
Edgar Schiebel and Marianne Hörlesberger. <i>About the Identification of Technology Specific Keywords in Emerging Technologies: The Case of "Magnetoelectronic"</i>	691
Jesper Wiborg Schneider, Birger Larsen and Peter Ingwersen. <i>Comparative Study between First and All-Author Co-Citation Analysis Based on Citation Indexes Generated from XML Data</i>	696
R. D. Shelton, Patricia Foland, and Roman Gorelsky. <i>Do New SCI Journals Have a Different National Bias?</i>	708
Henry Small and Phineas Upham. <i>Citation Structure of an Emerging Research Area: Organic Thin Film Transistors</i>	718
Alastair Smith. <i>Issues in "Blogmetrics" - Case Studies Using BlogPulse to Observe Trends in Weblogs</i>	726
David Stuart and Mike Thelwall. <i>University-Industry-Government Relationships Manifested Through MSN Reciprocal Links</i>	731
Bart Thijs and Wolfgang Glänzel. <i>A Structural Analysis of Benchmarks on Different Bibliometrical Indicators for European Research Institutes Based on their Research Profile</i>	736
Robert J.W. Tijssen and Thed. N. van Leeuwen. <i>Research Cooperation within Europe: Bibliometric Views of Geographical Trends and Integration Processes</i>	740
Liwen Vaughan and Justin You. <i>Content Assisted Web Co-Link Analysis For Competitive Intelligence</i>	745

Peter Vinkler. <i>Introducing the Contemporary Contribution Index for Characterizing the Recent, Relevant Impact of Journals</i>	753
Martijn S. Visser, Clara M. Calero Medina and Henk F.Moed. <i>Beyond Rankings: the Role of Large Research Universities in the Global Scientific Communication System</i>	761
Liying Yang, Steven A. Morris and Elizabeth M. Barden. <i>Mapping Institutions and Their Weak Ties in a Research Specialty: A Case Study of Cystic Fibrosis Body Composition Research</i>	766
Yang Zhiping, Fang Shu, Cheng Yunwei, Wang Chun,Wen Yi, Hu Zhengyi and Zheng Yi. <i>Profiles of Technological Capabilities of the Chinese Academy of Sciences (CAS). A Comparison of Patenting Activities of the CAS with other National Level Institutions</i>	776
Fuyuki Yoshikane, Takayuki Nozawa, Susumu Shibui and Takafumi Suzuki. <i>An Analysis of the Connection between Researchers' Productivity and their Co-authors' Past Attributions, Including the Importance in Collaboration Networks</i>	783
Weiping Yue, Yuntao Pan, Zheng Ma and Xishan Wu. <i>A Bibliometric Analysis of siRNA Fundamental Research and Patent Activities</i>	792
Michel Zitt,, Elise Bassecoulard, Ghislaine Filliatreau and Suzy Ramanana-Rahary. <i>Revisiting Country and Institution Indicators from Citation Distributions: Profile Performance Measures</i>	797
Alesia Zuccala and Peter van den Besselaar. <i>Mapping Review Networks: Exploring Research Community Roles and Contributions</i>	803
POSTER PAPERS	815
Isidro F. Aguillo, José Luís Ortega and Begoña Granadino. <i>BRIC: Academic Web Contents in the Four Largest Emerging Economies</i>	816
Rafael Aleixandre-Benavent, Gregório González-Alcaide, Alberto Miguel-Dasit, Miguel Castellano-Gómez and Juan Carlos Valderrama Zurian. <i>Citation Analysis and Impact Indicators of the Spanish Medical Journals (2001-2005)</i>	818
Farzaneh Aminpour and Payam Kabiri. <i>Research Performance in Isfahan University of Medical Sciences: Evaluation of Scientific Productivity</i>	820
Farzaneh Aminpour and Payam Kabiri. <i>Webometric Study on Iranian Universities of Medical Sciences</i>	822
Emma Angus, Mike Thelwall and David Stuart. <i>University Groups in Flickr: Tagging for Purpose or Pleasure?</i>	824
Judit Bar-Ilan. <i>The H-Index of H-Index and of Other Informetric Topics</i>	826
Maite Barrios, Angel Borrego, Andreu Vilagines, Marta Somoza and Candela Ollé. <i>Bibliometric Study of Psychological Research on Tourism</i>	828
Ana Isabel Bonilla, Tony Hernández and Isabel Gómez. <i>Scientometric Analysis on a Sample of Scientists in Astronomy, Particles Physics and Multidisciplinary Physics in arXiv.org (1996-2006)</i>	830
Manfred Bonitz. <i>Spontaneity of Consciousness: V.V. Nalimov's Most Ingenious Book</i>	832
R.K. Buter and E.C.M. Noyons. <i>Searching for Converging Research using Citations between Subject Categories: First Results and Challenges</i>	834
Rosaria Rosanna Cammarano, Maurella Della Seta. <i>A Review of New Web Tools for Citation Analysis and their Impact on Research</i>	836
Carolina Cañibano, Javier Otamendi and Inés Andújar. <i>Exploiting the Potential of Researchers CV Databases for Policy Making: Evidence from the Ramón y Cajal Programme in Spain</i>	838
Mª Angeles Coslado and Juan Miguel Campanario. <i>Contribution to the Impact Factor of Academic Journals in the Field of Education and Educational Psychology of Citations to Articles Authored by Editorial Board Members</i>	840
Dang Yaru, Wang Liya, Gao Feng, Li Kan, Huang Yue and He Fenglan. <i>Changes and Development of Key Indicators in JCR on the Web</i>	842
Lixin Chen, Zeyuan Liu and Liming Liang. <i>The Relationship of Price Index and Median Citation Age</i>	844
Ms. Sandhya Diwakar and Dr. KK Singh. <i>Capacity Building in Specialties towards National Health Research through Research Funding</i>	846
Daniela De Filippo, Elías Sanz-Casado and Isabel Gómez. <i>Impact of the Research Stays on the Scientific Output</i>	848
E. García-Carpintero, B. Granadino, N. Sastre and L.M. Plaza. <i>The Spanish Presence in Editorial Boards of International Scientific Journals: A Tool for the Promotion of Spanish Science</i>	850
Marianne Gauffriau, Peder Olesen Larsen, Isabelle Maye, Anne Roulin-Perriard, Markus von Ins. <i>40 Years Discussion on the Counting of Publications</i>	852

Juan Gorraiz, Christian Schlögl. <i>Comparison of two counting houses in the field of pharmacology and pharmacy: Web of Science versus Scopus</i>	854
B.M.Gupta, S.M.Dhawan and R. P. Gupta. <i>S&T Research in India: An Overview of its Research Output and Quality</i>	856
Yusef Hassan-Montero and Víctor Herrero-Solana. <i>Visualizing Library and Information Science from the Practitioner's Perspective</i>	858
Frank Havemann, Michael Heinz, Marion Schmidt, and Jochen Gläser. <i>Measuring Diversity of Research in Bibliographic-Coupling Networks</i>	860
Juan E. Iglesias and Carlos Pecharromán. <i>Comparing H-indices for Scientists in Different ISI Fields</i>	862
Peter Ingwersen, Jesper W. Schneider, Morten Scharff and Birger Larsen. <i>A National Research Profile-Based Immediacy Index and Citation Ratio Indicator for Research Evaluation</i>	864
Abdelali Kaaouachi. <i>Institutional Evaluation in Moroccan Higher Educational System: A Proposed Model</i>	866
Payam Kabiri and Farzaneh Aminpour. <i>Iranian Medical Research Performance: Does it Comply with the National Needs?</i>	868
Hildrun Kretschmer and Theo Kretschmer. <i>Distribution of Co-Author Pairs Frequencies of the Journal of Biological Chemistry Explained as Social Gestalt</i>	870
Mojca Kotar. <i>Author Cocitation Analysis of Polyamides Specialty 1974–1999 and the Modification of Multivariate Statistics</i>	872
Paula Leite and Jacqueline Leta. <i>Productivity and Prestige among Brazilian Scientists</i>	874
Jonathan M. Levitt and Michael Thelwall. <i>Two New Indicators Derived from the h-Index for Comparing Citation Impact: Hirsch Frequencies and the Normalised Hirsch Index</i>	876
Grant Lewison. <i>Breast Cancer Research, Death Rates and Life Expectancy in Iceland: Are they Linked?</i>	878
Grant Lewison. <i>Counting Citations: Fractionation of Addresses and “World-Scale”, a New Scalar</i>	880
Xia Lin. <i>New Visual Interfaces for Author Co-Citation Mapping</i>	882
Szu-chia Scarlett Lo. <i>Patent Indicators of Genetic Engineering Research of EU Countries - Overseas Performance: Preliminary Study</i>	884
Diana Lucio-Arias. <i>A Validation Study of HistCite™: Using the Discoveries of Fullerenes and Nanotubes</i>	886
Renato Fabiano Matheus. <i>Information Science Network – ISN: Social Network Analysis of Scientific Production of LIS Field in Brazil</i>	888
Katherine W. McCain. <i>The Relationship between Influence and Image: Two Views of the Oeuvre of Conrad Hal Waddington using Historiographic Mapping and Author Tri-Citation Image Analysis</i>	890
Katherine W. McCain and Scot Silverstein. <i>Using Historiographic Mapping to Trace Persistent Highly Visible Research Themes in Medical Informatics</i>	892
C.A. Macías-Chapula, J.A. Mendoza-Guerrero, I.P. Rodea-Castro and A. Gutiérrez-Carrasco. <i>Institutional Health Research Collaboration in Mexico. A Bibliometric Study</i>	894
Raúl I. Méndez-Vásquez, Eduard Suñén-Pinyol, Ginés Sanz and Jordi Camí. <i>Characterization of research groups in the cardio-cerebrovascular field. Spain 1996-2004</i>	896
Alberto Miguel-Dasit, Gregorio González-Alcaide, Juan Carlos Valderrama-Zurián, Adolfo Alonso-Arroyo and Rafael Aleixandre-Benavent. <i>Analysis of the Interdisciplinar Collaboration in Spanish Scientific Production on Radiology and Diagnostic Imaging</i>	898
Fernanda Morillo and Daniela de Filippo. <i>The Role of Madrid in the Spanish Regional Collaboration</i>	900
Rogério Mugnaini, Elías Sanz-Casado and Carlos García-Zorita. <i>Ways of Adequacy for Evaluation of Brazilian Scientific Production: National Impact versus International Impact</i>	902
Rogério Mugnaini, Rogério Meneghini and Abel Packer. <i>Citation Profiles in Brazilian Journals of the Scielo Database in Different Scientific Areas</i>	904
Ülle Must. <i>History Research in a Globalized World: a Bibliometric Approach</i>	906
Ryosuke L. Ohniwa, Aiko Hibino and Kunio Takeyasu. <i>Perspective Factor: Past, Present and Future of Life Sciences</i>	908
José Felipe Ortega, Jesús M. González-Barahona and Gregorio Robles. <i>In-Depth Analysis of Wikipedia Community</i>	910
Jose Luis Ortega, Isidro Agullo, Viv Cothey, Andrea Scharnhorst. <i>Exploring Visually the European Academic Web Space</i>	912

Ana Patricia Ortiz, William A. Calo, Carlos A. Suárez, Erik Suárez, Isabel Iribarren and Elías Sanz-Casado. <i>Basic Characteristics of Cancer-related Research in Puerto Rico: An Approach to Local and International Journals</i>	914
Pang Jing'an, Wang Lian and Cao Yan. <i>Quantification Evaluation of Technical Innovation Capabilities of Chinese Large and Medium Industrial Enterprises</i>	916
K. G. Pillai Sudhier. <i>Ranking Journals in Physics: An Informetric Study of the Citations in the Doctoral Theses of the Indian Institute of Science</i>	918
Marios Poulos, Nikolaos Korfiatis and George Bokos. <i>Towards the Construction of a Global Bibliometric Indicator</i>	920
Hampus Rabow. <i>A Bibliometric Analysis of Book Based Literature</i>	922
Cristina Ramo. <i>Research in Doñana National Park: Preliminary data</i>	924
Claude Robert, Concepción S. Wilson, Jean-François Gaudy and Charles-Daniel Arreto. <i>The Evolution of the Sleep Scientific Literature over 30 Years: A Bibliometric Analysis</i>	926
Gwendolyn Rogge and Ronald Rousseau. <i>Is the European Web British or American?</i>	928
Nadine Rons and Arlette De Bruyn. <i>Quantitative CV-Based Indicators for Research Quality, Validated by Peer Review</i>	930
Hervé Rostaing, Nicolas Barts, Valérie Léveillé. <i>Scientific Portfolio Analysis of a Scientific Area by a Competitive Position Approach</i>	932
Samile Andréa de Souza Vanz and Renato Fabiano Matheus. <i>Analyzing Grey Literature from Postgraduate Programs in Social Communication in Brazil: Network of Influence and Citation Analysis</i>	934
Yuan Sun, Masamitsu Negishi and Loet Leydesdorff. <i>National and International Dimensions of the Triple Helix in Japan: University-Industry-Government and International Co-Authorship Relations</i>	936
Dimitar T. Tomov and AbdulKader A. Murad. <i>International Communication Patterns in an Emerging Interdisciplinary Field - Applications of Geographic Information Systems in Public Health</i>	938
Ming-yueh Tsay. <i>An Analysis and Comparison of Citation Data between Journals of Physics, Chemistry and Engineering</i>	940
Yuen-Hsien Tseng, Yu-I Lin, Chun-Hsien Kuo and Yi-Yang Lee. <i>Verification of Increasing Trend Detection</i>	942
Ali Uzun. <i>Recent Trends in Renewable Energy Research: A Bibliometric Perspective</i>	944
Thanh-Trung Van and Michel Beigbeder. <i>From Web Citation to Web Co-citation: Discovering Relatedness on the Web</i>	946
Sonia Vasconcelos, Martha Sorenson and Jacqueline Leta. <i>English Proficiency: A Potential Science Indicator?</i>	948
Jakob Voss. <i>Common Statistics of Heterogeneous Tagging Systems</i>	950
Chun-Chieh Wang , Mu-Hsuan Huang and Dar-Zen Chen. <i>Innovative Capacity Evaluation of Main Countries Based on Patent Analysis</i>	952
Dietmar Wolfram. <i>Search Characteristics in Different Types of Internet-Based IR Environments: Are They the Same?</i>	954
Yasuhiro Yamashita, Sen Ueno, Hiroyuki Tomizawa and Masayuki Kondo. <i>Influence of the International Migration of Researchers on National Publications in Three Fields of Engineering</i>	956
Li-li Yang, Dong-hong Fu, Xiao-ping Bai, Wei-gang Fang, Shu-yin Yao. <i>Citation Analysis of the Chinese Journal of Medical Science Management</i>	958
Lin Yang Lili Yang and Chengzheng Zhao. <i>Assessment of Traditional Chinese Medicine for Substance-Related Disorders Based on Bibliometric Techniques</i>	960
Hairong Yu, Concepción S. Wilson, Mari Davis and Fletcher Cole. <i>Complex Data Modelling for Informetric Research</i>	962
Jian Zhang, Weizhong Zhu, Yunan Chen, Michael Vogeley and Chaomei Chen. <i>Analyzing the Impact of Sloan Digital Sky Survey on Astronomical Literature: A Multiple Perspective Approach</i>	964
Dangzhi Zhao. <i>Factor Rotation Methods in Author Co-citation Analysis: A Comparison</i>	966
M.A. Zulueta, G. García-Gómez, I. Doménech, M. Izquierdo and P. Moscoso. <i>Bibliometry Analysis of the Research on Women and Health</i>	968
AUTHOR INDEX	971

AUTHOR INDEX

A

- Aguillo, Isidro 608, 816, 912
Alcain, Mª-Dolores 672
Aleixandre-Benavent, Rafael 37, 818, 898
Alonso-Arroyo, Adolfo 898
Aubre, Sumeet 457
Aminpour, Farzaneh 820, 822, 868
Andújar, Inés 838
Angus, Emma 824
Archambault, Éric 45, 449
Arreto, Charles-Daniel 926
Ausloss, Marcel 365
Azagra-Caro, Joaquín M. 677

B

- Bai, Xiao-ping 958
Bailón-Moreno, Rafael 421
Barden, Elizabeth M. 766
Bar-Ilan, Judit 52, 826
Barjak, Franz 63
Barrios, Maite 828
Barts, Nicolas 932
Bassecoulard, Elise 74, 797
Batista, Pablo D. 184
Beigbeder, Michel 946
Besselaar, Peter van den 85, 803
Bhattacharya, Sujit 95
Böhmer, Susan 397
Bokos, George 920
Bonilla, Ana Isabel 830
Bonitz, Manfred 832
Bordons, Maria 226, 335, 552
Börner, Katy 112, 457
Bornmann, Lutz 149
Borrego, Angel 828
Bosire Onyancha, Omwoyo 100
Boyack, Kevin W. 112, 124, 437
Bozeman, Barry 305
Braam, Robert 136
Brodin Danell, Jenny-Ann 154
Bruyn, Arlette De 930
Burgoon, John 457
Burrell, Quentin L. 162
Buter, R.K. 834
Butler, Linda 170

C

- Calero-Medina, Clara 179, 599, 761
Calo, William A. 914
Camí, Jordi 896
Cammarano, Rosaria Rosanna 836
Campanario, Juan Miguel 840
Campiteli, Mônica G. 184
Cañibano, Carolina 838
Cao, Cong 427
Cao, Yan 916
Castellano-Gómez, Miguel 818

- Chaichio-Moreno, Juan Antonio 300
Chen, Chaomei 189, 964
Chen, Dar-Zen 952
Chen, Li 201
Chen, Lixin 844
Chen, Yunan 964
Cheng, Yunwe 776
Chinchilla-Rodríguez, Zaida 582
Chu, Heting 207
Coccia, Mario 216
Cole, Fletcher 962
Collazo-Reyes, Francisco 523
Corera-Álvarez, Elena 582
Coslado, Mª Angeles 840
Costas, Rodrigo 226
Cothey, Viv 231

D

- Danell, Rickard 154
Dang, Yaru 842
Daniel, Hans-Dieter 149
Davis, Mari 962
Debackere, Koenraad 659
Delgado López-Cózar, Emilio 421
Dhawan, S.M. 241, 856
Diwakar, Sandhya 846
Doménech, I. 968

E

- Eghe, Leo 245
Escudero, Cristina 300

F

- Fang, Shu 254, 776
Fang, Wei-gang 958
Fernández, M. Teresa 335
Fernández-Izquierdo, Francisco 266
Fernández-López, Juan Antonio 300
Filippo, Daniela de 848, 900
Filliatreau, Ghislaine 797
Foland, Patricia 708
Forti, Enrico 272
Franzoni, Chiara 272, 285
Frenken, Koen 85
Frietsch, Rainer 295
Fu, Dong-hong 958

G

- Gao, Feng 842
García, Preiddy-Efraín 266
García-Carpintero, E. 850
García-Gómez, G. 968
García-Romero, Antonio 300
García-Zorita, Carlos 266, 902
Garfield, Eugene 21
Gaudy, Jean-François 926
Gauffriau, Marianne 852

Gaughan, Monica	305
Giménez, Elea	672
Gingras, Yves	314, 449
Glänzel, Wolfgang	321, 408, 736
Gläser, Jochen	860
Gómez, Isabel	335, 830, 848
Gómez, Yuri Jack	346
Gómez-Crisóstomo, Rocío	582
González-Alcaide, Gregorio	37, 818, 898
González-Barahona, Jesús M.	910
Gorelskyy, Roman	708
Gorraiz, Juan	854
Granadino, Begoña	816, 850
Gupta, B.M.	241, 856
Gupta, R. P.	856
Gutiérrez-Carrasco, A.	894

H

Hällsten, Martin	685
Harnad, Stevan	27
Hassan-Montero, Yusef	858
Havemann, Frank	860
He, Fenglan	842
Heimeriks, Gaston	85
Heinz, Michael	860
Hellsten, Ina	365
Henadeera, Kumara	170
Hernández, Tony	830
Herrero-Solana, Víctor	858
Hérubel, Jean-Pierre V. M.	368
Hibino, Aiko	908
Hinze, Sybille	295
Holmberg, Kim	378
Hook, Peter A.	387
Hörlesberger, Marianne	691
Hornbostel, Stefan	397
Hu, Zhengyi	776
Huang, Mu-Hsuan	952
Huang, Yue	842

I

Iglesias, Juan E.	862
Ingwersen, Peter	696, 864
Ins, Markus von	852
Iribarren-Maestro, Isabel	401, 914
Izquierdo, M.	968

J

Janssens, Frizo	321, 408, 659
Jansz, M.	542
Jiménez-Contreras, Evaristo	421
Jin, Bihui	427

K

Kaaouachi, Abdelali	866
Kabiri, Payam	820, 822, 868
Ke, Weimao	457
Klavans, Richard	112, 437
Kondo, Masayuki	956
Korfiatis, Nikolaos	920

Kotar, Mojca	872
Kretschmer, Hildrun	870
Kretschmer, Theo	870
Krichel, Thomas	207
Kuo, Chun-Hsien	942

L

Lambotte, Renaud	365
Larivière, Vincent	45, 449
LaRowe, Gavin	457
Larsen, Birger	696, 864
Larsen, Katarina	463
Lascurain-Sánchez, María Luisa	266, 401
Lee, Yi-Yang	942
Leeuwen, Thed N. van	469, 740
Leite, Paula	874
Lelu, Alain	74
Leta, Jacqueline	480, 874, 948
Léveillé, Valérie	932
Levitt, Jonathan	485, 876
Lewison, Grant	489, 878, 880
Leydesdorff, Loet	499, 936
Li, Baoli	285
Li, Kan	842
Liang, Liming	508, 844
Libkind, I.	542
Lin, Xia	882
Lin, Yu-I	942
Liu, Yuxian	514
Liu, Zeyuan	844
Lo, Szu-chia Scarlett	884
Lucio-Arias, Diana	886
Luna-Morales, Ma. Elena	523

M

Ma, Zheng	201, 792
Macías-Chapula, C.A.	894
Markusova, V.	542
Martinez, Alexandre S.	184
Martín-Moreno, Carmen	266
Matheus, Renato Fabiano	888, 934
Mauleón, Elba	552
Maye, Isabelle	852
McCain, Katherine W.	558, 890, 892
Meho, Lokman	568
Méndez-Vásquez, Raúl I.	896
Mendoza-Guerrero, J.A.	894
Meneghini, Rogério	904
Meyer, Martin	631
Miguel-Dasit, Alberto	37, 818, 898
Moed, Henk F.	761
Moneda-Corrochano, Mercedes	421
Moor, Bart De	408, 659
Moreno-Díaz-del-Campo, Francisco-Javier	266
Morillo, Fernanda	335, 900
Morris, Steven A.	766
Moscoso, P.	968
Moya-Anegón, Félix	582
Mugnaini, Rogério	902, 904
Muñoz-Fernández, Francisco José	582

Muñoz-Molina, Antonio	582
Murad, AbdulKader A.	938
Must, Ülle	906

N

Nabiullin, Ahat A.	594
Navarrete-Cortés, José	300
Navarro-Molina, Carolina	37
Negishi, Masamitsu.....	936
Noyons, Ed C.M.	179, 599, 834
Nozawa, Takayuki	783

O

Ocholla, Dennis N.....	100
Ohniw, Ryosuke L.	908
Olesen Larsen, Peder.....	852
Ollé, Candela.....	828
Ortega, José Felipe.....	910
Ortega, Jose Luis.....	608, 816, 912
Ortiz, Ana Patricia	914
Otamendi, Javier	838

P

Packer, Abel.....	904
Pan, Yuntao.....	201, 792
Pang, Jing'an.....	916
Patel, Pari	295
Pecharromán, Carlos	862
Peritz, Bluma C.....	52
Pillai Sudhier, K.G.	918
Plaza, L.M.....	850
Polanco, Xavier	613
Ponomariov, Branco.....	305
Poulos, Marios	920
Pouris, Anastassios.....	619
Pouris, Anthipi	619
Povedano, Elisa.....	266

R

Rabow, Hampus.....	922
Rafols, Ismael	631
Ram, Ashwin.....	285
Ramanana-Rahary, Suzy	643, 797
Ramo, Cristina	924
Rao, I.K.Ravichandra.....	655
Robert, Claude	926
Robinson, Simon.....	63
Robles, Gregorio	910
Rodriguez, Victor.....	659
Rogge, Gwendolyn.....	928
Román-Román, Adelaida	266, 672
Romero-de-Pablos, Ana	677
Rons, Nadine	930
Rostaing, Hervé.....	932
Roulin-Perriard, Anne	852
Rousseau, Ronald.....	427, 514, 643, 928
Rubio-Liniers, Cruz.....	266
Ruiz-Baños, Rosario	421
Ruiz-Pérez, Rafael	421
Russell, Jane M.	523

S

Sandström, Ulf	685
Sanjuan, Eric	613
Sanz, Ginés	896
Sanz-Casado, Elías	266, 401, 848, 902, 914
Sastre, N.....	850
Scharff, Morten	864
Scharnhorst, Andrea	365, 912
Schiebel, Edgar	691
Schlögl, Christian.....	854
Schmidt, Marion	860
Schneider, Jesper W.....	696, 864
Seta, Maurella Della.....	836
Shelton, R. D.....	708
Shibui, Susumu	783
Silverstein, Scot	892
Simpkins, Christopher L.	285
Singh, Dr. KK	846
Small, Henry	718
Smith, Alastair	726
Sobrero, Maurizio	272
Somoza, Marta	828
Song, Il-Yeol.....	189
Sorenson, Martha	948
Souza Vanz, Samile Andréa de.....	934
Stuart, David	731, 824
Su, Cheng	201
Suárez, Carlos A.	914
Suárez, Erik	914
Sufén-Pinyol, Eduard	896
Suttmeyer, Richard P.	427
Suzuki, Takafumi	783

T

Takeyasu, Kunio	908
Teixeira, Flávio Martins.....	480
Thelwall, Mike	378, 485, 731, 824, 876
Thijs, Bart	321, 736
Tijssen, Robert J.W.	469, 740
Tomizawa, Hiroyuki	956
Tomov, Dimitar T.	938
Torres-Salinas, Daniel.....	421
Tsay, Ming-yueh	940
Tseng, Yuen-Hsien.....	942

U

Ueno, Sen	956
Upham, Phineas	718
Uzun, Ali	944

V

Valderrama-Zurián, Juan Carlos	37, 818, 898
Van, Thanh-Trung	946
Vargas-Quesada, Benjamín	582
Varshavsky, A.	542
Vasconcelos, Sonia	948
Vaughan, Liwen	745
Vilagines, Andreu	828
Vinkler, Peter	753
Visser, Martijn S.	761

Vogeley, Michael	964
Voss, Jakob	950

W

Wagner, Caroline	499
Wang, Chun	776
Wang, Chun-Chieh.....	952
Wang, Lian.....	916
Wang, Liya.....	842
Wen, Yi	776
Wilson, Concepción S.....	926, 962
Wolfram, Dietmar	954
Wu, Yishan	201, 792

X

Xiao, Guo-hua	254
---------------------	-----

Y

Yamashita, Yasuhiro	956
Yang, Kiduk.....	568
Yang, Lili	958, 960
Yang, Lin	960
Yang, Liying	766
Yang, Zhiping	776
Yao, Shu-yin	958
Yoshikane, Fuyuki	783
You, Justin	745
Yu, Hairong.....	962
Yuan, Sun.....	936
Yue, Weiping	792

Z

Zhang, Jian.....	964
Zhang, Xian.....	254
Zhao, Chengzheng	960
Zhao, Dangzhi	966
Zheng, Yi	776
Zhu, Weizhong.....	189, 964
Zitt, Michel	74, 643, 797
Zuccala, Alesia.....	803
Zulueta, M.A.....	968

» KEYNOTES

ISSI 2007 - Madrid

From The Science of Science to Scientometrics. Visualizing the History of Science with *HistCite* Software

Eugene Garfield

garfield@codex.cis.upenn.edu

Chairman Emeritus, Thomson ISI,
3501 Market Street, Philadelphia PA 19104 (USA)

Abstract

While ISSI was founded in 1993, scientometrics and bibliometrics are now at least half a century old. Indeed, the field can be traced to early quantitative studies in the early 20th Century. In the thirties, it evolved to the “science of science.” The publication of J. D. Bernal’s *Social Function of Science* in 1939 was a key transition point but the field lay dormant until after World War II, when DJD Price’s books *Science Since Babylon* in and *Little Science, Big Science* were published in 1961 and 1963. His role as the “father of scientometrics” is clearly evident by using the *HistCite* software to visualize his impact as well as the subsequent impact of the journal *Scientometrics* on the growth of the field. *Scientometrics* owes its name to V. V. Nalimov, the author of *Naukometriya*, and to Tibor Braun who adapted the neologism for the journal. The primordial paper on citation indexing by Garfield which appeared in *Science* 1955 became a bridge between Bernal and Price. The timeline for the evolution of scientometrics is demonstrated by a HistCite tabulation of the ranked citation index of all the 100,000 references cited in the 3,000 papers citing Price.

Keywords

history of scientometrics; etymology of science; Derek J.D. Price; V. V. Nalimov; J. D. Bernal; Science of Science; HistCite; algorithm; historiography; bibliometrics.

Introduction

When Henk Moed asked me to present a keynote address to this Eleventh International Conference of the International Society for Scientometrics and Informatics (ISSI) I had mixed feelings. I had previously planned to participate by simply describing my current work on algorithmic historiography. The paper I originally planned to submitted was an up-to-date description of the *HistCite* system (<http://www.histcite.com/>). Briefly stated, *HistCite*™ is a software system which generates chronological maps of bibliographic collections resulting from subject, author, institutional or source journal searches of the *ISI Web of Science*.® *WoS* export files are created in which all cited references for each source document are captured. The software generates chronological historiographs highlighting the most-cited works in the retrieved collection. Other listings include rankings by author, journal, institution, or vocabulary.

But Henk thought that this might be a good chance to provide the current ISSI membership with some personal reflections on the origins of scientometrics, especially as it is now two decades since the first ISSI conference held in Belgium in 1987 and 14 years since ISSI was founded in Berlin. It is noteworthy that the term “scientometrics” itself was not included in the title of the 1987 meeting which was called the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval. Twenty years earlier, Alan Pritchard had coined the term bibliometrics in his paper on statistical bibliography which defined the term bibliometrics. (Pritchard, 1969).

Most of us have been exposed to the macro history of scientometrics. We recognize names like Derek de Solla Price and V.V. Nalimov and perhaps earlier pioneers in measurement such as Alfred Lotka and George K. Zipf. If you search the *Web of Science* for the past century, these names will pop up very quickly. But when you search year-by-year you obtain a very different micro-perspective. Today, I would like to recall for you aspects of the micro and macro impact of Derek Price’s work, since he is usually considered “the father of Scientometrics.” But this simplistic metaphor for the history and evolution of his role in scientometrics, does not adequately reflect the influences of earlier statistically and quantitatively oriented scholars.

In the foreword to the second edition of “Little Science, Big Science,” (Merton and Garfield, 1986) Robert K. Merton and I identified Derek as the father of scientometrics because he was perceived, in the western world, to have made the greatest impact on the use of quantitative indicators in formulating science policy. The first edition of the 1963 book was aptly identified later as a *Citation Classic* (Price, 1983) but at the time the book was written, Derek had not even encountered the term scientometrics, which was coined by the Russian mathematician-philosopher-polymath, V. V. Nalimov. “Scientometrics” is the English translation of the title word of Nalimov’s classic monograph *Naukometriya*,ⁱ (Nalimov and Mul’chenko, 1973) which was relatively unknown to western scholars even after it was translated into English. Without access to the internet and limited distribution, it was rarely cited. However, the term became better known once the journal *Scientometrics* appeared in 1978. Stephen Bensman in a tribute to Tibor Braun recently reminded us how the journal became a bridge between the East and West. (Bensman & Kraft, in press)

Let me remind you of some historical facts. Price’s “*Science Since Babylon*” (Price 1986) was published six years after my 1955 paper in *Science* (Garfield, 1955). The first edition of *Little Science, Big Science* appeared two years later in 1963. The opening page is called a “prologue to a science of science.” If Derek was aware of my paper, he did not cite it then. Even in his classic 1965 *Networks* paper in *Science* (Price, 1965) he referred to the 1963 *Genetics Citation Project* and my 1964 *Science* paper by which time we had made personal contact (Garfield, 1964). But even earlier, in 1962, I had written to J.D. Bernal and Robert K. Merton about the experimental *Science Citation Index* which resulted from that project. I met Bernal briefly at the International Conference on Scientific Information in Washington in 1958. It was not until 1983, in his Citation Classic commentary (Price, 1983) cited above, that Derek notes that he was “stimulated much by Robert Merton’s writings in the sociology of science, by Eugene Garfield’s new book on citation indexing, and by rereading Desmond Bernal’s books which had prepared my mind for the initial sensitivity that led me to this field in the first place.” Of course, Derek could not have read my book at that time because it did not come out until 1979. Perhaps he should have used the term “work” instead.

In the preface to Volume 3 of my *Essays of an Information Scientist*, (Price DJD, 1980) Derek himself related how we first encountered each other when he was a member of the National Science Foundation’s Science Information Council. He reports how I tried to get the NSF to support printing and distribution of the *Science Citation Index*:

From that day to the present....I have found megavitamins for my intellectual diet on the cutting room floor of ISI’s computer room. Bit by bit we have begun to understand how citations work and in the course of this, there has emerged a new sort of statistical sociology of science that has thrown light on many aspects of the authorship, refereeing, and publication of scientific research papers. The Society of Social Studies in Science now has an annual meeting devoted to this new method of understanding science that has grown, almost as an accidental by-product, from the indexing technology developed by the Institute for Scientific Information. Our initial intuitive perceptions have turned out to be correct.

The early 4S group ultimately became the Society for the Social Studies in Science (4S) which together with ISI sponsors the annual Bernal Award. The Society’s interest in scientometrics has waned considerably in recent years, perhaps in part because of the growth of ISSI which understandably is not as preoccupied with the history and sociology of science as is 4S.

The first co-citational link between Garfield and Price was made in the early sixties by the mathematical statistician, John W. Tukey (Tukey 1962). Between 1955 and 1964 he was the only author who co-cited me and Derek. Keep in mind that Tukey was not a scientometrician. Like myself at the time, he was primarily interested in helping scientists to keep in touch with the literature. He and Joshua Lederberg played a key role, especially through the Weinberg Committee report, in promoting the idea of citation indexes as a new and promising method for information retrieval. No one was then actively talking about citation indexing as a scientometric tool per se. Alan Pritchard’s paper on “Statistical Bibliography,” mentioned earlier, did not appear until 1969.

Another early science policy scholar was the Yugoslav Stevan Dedijer. (Dedijer, 1962) Like Tukey he was aware of the work by Derek Price but in those early years there were only vague references to the use of bibliometric data for science policy purposes. Rather, the term “science of science” was used by Price , (Price 1975), Maurice Goldsmith), and others to describe the pioneering work of J.D. Bernal and its offshoots. However, the term did not gain favor even though the Society for the Social Study of Science (4S) was formed in 1975. I plan to present a more detailed analysis of Bernal’s work at the forthcoming celebration of his 100th birthday in Ireland in September.

Using citations to the work of Price as one indicator of the growth of this field here is the year-by-year graph of citations to Derek’s work based on using the histogram feature of *HistCite* or *Web of Science* (*WoS*).

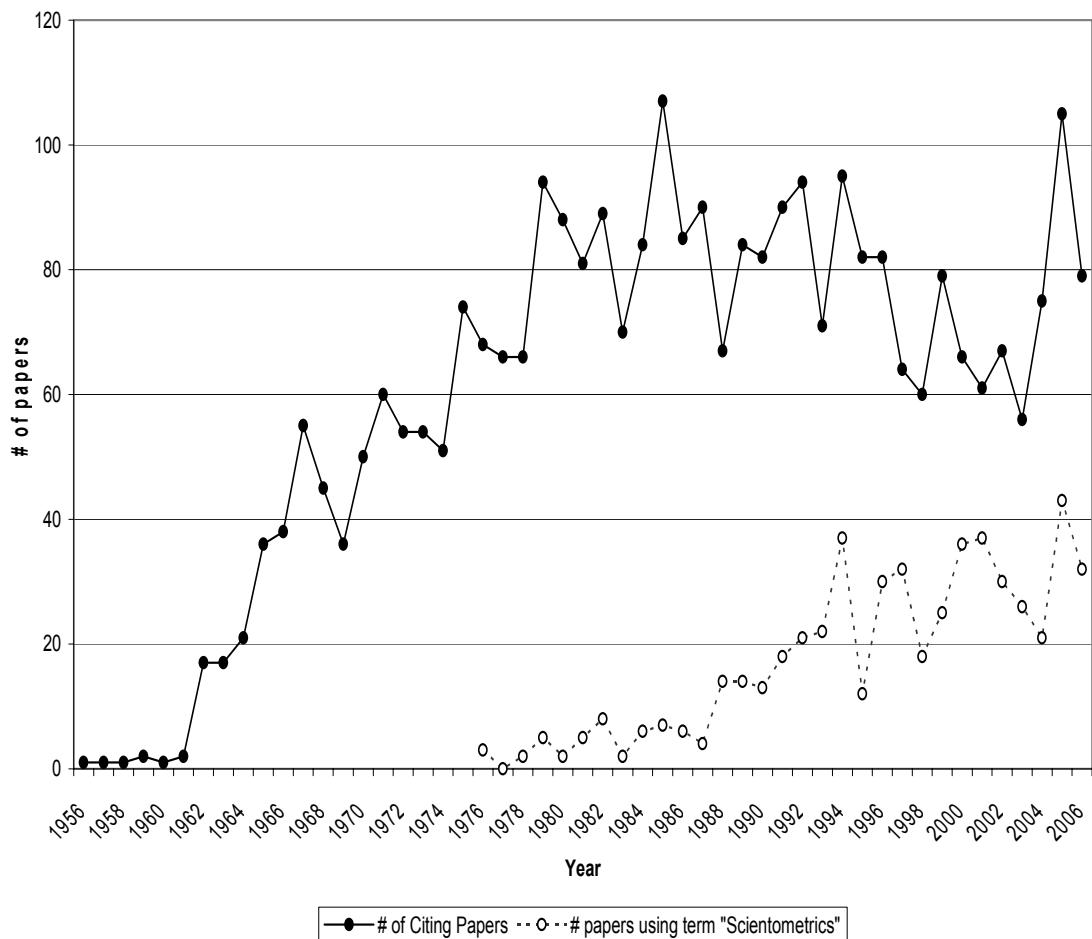


Figure 1. Papers citing Price versus Papers using the term “scientometrics,” from 1956-2006

In contrast to the visible growth in citations to Price’s work, an analysis of papers published in *WOS* containing the term scientometric(s) does not reveal the growth of the topic because the general term is displaced by more specific terminology as the field evolved (Figure 1).

To continue this brief discussion of the work of Derek Price, the following historiograph displays the linkages between the 35 most-cited works of the *HistCite* collection. Each of these papers was cited at least 107 times.

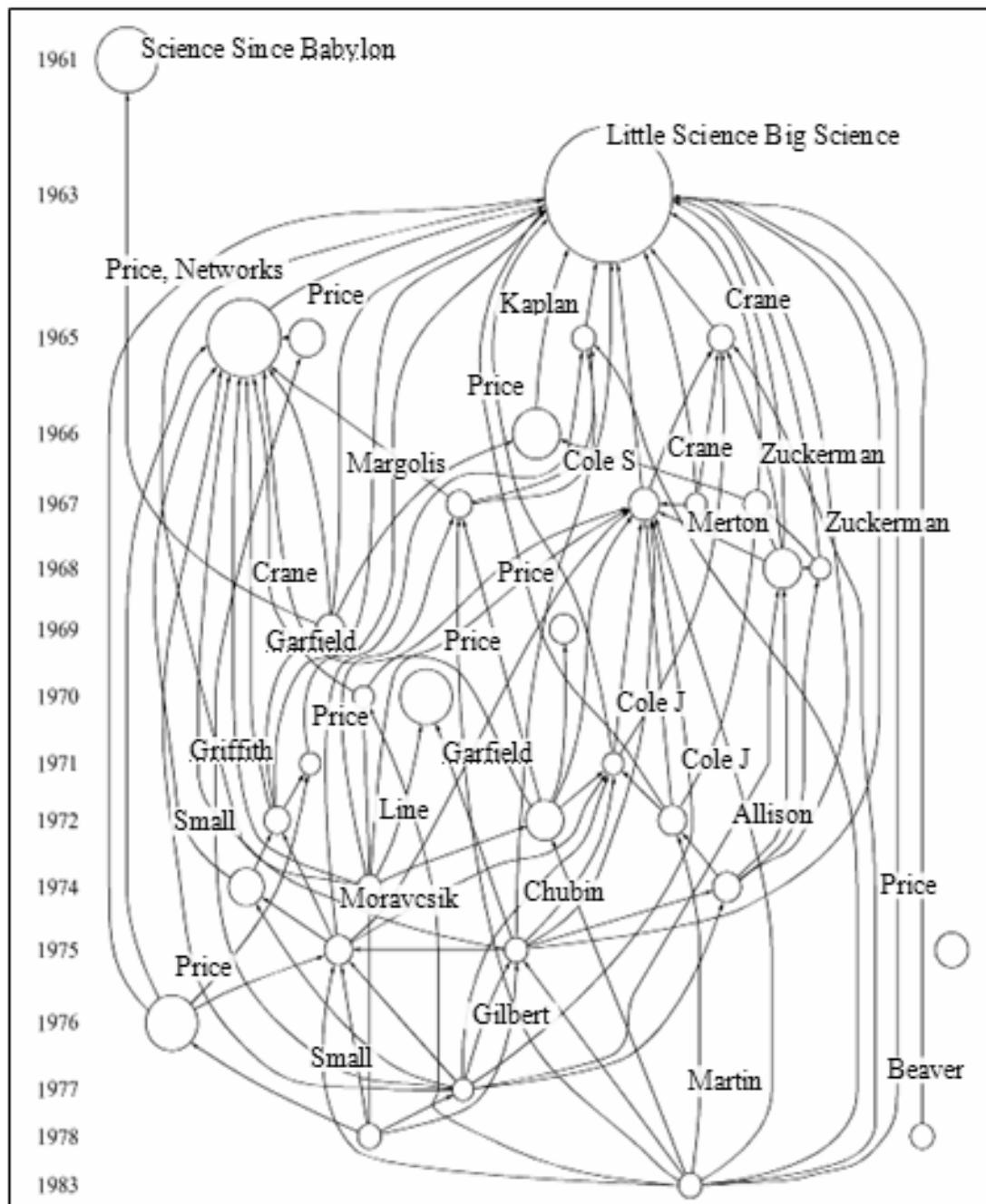


Figure 2. Historiograph of 33 most-cited works in the collection if papers citing Price from 1956-2006

	Author, year, reference	Cites
1	COLE FJ, 1917, SCI PROGR, V11, P578	36
2	LOTKA AJ, 1926, J WASHINGTON ACADEMY, V16, P317	213
3	GROSS PLK, 1927, SCIENCE, V66, P385	39
4	BRADFORD SC, 1934, ENGINEERING-LONDON, V137, P85	69
5	BERNAL JD, 1939, SOCIAL FUNCTION SCI	42
6	BUSH V, 1945, ATLANTIC MONTHLY, V176, P101	65
7	BRADFORD SC, 1948, DOCUMENTATION	84
8	VICKERY BC, 1948, J DOC, V4, P198	24
9	ZIPF GK, 1949, HUMAN BEHAVIOR PRINCIPLE	86
10	FUSSLER HH, 1949, LIBRARY Q, V19, P19	40
11	BARBER B, 1952, SCIENCE SOCIAL ORDER	36
12	LEHMAN HC, 1953, AGE ACHIEVEMENT,	33
13	SIMON HA, 1955, BIOMETRIKA, V42, P425	76
14	GARFIELD E, 1955, SCIENCE, V122, P108	57
15	PRICE DJD, 1956, DISCOVERY, V17, P240	28
16	MERTON RK, 1957, AM SOCIOl REV, V22, P635	76
17	MERTON RK, 1957, SOCIAL THEORY SOCIAL	48
18	SHOCKLEY W, 1957, P IRE, V45, P279	39
19	POPPER K, 1959, LOGIC SCI DISCOVERY	39
20	BURTON RE, 1960, AM DOC, V11, P18	69
21	WESTBROOK JH, 1960, SCIENCE, V132, P1229	27
22	KENDALL MG, 1960, OPERATIONAL RESEARCH, V11, P31	25
23	PRICE DJD, 1961, SCI SINCE BABYLON, P1	337
24	MERTON RK, 1961, P AM PHILOS SOC, V105, P470	35
25	BARBER B, 1961, SCIENCE, V134, P596	30
26	KUHN TS, 1962, STRUCTURE SCI REVOLUTION	199
27	MACHLUP F, 1962, PRODUCTION DISTRIBUT	41
28	ROGERS EM, 1962, DIFFUSION INNOVATION,	27
29	PRICE DJD, 1963, LITTLE SCIENCE BIG SCIENCE, P1	1454
30	KESSLER MM, 1963, AM DOC, V14, P10	61
31	GARFIELD E, 1963, AM DOC, V14, P289	28
32	GARFIELD E, 1963, AM DOC, V14, P195	27
33	GARFIELD E, 1964, USE CITATION DATA WR,	51
34	GARFIELD E, 1964, SCIENCE, V144, P649	37
35	CLARKE BL, 1964, SCIENCE, V143, P822	31
36	PRICE DJD, 1964, SCIENCE, V144, P655	30
37	PRICE DJD, 1965, SCIENCE, V149, P510	499
38	HAGSTROM WO, 1965, SCIENTIFIC COMMUNITY	214
39	PRICE DJD, 1965, TECHNOL CULT, V6, P553	122
40	CRANE D, 1965, AM SOCIOl REV, V30, P699	63
41	KAPLAN N, 1965, AM DOC, V16, P179	50
42	PRICE DJD, 1965, NATURE, V206, P233	33
43	PRICE DJD, 1966, AM PSYCHOL, V21, P1011	213
44	BAYER AE, 1966, SOCIOl EDUC, V39, P381	53
45	CARTTER AM, 1966, ASSESSMENT QUALITY G,	42
46	STORER NW, 1966, SOCIAL SYSTEM SCI,	39
47	SCHMOOKLER J, 1966, INVENTION EC GROWTH,	33
48	BENDAVID J, 1966, AM SOCIOl REV, V31, P451	29
49	STORER NW, 1966, SOCIAL SYSTEM SCIENC,	26
50	MAY KO, 1966, SCIENCE, V154, P1672	24
51	COLE S, 1967, AM SOCIOl REV, V32, P377	91
52	MARGOLIS J, 1967, SCIENCE, V155, P1213	62
53	ZUCKERMAN H, 1967, AM SOCIOl REV, V32, P391	61
54	CRANE D, 1967, AM SOCIOl, V2, P195	44
55	LEIMKUHLER FF, 1967, J DOC, V23, P197	40
56	PRICE DJD, 1967, SCI TECHNOL, V70, P84	33
57	MERTON RK, 1968, SCIENCE, V159, P56	128
58	ZIMAN J, 1968, PUBLIC KNOWLEDGE SOC	68
59	ZUCKERMAN H, 1968, AM J SOCIOl, V74, P276	47
60	BROOKES BC, 1968, J DOC, V24, P247	40
61	MULLINS NC, 1968, AM SOCIOl REV, V33, P786	38
62	MERTON RK, 1968, SOCIAL THEORY SOCIAL	37
63	COLE S, 1968, AM SOCIOl REV, V33, P397	32
64	WATSON JD, 1968, DOUBLE HELIX,	24
65	CRANE D, 1969, AM SOCIOl REV, V34, P335	73
66	PRICE DJD, 1969, P ISRAEL ACAD SCI HU, V4, P98	69
67	PRITCHARD A, 1969, J DOC, V25, P348	47
68	FAIRTHORNE RA, 1969, J DOC, V25, P319	46
69	BROOKES BC, 1969, NATURE, V224, P953	40
70	MACRAE D, 1969, AM SOCIOl REV, V34, P631	34
71	PRICE DJD, 1969, FACTORS TRANSFER TEC, V1, P91	30

Figure 3. Time Line for the History of Scientometrics

The chronological listing of the 200 most-cited works, based on over 102,000 cited references in the collection of 3083 citing papers provides a fairly accurate historical timeline of the field.

Starting with F. J. Cole in 1917, AJ Lotka in 1926, Gross & Gross in 1927, Samuel Bradford in 1934, and then Bernal in 1939. Vannevar Bush's classic, "As we may think" appeared in 1945 at the end of World War II (Bush, 1945). A decade later, we find the work of Herb Simon in 1955, and in the same year, the paper by yours truly. Then in 1956 Derek's first paper on "the exponential growth in science," appears in 1956 (Price, 1956). I won't continue to recite all the names that are recalled in this exercise but I believe this list of works cited 30 or more times in the Price *HistCite* collection demonstrates the simple notion that bibliographic history is recapitulated rather well by the collective bibliographic memory of the scholars who have contributed to the literature, both at the macro and micro level of analysis.

References

- Bensman, SJ & Kraft, DH. (In Press.) Happy 75th Birthday, Tibor Braun, *Scientometrics*.
- Bush, V. (1945). As we may think it. *Atlantic Monthly*, 176:101.
- Merton RK & Garfield E, (1986) Foreword to Little science, Big science.. and beyond, New York: Columbia University Press, 301 pgs.
- Nalimov VV and Mul'chenko ZM. (1969). Naukometriya. Izuchenie nauki kak informatsionnogo protessa (Scientometrics. Study of science as an information process.) Moscow: Nauka, 192 pgs. (Available in English on microfilm: Measurement of science. Study of the development of science as an information process. Washington, DC: Foreign Technology Division, U.S. Air Force Systems Command, 13 October 1971.196pgs.). Retrieved March 23, 2007 from: <http://www.garfield.library.upenn.edu/nalimov/nalimovmeasurementofscience/book.pdf>
- Pritchard A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, 25, 348–349
- Price DJD. (1956). The exponential curve of science. *Discovery* 17(1):240-243.
- Price DJD. (1963). *Little Science, Big Science*. New York: Columbia University Press.
- Price Derek J. deSolla, (1965). Networks of Scientific Papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, 149(3683):510-515.
- Price DJD. (1976). *Science since Babylon*. New Haven: Yale University Press.
- Price DJD. (1975). *Science since Babylon*. New Haven, CT: Yale University Press.
- Price, DJD. (1980). Foreward to *Essays of an Information Scientist*, Volume 3. Philadelphia: ISI Press
- Price DJD. (1983). This week's Citation Classic, *Current Contents* No. 29, pg. 18. Retrieved March 23, 2007 from: <http://garfield.library.upenn.edu/classics1983/A1983QX23200001.pdf>
- Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science* 122(3159):108-11. Retrieved March 23, 2007 from: http://www.garfield.library.upenn.edu/papers/science_v122v3159p108y1955.html
- Garfield, E. (1964). *Science Citation Index -- A New Dimension in Indexing*. *Science*, 144(3619): 649-54.
- Tukey JW. (1962). Keeping Research in Contact with Literature - Citation Indexes and beyond. *IEEE Transactions on Engineering Writing and Speech*. EWS5(2):78. Dedijer S. (1962). Measuring Growth of Science, *Science*. 138(3542):781.

Open Access Scientometrics and the UK Research Assessment Exercise

Stevan Harnad

<http://www.crsc.uqam.ca/>

Institut des sciences cognitives Université du Québec à Montréal
Montréal, Québec (Canada)

<http://www.ecs.soton.ac.uk/~harnad/>

Department of Electronics & Computer Science University of Southampton
Highfield, Southampton (UK)

Abstract

Scientometric predictors of research performance need to be validated by showing that they have a high correlation with the external criterion they are trying to predict. The UK Research Assessment Exercise (RAE) together with the growing movement toward making the full-texts of research articles freely available on the web offer a unique opportunity to test and validate a wealth of old and new scientometric predictors, through multiple regression analysis: Publications, journal impact factors, citations, co-citations, citation chronometrics (age, growth, latency to peak, decay rate), hub/authority scores, h-index, prior funding, student counts, co-authorship scores, endogamy/exogamy, textual proximity, download/co-downloads and their chronometrics, etc. can all be tested and validated jointly, discipline by discipline, against their RAE panel rankings in the forthcoming parallel panel-based and metric RAE in 2008. The weights of each predictor can be calibrated to maximize the joint correlation with the rankings. Open Access Scientometrics will provide powerful new means of navigating, evaluating, predicting and analyzing the growing Open Access database, as well as powerful incentives for making it grow faster.

Keywords

Open Access; Research Assessment Exercise; United Kingdom

Scientometrics probably began as Learned Journals took over from peer-to-peer scholarly letter-writing in the 17th century (Guédon 2002), but it came into its own in our own ‘publish-or-perish’ era. With Garfield’s (1955) contributions to citation counting and indexing in the 1950’s, the academic bean-counting of publications for performance evaluation and funding came to be supplemented by citation counting. It was no longer enough just to publish in bulk: it had to be demonstrable that your publications were also heavily used, hence useful and important. A direct indicator of usage was the fact that your research was cited by subsequent research (Moed 2005).

Of course ‘heavy’ varies with the field. There are industrial-scale research areas and sparse esoteric ones, having only a few peers worldwide (rather as in the letter-writing era). So it was always true (if not always taken into account) that citation counts would have to be used with caution, always comparing like with like rather than turnips with truffles. It makes no sense to point out that a paper or author in cancer research has more citations than a paper or author in Finno-Ugric philology. Nor that a full professor in the fullness of his years has more citations than a fresh post-doc. Journals too are hard to compare, unless their subject matter is closely equated. And although for a while it became the most popular bean-counting method, using the ‘journal impact factor’ (the average number of citations per article) to weigh publications is like using the average graduating marks of their secondary schools to weigh incoming university applicants. One wants the applicant’s own exact marks, and then one wants to know what those marks mean.

Psychometrics and Test Validation.

In weighing the meaning of metrics, scientometrics can perhaps take a lesson from psychometrics (Kline 2000). In the field of aptitude testing, the tests are constructed and validated against external criteria. One first starts by inventing test items, picking items that agree with one another in polarity: A set of items is given to a large number of testees, and their average scores on half of the test items are compared with their average scores on the other half. If the correlation is not very high, then the

test does not even agree with itself, let alone predict something external to itself. Split-half correlations are called a test's reliability. Test/re-test correlations are a further measure of reliability.

Now suppose we have a reliable test, with split-halves and test/re-test being highly correlated: Is the test valid? Does it measure anything other than itself? After all, in aptitude testing, we cannot afford to define aptitude 'operationally' -- as simply amounting to the score on the reliable test that I happen to have constructed and baptized a test of aptitude. Aptitude tests do not have face validity. We have to show that their score is predictive, in that it correlates with something else that we already agree to call aptitude.

In this respect, aptitude testing is like weather-forecasting. It is not enough to show that measures of barometric pressure are reliable, and tend to agree with themselves when measured repeatedly or in different ways. It has to be shown that they predict rain. If there is a high correlation between barometric pressure and probability of precipitation, then barometric pressure is validated as a predictor of impending rain. In the same way, we can correlate our aptitude test scores with some external measure of aptitude (age-related school performance norms perhaps, or postgraduation job performance, or human judgments of individuals' relative intelligence). This external variable against which psychometricians validate their test scores is called the 'criterion', and the measure of the validity of a test is essentially the size of its correlation with its criterion (or criteria).

Now note the difference between psychometrics and meteorology: Barometric pressure either does or does not correlate with the likelihood of rain. If it doesn't, then we need to look for another predictor. In the case of aptitude tests, if we have an unreliable test, one that does not agree with itself, we can keep constructing new items and discarding old ones until we have made the test internally reliable. Then, if a reliable test has a low correlation with its external criterion, we can try discarding old items and constructing new ones in order to try to raise the test's correlation with its criterion to the level of predictive validity we need for normative use. In meteorology, if the correlation between pressure and rain is not high enough, we can't change the measure of pressure, but we can add more predictors, for example, current humidity, or the speed and direction of an approaching cold front. The measures can then be combined in a multiple regression equation, in which, say, 3 predictors, each with its own regression (or 'beta') weight, are used jointly to predict the criterion.

The same can be done with psychometric test validation. If the validity of a single test is too low, we can add more tests : tests of verbal ability, tests of numerical ability, tests of spatial ability – to see whether, jointly, they do a better job of predicting our criterion.

Multivariate Metrics.

Now contrast this with the case of scientometrics, and citation counting in particular. Although I don't know whether it has been tested formally, let us assume that citation counts are indeed reliable, in that if you randomly split an author's works in half, there will be a good split-half correlation, and that if you take successive citation counts in two different time windows, they will be sufficiently correlated with one another too.

Now we proceed to validation : Correlations between citations and various criteria (e.g., performance indicators, such as funding, productivity, prizes) have been measured, much the way correlations between psychometric scores and their criteria have been measured. And those scientometric correlations have turned out to be whatever they have been : sometimes higher sometimes lower, but never, I would suggest, stellar – never approaching the kind of predictivity there is in meteorology or aptitude psychometrics or medical biometrics. And the reason validity is not higher is that we cannot really discard old test items and construct new ones to improve the correlation of citations with their criterion, as we do in psychometrics. Citation counts are citation counts. So we have thus far been rather passive about the validation of our scientific and scholarly performance metrics, taking pot-luck rather than systematically trying to increase their validity, as in psychometrics.

For a long time we simply stuck superstitiously with the journal citation impact factor. As noted, in evaluating individual papers or authors the average citation count of the journal in which their paper was published is a rather blunt instrument. Performance evaluation committees have since begun to look also at exact citation counts for papers or authors. There have been some gestures toward using co-citations too (after all, it should make a difference if a paper or author is co-cited with a graduate student versus a Nobel Laureate). The Institute for Scientific Information developed an ‘immediacy index’ for journals ; in principle, similar time-based measures could index an individual paper’s or author’s citation growth, latency to peak, and decay rate. Let’s call those ‘citation chronometrics’. Then there were ‘hub’ and ‘authority’ scores, measuring citation fan-out and fan-in (Kleinberg 1999). The authority score is a measure similar to Google’s PageRank, that recursively weights incoming citations by their own respective incoming citation weights (Page et al. 1999). Lately we have also had Hirsch’s (2005) h-index. Online download counts have likewise entered the arena. Many other metrics are possible too: co-authorship, endogamy/exogamy (how narrow or wide and even cross-disciplinary is a paper’s or author’ citation fan-in, fan-out, and wider interconnectivity?), textual proximity (degree of textual overlap, and latent semantic distance metrics ; Landauer et al 1998), download chronometrics, perhaps eventually even co-download measures. But for some reason, we have so far tended to use these metrics one at a time, as if both predictor and criterion were univariate, rather than trying to combine them into multivariate measures that might have higher predictive power.

The first thing psychometricians would do with a ‘battery’ of univariate metrics would be to systematically validate them against external criteria that already have some face validity for us : There are other classic performance measures, such as funding, doctoral student counts, and prizes -- but, frankly, using those would be circular, as they have not been externally validated either. What psychometricians sometimes do first with their batteries of diverse metrics is to look at their intercorrelational structure through principal-component and factor analyses. What has emerged, somewhat surprisingly, from factor-analyzing many diverse aptitude tests jointly is a single primary factor – the General Intelligence or ‘G’ factor – that is common to all of them. There are those who think G is some sort of artifact of the test construction and correlational methods used, but most psychometricians think G is actually an empirical finding : that there is a single basic human intellectual ability underlying all the other special abilities (Kline 2000).

Aptitude tests differ in the size of their G ‘loadings’, but all of them have some positive G loading. No one test, however, is a direct measure of G alone. It requires a battery of tests to get a picture of G (and so far individuals are not characterised as having a given ‘G Score’, but as having a composite of scores on a variety of different aptitude tests). Intelligence is multidimensional even if there is a shared underlying factor. Hence aptitude tests have to be multidimensional. But aptitude testing has the advantage over scientometrics that most of the various specific aptitude tests (verbal reasoning, mathematical reasoning, spatial visualization, short-term memory, reaction time, musical ability, motor coordination, etc., some of them psychometric, some of them biometric) have been validated against their respective criteria. This is not true in scientometrics: Citation and other metrics have been used, and their pairwise correlations with criteria have been calculated and reported, but nothing like the systematic validation process that goes into the construction and use of aptitude tests – calibrating and optimizing on the basis of data from generation after generation of students -- has been done with scientometric measures. Nothing like standardized ‘norms’ or benchmarks has as yet emerged from scientometrics.

It has to be said that this is partly because the database containing the most important of the scientometric indicators – citations -- has been in the proprietary hands of one sole database provider for decades, with parts of it temporarily leased (at no small cost) to those who wished to do some data-mining and analyses. This is about to change, with the onset of the ‘Open Access’ era:

Open Access.

Until now, the reference metadata and cited references of the top 25% of the c. 24,000 peer-reviewed journals published worldwide, across disciplines and languages, have been systematically fed (by the journal publishers) to the Institutute for Scientific Information (ISI), to be extracted and stored. But

soon this will change. It has been discovered (belatedly) that the Web makes it possible to make the full-text (not just the reference metadata and cited reference) of every single one of the 2.5 million articles published annually in those 24,000 journals (not just the top 25%) freely accessible online to all users (not just those that can afford paid access to the journals and the ISI database).

Open Access (OA) means free online access to all peer-reviewed journal articles. In the paper era, because of the cost of generating and disseminating print-on-paper, OA was unthinkable. Paper access could not be provided free for all, because subscription income was needed to cover the costs of peer review and publishing. Moreover, the paper medium did not make it possible to (literally) put the entire peer-reviewed journal corpus at the fingertips of all users, everywhere, at all times (in the way the Web already does today – but with other forms of content, such as E-bay product blurbs, blogs, and pornography).

Awareness of the Web's potential for providing OA to the research corpus is arriving belatedly for a number of reasons, but the chief two reasons are that for a long time most researchers neither realized (nor believed) (1) that OA was fully within their reach, nor (2) that it could bring them substantial benefits.

(1) Reachability of OA.

That OA was fully within researchers' reach had in fact already been demonstrated decades earlier, first by the computer scientists who invented Unix and the Internet, and immediately began storing and sharing their papers on 'anonymous FTP sites' (in the '80's, and then on websites, once the web was invented). Next, some branches of physics -- notably high energy physics (in which there had already been a culture, even in the paper era, of systematically sharing one another's prepublication preprints) -- made the natural transition to first sharing preprints via email and then via the web : In the early '90s, physicists began self-archiving their papers in electronic form ('eprints') -- both before (preprints) and after peer-review (postprints) -- in a central web archive (long called XXX and eventually Arxiv). In computer science, meanwhile, where self-archiving had been going on even longer, but on authors' local FTP and websites rather than centrally, a central harvester of those local websites, Citeseer, was created in the late '90s that not only gathered together all the computer science papers it could trawl from the web, but extracted and linked their reference lists, generating a rudimentary citation count for each article (Giles et al. 1999). Shortly thereafter, Citebase was created to do the same (and more) for the contents of Arxiv (Brody 2003). (We will return to Citebase, shortly.)

(2) Benefits of OA.

But apart from these two bursts of spontaneous self-archiving in computer science and some areas of physics, other disciplines did not pick up on the power and potential of the online medium for enhancing the usage of their work. Lawrence (one of the co-inventors of Citeseer) published a study in Nature in 2001, showing that articles that were made freely available on the Web were cited more than twice as much as those that were not ; yet most researchers still did not rush to self-archive. The finding of an OA citation impact advantage was soon extended beyond computer science, first to physics (Harnad & Brody 2004), and then also to all 10 of the biological, social science, and humanities disciplines so far tested (Hajjem et al 2005) ; yet the worldwide spontaneous self-archiving rate continued to hover around 15%.

If researchers themselves were not very heedful of the benefits of OA, however, their institutions and research funders – co-beneficiaries of their research impact – were: To my knowledge, the department of Electronics and Computer Science (ECS) at University of Southampton was the first to mandate self-archiving for all departmental research articles published: These had to be deposited in the department's own Institutional Repository (IR) (upgraded using the first free, open source software for creating OA IRs, likewise created at Southampton and now widely used worldwide).

Southampton was persuaded to do what it did in part because of something unique to the United Kingdom : The Research Assessment Exercise (RAE), in which research performance indicators from

every department in every UK university are evaluated and ranked every six years or so, with the departments receiving substantial top-sliced research funding as a function of their RAE rank.

The UK Research Assessment Exercise (RAE).

The RAE is a very cumbersome, time-consuming and expensive undertaking, for the researchers as well as the assessors. It requires preparing and submitting piles and piles of paper containing indicators of research productivity and performance (funding, students, publications, applications, industrial spin-offs, etc.) along with each researcher's four best papers, which are then 'peer-reviewed' by an RAE panel for each discipline. Of course, these papers are already published, hence have already been peer-reviewed; moreover, those that were published in the better journals had already been peer-reviewed by the best experts in the world in their field, not a generic RAE panel. But the interesting thing about the RAE outcome was that although (for no particular reason) it explicitly forbade citation counts among its performance indicators, the RAE rankings nevertheless turned out to be highly correlated with the total citation counts for the submitted researchers, and this was found to be true in every one of the RAE disciplines for which the correlation was tested (Oppenheim 1996).

The question then arises : If the RAE ranks that result from this complicated and time-consuming process are highly correlated with citations that are not even explicitly counted, why not just count the citations? If the correlation is high enough, the cumbersome process can be dropped, or at least simplified. Professor Charles Oppenheim of Loughborough University (who did most of the RAE/citation correlation studies), the University of Southampton research group that did most of the OA citation advantage studies, and many others in the UK, accordingly urged that the RAE should be dropped or simplified in favor of metrics, and it was subsequently decided to do so: The next RAE in 2008 will be a parallel exercise, conducted the old, complicated way, alongside a new, stream-lined, metrics-based way (Harnad 2006). But another complication has arisen:

In many of the hard-science disciplines there was a metric that correlated with the RAE outcome even more highly than citations: *prior research funding*. One could hence make an even stronger argument for basing the RAE rank entirely on prior research funding in those disciplines (where the correlation was close to 1.0). The problem, however, is that research funding *is* explicitly submitted as a performance indicator. So its tight correlation with the RAE outcome could well be because the RAE panels, in making their evaluations, were strongly influenced (indeed biased) by the amount of prior funding received – perhaps more influenced than by other factors, such as, for example, the quality of the submitted papers (which the panels were not necessarily even best-qualified to judge). Citations, in contrast, were not submitted or counted; so their correlation with the outcome, though not as high as that of prior funding, was ‘unbiased’. (Moreover, the RAE is meant to be an independent, top-sliced component of UK research funding, alongside classical competitive proposal-based funding, in a *dual* funding system. If prior funding were given too heavy a weight in the RAE ranking, that would merely amount to collapsing the dual funding system into a single proposal-based system, and just adding a multiplier effect to the proposal-based funding, amplifying any Matthew Effect, with the rich just getting richer and the poor poorer.)

There is a lesson to be learned here, and we will return to it, but first, back to the problem of the low spontaneous OA self-archiving rate worldwide, despite the evidence that it doubles citations. Part of the reasoning that drove Southampton's ECS department to adopt the world's first self-archiving mandate was that it would increase the visibility and usage of the department's research, thereby increasing its citation impact, and, in turn, the department's RAE rank. But there was also a lot of enthusiasm in the department for OA self-archiving, arising from studies based on citation-linking the articles deposited in the Physics Arxiv, through Citebase, the scientometric search engine created by Tim Brody, then a doctoral student in the department (Brody 2003; Harnad et al. 2003; Harnad & Brody 2004).

Citebase.

In conjunction with citation data from the ISI database (from a leased ISI CD-ROM), Citebase was used to compare the citation counts for articles in the same physics journals and years that were and

were not self-archived in Arxiv. Lawrence's OA citation advantage in computer science was confirmed in physics. Brody then went on to compare the download counts for articles in Arxiv within the first six months after they were deposited, and their citation counts a year or more later – and found a significant correlation : Early citation counts predict later citation counts, but download counts predict citations even earlier. A download/citation correlator was designed that allowed the user to set the time-window for calculating the correlation between downloads and citations (Brody et al 2006).

Citebase accordingly began adding more and more metrics on which it could rank the results of searches, including (for either the paper or the author): dates (age), citations, downloads, hub scores, authority scores, co-citedness scores. These rankings were of course for demonstration purposes only, and unvalidated. They were also very noisy, as authors' names were not uniquely sorted, the download data only came from the UK mirror site of Arxiv (as the primary US Arxiv site did not allow us to analyze their download statistics), not all articles in Arxiv had been successfully linked, and of course Arxiv's coverage of physics is not complete. Nevertheless, Citebase demonstrated the potential power of citation-based navigation as well as (univariate) scientometric ranking.

A Scientometric Ranking Engine for the RAE.

The natural next step – apart from increasing Citebase's coverage – is to add more of the ‘vertical’ metrics that currently allow univariate ranking, one variable at a time, but to turn them instead into ‘horizontal’ multivariate metrics, in a dynamic multiple-regression engine, in which each of the individual vertical predictors can be given a weight and combined in a multivariate linear equation. This can of course be done by simply picking one (or more) of the univariate metrics themselves, and using it as the criterion. That allows some rudimentary assignment of the beta weights on each remaining predictor. Or the weights can be hand-calibrable, so the user can explore the ranking yielded by different (normalized) hand-adjusted weights.

But the best solution of all is to provide an external criterion for validating the weights on the battery of predictor metrics. And there is, in the case of RAE 2008, a natural criterion, namely, the ranks generated by the parallel panel-based (‘peer review’) exercise itself. For all submitted authors (or papers), a Citebase-like multiple regression engine can be fed their publication counts, their articles’ journal impact factors, citation counts, co-citation counts, citation chronometrics (age, growth, latency to peak, decay rate), hub/authority scores, h-index, prior funding, student counts, co-authorship scores, endogamy/exogamy scores, textual proximity scores, download and co-download counts and chronometrics, and more. The multiple regression of all those predictor metrics onto the panel rankings as the validation criterion will generate the default beta weights (different in each discipline) for each metric (Harnad 2006). This will give an indication of the validity of the predictor metrics as a whole, relative to the panel-based rankings as the criterion. It will also allow adjustments, to calibrate the weights on the predictors so as to fine-tune the outcome or eliminate biases (e.g., to reduce or remove the Matthew Effect of prior funding; or there may be other factors to amplify, reduce, minimize, or eliminate). The battery could also be factor-analyzed to look for a smaller number of underlying factors – possibly even a ‘G-factor’, if there is one.

The UK RAE is merely one glimpse of the possibilities opened up by OA scientometrics (Shadbolt et al. 2006). There is no more reason for these metrics to remain inaccessible to all users in the online age than for the research articles themselves to remain inaccessible. The data-mining potential of an OA corpus is enormous, not just for research evaluation by performance assessors, but for search and navigation by researcher-users, students, and even the general public. Historians of knowledge, as well as analysts trying to predict the future course of knowledge will gain a wealth of powerful new tools to do so. The only thing missing is the primary OA database: the articles themselves. The main allies in demonstrating to the research community the benefits of providing that primary OA database today are the self-same scientometric tools that are waiting to mine it.

References

- Brody, T. (2003) Citebase Search: Autonomous Citation Database for e-Print Archives. *ECS Technical Report* <http://eprints.ecs.soton.ac.uk/10677/>

- Brody, T., Harnad, S. and Carr, L. (2006) Earlier Web Usage Statistics as Predictors of Later Citation Impact. *Journal of the American Association for Information Science and Technology (JASIST)* 57(8) pp. 1060-1072. <http://eprints.ecs.soton.ac.uk/10713/>
- Garfield, E., (1955) Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science* 122: 108-111. [http://www.garfield.library.upenn.edu/papers/science_v122\(3159\).p108y1955.html](http://www.garfield.library.upenn.edu/papers/science_v122(3159).p108y1955.html)
- Giles, C.L. , K. Bollacker, and S. Lawrence (1998) "CiteSeer: An Automatic Citation Indexing System," *Digital Libraries '98: Third ACM Conf. on Digital Libraries*, ACM Press, New York, 1998, pp. 89-98. <http://www.neci.nj.nec.com/homepages/lawrence/citeseer.html>
- Guédon, Jean-Claude, (2002) *In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing*. Washington, DC: The Association of Research Libraries <http://www.arl.org/resources/pubs/mmpceedings/138guedon.shtml>
- Hajjem, C., Harnad, S. and Gingras, Y. (2005) Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact. *IEEE Data Engineering Bulletin* 28(4) pp. 39-47. <http://eprints.ecs.soton.ac.uk/12906/>
- Harnad, S. (2006) Online, Continuous, Metrics-Based Research Assessment. Technical Report, ECS, University of Southampton. <http://eprints.ecs.soton.ac.uk/12130/>
- Harnad, S. & Brody, T. (2004) Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals, *D-Lib Magazine* 10 (6) June <http://www.dlib.org/dlib/june04/harnad/06harnad.html>
- Harnad, S., Carr, L., Brody, T. & Oppenheim, C. (2003) Mandated online RAE CVs Linked to University Eprint Archives: Improving the UK Research Assessment Exercise whilst making it cheaper and easier. *Ariadne* 35 (April 2003). <http://www.ariadne.ac.uk/issue35/harnad/>
- Hirsch, Jorge E., (2005), "An index to quantify an individual's scientific research output" *Proceedings of the National Academy of Sciences* 102(46) 16569-16572. <http://www.pnas.org/cgi/content/abstract/102/46/16569>
- Kleinberg, Jon, M. (1999) Hubs, Authorities, and Communities. *ACM Computing Surveys* 31(4) http://www.cs.brown.edu/memex/ACM_HypertextTestbed/papers/10.html
- Kline, Paul (2000) *The New Psychometrics: Science, Psychology and Measurement*. Routledge
- Lawrence, S. (2001) Online or Invisible? *Nature* 411 (6837): 521. <http://www.neci.nec.com/~lawrence/papers/online-nature01/>
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.
- Moed, H. F. (2005) *Citation Analysis in Research Evaluation*. NY Springer.
- Oppenheim, Charles (1996) Do citations count? Citation indexing and the research assessment exercise, *Serials*, 9:155-61, 1996. <http://uksg.metapress.com/index/5YCDB0M2K3XGAYA6.pdf>
- Page, L., Brin, S., Motwani, R., Winograd, T. (1999) *The PageRank Citation Ranking: Bringing Order to the Web*. <http://dbpubs.stanford.edu:8090/pub/1999-66>
- Shadbolt, N., Brody, T., Carr, L. and Harnad, S. (2006) The Open Research Web: A Preview of the Optimal and the Inevitable, in Jacobs, N., Eds. *Open Access: Key Strategic, Technical and Economic Aspects*, chapter 21. Chandos. <http://eprints.ecs.soton.ac.uk/12453/>

» FULL PAPERS AND RESEARCH-IN-PROGRESS PAPERS

ISSI 2007 - Madrid

Full-Text Publications in Peer-Reviewed Journals Derived from Presentations at Two ISSI Conferences

Rafael Aleixandre-Benavent*, Gregorio González-Alcaide*, Alberto Miguel-Dasit**, Carolina Navarro-Molina* and Juan Carlos Valderrama-Zurián*

**Rafael.Aleixandre@uv.es, gregorio.gonzalez@uv.es, canamo@uv.es, juan.valderrama@uv.es*
Instituto de Historia de la Ciencia y Documentación López Piñero, CSIC-Universidad de Valencia, (Spain)

** *amdasit@hotmail.com*
Unidad de Resonancia Magnética. Hospital de La Plana, Vila-Real, Castellón (Spain)

Abstract

The purpose of this study is to analyse the bibliometric characteristics of the full presentations and posters at the 5th and 8th Conferences of the International Society for Scientometrics and Informetrics (ISSI), held in 1995 and 2001 respectively, which were subsequently published in peer-reviewed journals covered by the Science Citation Index, Social Science Citation Index and LISA databases. 25.1% of all the papers presented at the two conferences were published, with the full papers of the 8th Conference attaining the highest percentage. Scientometrics was the journal that published the highest proportions. This relatively low overall rate of publication deprives the scientific community of potentially interesting results and points up the role of the ISSI Conference proceedings as a primary source of information.

Keywords

ISSI conferences; derived publications; publication rate

Background

Scientists commonly seek to exploit and enhance their reputations by taking part in conferences. For one thing, it is an investment for their research, since it allows them to extend their network of collaborators to include other peers; and for another, conferences are the ideal forum for discussing research that is in progress and new trends.

However, by no means all the information presented at conferences gets published in journals, and that makes it difficult for researchers to access it. Although many conference organisers publish proceedings, the content of these is often restricted to abstracts of the papers and posters, while the full texts remain unavailable. Hence, some researchers seek subsequently to publish an augmented version of their contributions and in that way ensure that their research reaches the widest possible audience.

Post-conference publication has been studied for various areas of science, among them library science (Fennewald, 2005), pharmacy (Byerly et al, 2000), marine biology (Bird & Bird, 1999) and several medical fields (Miguel-Dasit et al, 2006a).

The purpose of the present study is to perform a bibliometric analysis of publications in peer-review journals which are derived from the full papers and the posters presented at two conferences of the International Society for Scientometrics and Informetrics (ISSI): namely the 5th Conference, held at River Forest, Illinois, USA, in 1995, and the 8th Conference, held in Sydney, Australia, in 2001. The ISSI has been holding its biennial conferences all over the world since 1987 (in Belgium, Canada, India, Germany, USA, Israel, Mexico, Australia, China and Sweden), and there have been ten of them in the past 19 years. We chose 1995 and 2001 for two reasons: the need to leave a sufficient gap of time after the conferences for research in progress to be terminated and the results published; and on the other hand, the need to dispose of data that went back far enough to allow comparative analyses.

Method

All the data was drawn from the volumes *Fifth International Conference of the International Society for Scientometrics and Informetrics, Proceedings, 1995* (Koenig & Bookstein, 1995), and *Proceedings of the 8th Conference on Scientometrics and Informetrics* (Davis & Wilson, 2001).

In order to manage the information, a database was set up in Microsoft Access and the following data items were entered: author, title, author's affiliation, type of presentation (full paper or poster) and the year of the conference. After that, we set out to trace the subsequent publications in *Science Citation Index* (SCI) and *Social Science Citation Index* (SSCI), available through the *Web of Science*, and in *Library and Information Science Abstracts* (LISA), which is specific to information science and which we consulted through the *Cambridge Scientific Abstracts* (CSA) platform. The search technique used for retrieving the published articles consists basically in interrogating the databases with the surnames and initials of the lead authors in Boolean combination with keywords from the title. When this procedure produced no hits, the databases were interrogated again with the names of the co-authors (Callaham et al., 2002; Miguel-Dasit et al., 2006). In order to determine the correspondence between a conference presentation and the subsequent publication we follow these criteria: a) The authors' names were the same in both the article and a presentation; b) The author affiliations coincided; c) The title and keywords coincided.

When the attributes of an article coincided with those of a presentation, it was added to the database with its title, authors, journal of publication, the journal's impact factor as assessed in *Journal Citation Reports* (JCR), the quartile in which the journal was placed, subject area, countries of the institutions to which the authors were affiliated, collaboration index (mean number of authors or institutions per document) and types of collaboration (domestic versus international) (Glazebrook & Lange, 2002).

Findings

191 presentations were made at the two conferences analysed: 93 (49%) at the 5th Conference and 98 (51%) at the 8th Conference. At the 5th Conference, there were more full papers than posters (69% versus 31%); whereas at the 8th Conference, the percentages were more even (73% and 26% respectively) (Table 1). Out of all the presentations at both Conferences, 48 (25%) were later published and then picked up by the SCI, SSCI and LISA. 11 of these were from the 5th Conference (12% of the presentations at that conference) and 37 (38%) from the 8th Conference. As for the type of presentation, the full papers from the 8th Conference scored the highest percentage of publication (42% of the full papers presented). The lowest percentage was of the full papers from the 5th Conference (7.8% of the full papers presented).

Table 1. General data

	5th Conference	8th Conference	Total
<i>Full papers</i>	64 (68.8%)	72 (73.5%)	136 (71.2%)
<i>Full papers published</i>	5 (7.8%)	30 (41.7%)	35 (25.7%)
<i>Posters</i>	29 (31.2%)	26 (26.5%)	55 (28.8%)
<i>Posters published</i>	6 (20.7%)	7 (26.9%)	13 (23.6%)
Total presentations	93	98	191
Total published	11 (11.8%)	37 (37.7%)	48 (25.1%)

The distribution of the presentations and subsequent publications is given in Table 2 according to country of origin. They came from 22 different countries at the 5th Conference and 28 countries at the 8th Conference. The countries with most presentations at the 5th Conference were India (n=21), USA (n=14) and Spain (n=9); at the 8th Conference, they were India (n=25) and France (n=10). The countries with the greatest number of subsequent publications from the 5th Conference were Germany, which published 28.6% of its presentations, and the USA with 21%. After the 8th Conference, all the presentations from Belgium, Canada, Finland, Hungary, Iran, Israel, Japan, Norway and Taiwan were published. 15 of the 22 countries whose researchers presented work at the 5th Conference did not publish any of it; whereas this happened to only 7 of the 24 countries represented at the 8th Conference.

As for the impact factor of the derived articles, only 2 from the 5th Conference (18.2%) were published in journals in the top quartile of the JCR classification by subject field, whereas from the 8th

Conference there were 34 (77,3% of those published). The countries with the largest number of articles published in top quartile journals were the USA and India (n=4, respectively), Belgium, France and Netherlands (n=3, respectively).

Most of the derived papers were published the year following their respective Conferences (45.5% of the ones from the 5th Conference in 1996; 75.7% of those from the 8th Conference in 2002) (Table 3). Something striking is the speed with which some of the papers from the 5th Conference were published, because 4 of them (36.4%) appeared in the year of the Conference itself (1995). On the other hand, a few were published 4 or even 5 years later than their respective Conferences.

Table 2. Publication rates according to country of origin

Countries	5th Conference				8th Conference			
	FP	P	PA	AP-1 st Q	FP	P	PA	AP-1 st Q
Australia	5	-	-	-	6	2	3	-
Austria	-	-	-	-	1	-	-	-
Belarus	-	-	-	-	-	1	-	-
Belgium	1	2	1	-	3	-	3	3
Brazil	1	1	-	-	-	-	-	-
Bulgaria	-	1	-	-	-	-	-	-
Canada	1	-	-	-	2	1	3	2
China	-	3	1	-	3	3	2	1
Cuba	-	1	-	-	1	-	-	-
Denmark	1	-	-	-	2	-	1	1
Estonia	-	-	-	-	1	-	-	-
Finland	-	-	-	-	-	1	1	1
France	4	2	-	-	7	3	4	3
Germany	6	1	2	1	5	1	1	1
Hungary	2	-	-	-	2	-	2	2
India	12	9	1	-	14	11	5	4
Indonesia	-	-	-	-	1	-	-	-
Iran	-	-	-	-	1	-	1	-
Israel	2	-	-	-	3	-	3	2
Italy	-	1	-	-	-	-	-	-
Japan	-	-	-	-	2	-	2	2
Mexico	4	-	-	-	4	1	2	2
New Zeland	-	-	-	-	2	1	-	-
Norway	-	-	-	-	1	-	1	-
Poland	2	-	-	-	-	-	-	-
Russia	1	-	-	-	2	-	1	1
South Africa	-	-	-	-	1	1	-	-
Spain	7	2	1	-	-	2	1	1
Sweden	1	-	-	-	-	-	-	-
Taiwan	-	-	-	-	-	1	1	1
Netherlands	2	3	2	1	6	-	3	3
UK	3	-	-	-	3	-	-	-
USA	11	3	3	-	7	-	4	4
Total*	66	29	11	2	80	29	44	34

FP: Full papers; P: Poster; PA Published Articles; AP-1stQ: Articles in first quartile in the subject areas of JCR-ranked journals // * The sum total of the columns does not accord with Table 1 because some studies came from authors in more than one country.

Table 3. Publication rates according to year of publication of the articles

5th Conference				8th Conference			
Year	Total	FP	P	Year	Total	FP	P
1995	4 (36.4%)	2 (40%)	2 (33.3)	2001	2 (5.4%)	2 (6.7%)	-
1996	5 (45.5%)	2 (40%)	3 (50)	2002	28 (75.7%)	24 (80%)	4 (57.1%)
1997	-	-	-	2003	2 (5.4%)	2 (6.7%)	-
1998	1 (9.1%)	-	1 (16.7)	2004	2 (5.4%)	1 (3.3%)	1 (14.3%)
1999	1 (9.1%)	1 (20)	-	2005	2 (5.5%)	1 (3.3%)	1 (14.3%)
2000	-	-	-	2006	1 (2.7%)	-	1 (14.3%)
Total	11	5	6	Total	37	30	7

FP: Full papers; P: Posters

Table 4. Distribution of journals, countries of publication and country from which the abstract was submitted

Journal	Country of Edition	No. Articles				Country from which the abstract was submitted		
		5 th Conf.	IF (mean)	8 th Conf.	IF (mean)	5 th Conf.	8 th Conf.	
<i>Scientometrics (JCR)</i>	Netherlands Hungary	4	0.478	28	0.866	China India Netherlands USA	Belgium (2) Canada (2) China Denmark Finland France (4) Germany Hungary (2) India (5)	Israel (2) Japan Mexico (2) Netherlands (2) Norway Spain Taiwan USA (3)
<i>Information Processing & Management (JCR)</i>	USA	2	0.697	2	1.185	Germany	Russia	USA
<i>JASIS (JCR)</i>	USA	1	1.231	2	1.528	Spain	Canada	Japan
<i>Behavioral & Social Sciences Librarian (JCR)</i>	USA	1	0	-	-	USA		-
<i>Publishing Research Quarterly (JCR)</i>	USA	1	0.078	-	-	USA		-
<i>Research Policy (JCR)</i>	Netherlands	1	1162	-	-	Belgium Netherlands		-
<i>Serials Librarian</i>	USA	1	-	-	-	Belgium		-
<i>CJILS</i>	Canada	-	-	1	0.308	-	Australia	China
<i>Current Science</i>	India	-	-	1	0.533	-		India
<i>Libri</i>	Germany	-	-	1	0.123	-	Australia	Iran
<i>Journal of Information Science</i>	UK	-	-	1	1.080	-		Israel
<i>Research Evaluation</i>	UK	-	-	1	0.308	-		Germany
Total	-	11	-	37	-	-	-	-

JASIS: Journal of the American Society for Information Science

CJILS: Canadian Journal of Information and Library Science

JCR: journals included in Journal Citation Reports

Table 4 shows the distribution of the articles over the 10 journals in which they were published, as well as the country from which the abstract was submitted. The journal that published the most was Scientometrics (n=32, 66.7%), followed by Information Processing & Management (n=4, 8.3%) and the Journal of the American Society for Information Science (n=3, 6.2%). This last was the one with the highest average impact factor (FI=1.528). The other 9 journals only accounted for 1 article each.

Table 5 shows the distribution of the number of authors per paper. At the 5th Conference, 40% of the papers were presented by 1 sole author and 34.4% by 2; whereas at the 8th Conference, there were more dual than solo authors: 39.8% to 33.7%. Among the subsequent publications, articles by 2 authors also predominated (39.6%), followed by those with a single author (31.2%). The collaboration index was approximately 2 authors per study in most of the categories, the lowest figure being for the full papers at the 5th Conference (1.97) and the highest for the posters at the 8th Conference (2.35).

Table 5. Author collaboration patterns

Authors per paper	5 th Conference		8 th Conference		PA
	FP	P	FP	P	
1	26	11	28	5	15
2	23	9	27	12	19
3	10	7	8	5	8
4	1	2	4	3	2
5	4	0	3	1	3
6	0	0	2	0	1
Total	64	29	72	26	48
Col. Index	1,97	2	2,07	2,35	2,1

FP: Full papers; P: Posters; PA Published Articles

An analysis of the author affiliations (Table 6) shows that papers submitted from just one institution (no collaboration, or type 1) were in the majority: 78.5% of the presentations at the 5th Conference and 65.3% at the 8th Conference. The collaboration percentages at the 8th Conference were higher than at the 5th, as regards both domestic (26.5% versus 18.3%) and international (8.2% versus 3.2%) collaborations. As for the published articles, among those derived from the 5th Conference the domestic collaborations predominated (45.5%); whereas the largest percentage from the 8th Conference was of articles prepared without collaboration (56.8%). Approximately 17% of the published articles were the fruit of an international collaboration (type 3), which is a much higher percentage than for this type of collaboration among the papers given at the Conferences.

Discussion

The biennial meetings of the ISSI allow researchers in its area to disclose and exchange ideas. Nevertheless, it is publication in peer-reviewed journals and indexing in the science bibliographic databases that remain the ultimate validation of these research findings, because publication implies a more rigorous review of the design, methods and conclusions.

Some authors consider that a high percentage of published papers from a conference is an indication of its high quality (Miguel-Dasit et al., 2006a). However, publication customs differ from one field of science to another, and in some of them the conference proceedings are the ultimate stage and are even more cited than the journals (Aleixandre et al., 2004).

The overall percentage of derived publications from the ISSI Conferences was higher than from other areas such as library science, where Fennewald (2005) found only 13% of derived publications from the Ninth ACRL Conference (Association of College Research Libraries). In this regard, the ISSI

Conferences are in an intermediate position compared with others studied, although most of the other studies are not really comparable because they were drawn from the area of the health sciences. For example, the publication rates from radiological meetings range from 9% to 37% (Marx et al., 1999; Bydder et al., 2004; Miguel-Dasit et al., 2006b). In another medical field, urology, the percentage was 37.7% (Longena, 2004); whilst in paediatrics, emergency medicine and gynaecology, it was higher (51%, 62% and 75% respectively) (Riordan, 2000; Callaham et al., 2002; Gandhi & Gilbert, 2004). In a sample of pharmacy conferences, it ranged from 11% to 33% (Byerly et al., 2000).

Table 6. Distribution of abstracts and derived articles according institutional collaboration

Institutional Collaboration	5th Conference			8th Conference		
	FP / PA	P / PA	Total (FP+P) / Total PA	FP / PA	P / PA	Total (FP+P) / Total PA
Type 1	50 / 2	23 / 2	73 (78.5%) / 4 (36.3%)	47 / 17	17 / 4	64 (65,31%) / 21 (56,76%)
Type 2	11 / 2	6 / 3	17 (18.3%) / 5 (45.5%)	17 / 9	9 / 1	26 (26,53%) / 10 (27,02%)
Type 3	3 / 1	0 / 1	3 (3.2%) / 2 (18.2%)	8 / 4	0 / 2	8 (8,16%) / 6 (16,22%)
Total	64 / 5	29 / 6	93 / 11	72 / 30	26 / 7	98 / 37

FP: Full papers; P: Posters; PA Published Articles

Type 1: no collaboration (abstracts submitted from just one institution)

Type 2: abstracts in domestic collaboration

Type 3: abstracts in international collaboration

There are good reasons why not all the research presented at conferences is published as articles: there is insufficient space for it in the top journals, the selection process there is far more rigorous, and the fact is that much of the work presented at conferences is not original, relevant or of high enough quality to justify publication. The reasons why the conferences of some learned societies achieve higher rates of publication than others are unclear, and they are difficult to pin down because they are multifactorial. One reason, perhaps the main one, is certainly the quality of the papers; the peer-review process of the journals is extremely demanding. In addition, it seems intuitive that new findings are more likely to be published than confirmatory observations. Other reason, which is relevant to the general nature of ISSI, is the fact that most papers are quantitative or mathematical, where the author wants to introduce a new calculation or bibliometric method. In those cases, a precise conference paper could be sufficient enough. Some of the reasons cited for rejection after peer review are unsound methods, negative results, small sample size and the existence of other papers with similar findings (Krzyzanowska, 2003; Timmer, 2002; Riordan, 2000).

Another factor that weighs heavily in the potential for subsequent publication is the prior selection work done by the conference scientific committees. Hence, a higher rate of success in publishing in peer-reviewed journals is related to a high level of competitiveness between submitters and to a rigorous selection process. At the same time, it is possible that some of the presentations remain unpublished because they only report ‘work in progress’ that may carry on for more than another five years before it is finally publishable — or which is never completed (Riordan, 2000).

In a survey of authors who had presented papers at conferences but did not submit them to a journal afterwards, 47% of respondents stated that they did not have time to do the work of preparing a manuscript and 17% cited difficulties with their co-authors. Obviously, this latter kind of difficulty as the work progresses could be avoided by carefully defining the role of each co-author at the start (Sprague et al., 2003; Marx et al., 1999). Other reason mentioned by Weber (1998) and by Gandhi & Gilbert (2004) had to do with time constraints due to other responsibilities (professional obligations, teaching, administrative duties), and because the priority given to publishing the results declined progressively with the passage of time.

There are also more personal reasons why researchers did not submit articles. For instance, some researchers may have considered that their work was already reported in sufficient detail at the conference and that publication in the proceedings was enough, without need to re-publish it in a journal. And then there is the pessimistic attitude of some presenters regarding the likelihood of journal acceptance; they were under the impression that their work was interesting enough for a conference paper but that it was not of sufficiently high quality to pass the peer-review process for publication.

The subsequent rate of publication was quite similar in both oral (25.7%) and poster presentations (23.6%), in opposition to other studies that found in full papers a greater chance of getting published than in posters (Carroll et al., 2003; Juzych et al., 1993; Gandhi and Gilbert, 2004).

Derived publication usually occurs within two years of the Conference. The time period is similar to that found in one study (Yentis et al., 1993) but longer than in others (Marx et al., 1999; Nguyen et al., 1998; Juzych et al., 1999). This is not at all surprising, and indeed some authors are of the opinion that if at least an abstract of a presentation is not published by three years after a conference, its data should be viewed in the context of the multiple uncertainties that plague unpublished reports (Marx et al., 1999). The long delay that sometimes occurs between the meeting presentation and full-text publication may be due to a desire on the part of the authors to augment and improve it by revising its design, methodology and conclusions before submitting it to an international peer-reviewed journal of the highest quality and impact factor (Miguel-Dasit et al., 2006a).

The present study has some limitations that may possibly have biased the results:

1. We have only analysed two of the ten ISSI Conferences held so far. The behaviour of our variables at other Conferences therefore remains unknown.
2. A second limitation concerns the comprehensiveness of the SCI, SSCI and LISA databases. Although they have international coverage and one of them is specific to information science, they may possibly have missed some articles published in domestic journals they do not index. Articles published in journals outside the scope of the SCI, SSCI and LISA databases were treated in our analysis as unpublished.
3. It may be that some presentations had undergone important changes to their title, authorship or abstract which made them unrecognisable in the databases.

Conclusions

To sum up, a quarter of the presentations at the two ISSI biennial meetings were ultimately published in peer-reviewed journals and cited in international databases. The journal that published the most was Scientometrics. The low overall rate of publication can deprive the scientific community of potentially interesting results, and also prevents them from being indexed in bibliographic databases. However, if the ISSI is considered to be a very strong community, which is made up of a close-nit number of academics, then most will attend the conference and know what is being done in the field regardless of subsequent publication rates. We would therefore encourage all researchers whose abstracts are accepted for presentation to ISSI meetings to complete and submit a manuscript for journal publication, because the peer-review selection process of the journals identifies research as being worthy of publication and therefore important. On the other hand, the low percentage of journal publications shows how important is the function of the ISSI Conference Proceedings as a primary source of information, and the need not to overlook them when searching for exhaustive information in the field of informetrics and scientometrics.

There are still unresolved questions that might well be addressed in further studies. It would be of interest, for example, to analyse the relationship between certain attributes of conference papers and their later publication: for example originality, soundness of method and positive versus negative results. One would also like to know more about the criteria that are used in selecting the papers for the ISSI Conferences.

References

- Aleixandre, R., Valderrama, J.C., Desantes, J.M., Torregrosa, A.J. (2004). Identification of information sources and citation patterns in the field of reciprocating internal combustion engines. *Scientometrics*, 59 : 321-336.
- Bird, J.E., Bird, M.D. (1999). Do peer-reviewed journal papers result from meeting abstracts of the Biennial Conference on the biology of Marine Mammals?. *Scientometrics*, 46 : 287-297.
- Bydder, S.A., Joseph, D.J., Spry, N.A. (2004). Publication rates of abstracts presented at annual scientific meetings: How does the Royal Australian and New Zealand College of Radiologists compare? *Australasian Radiology*, 48 : 25-28.
- Byerly, W.G. Rheney, C.C., Connelly, J.F., Verzino, K.C. (2000). Publication rates of abstracts from two pharmacy meetings. *Annals of Pharmacotherapy*, 34 (10) : 1123-1127.
- Callaham, M., Wears, R.L., Weber, E. (2002). Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA*, 287 : 2847-2850.
- Carroll, A.E., Sox, C.M., Tarini, B.A., Ringold, S., Christakis, D.A. (2003). Does presentation at the pediatric Academic Societies' annual meetings predict subsequent publication?, *Pediatrics*, 112 : 1238-1241.
- Davis, M., Wilson, C.S. (eds) (2001). Proceedings of the 8th Conference on Scientometrics and Informetrics. Sydney: Bibliometric and Informetric Research Group. University of New South Wales.
- Fennewald, J. (2005). Perished or published: the fate of presentations from the Ninth ACRL Conference. *College & Research Libraries*, 66 (6) : 517-525.
- Gandhi, S.G., Gilbert, W.M. (2004). Society of Gynecologic Investigation. What gets published?. *Journal of the Society of Gynecologic Investigation*, 11 : 526-565.
- Glanzel, W.; de Lange C. (2002). A distributional approach to multinationality measures of international scientific collaboration. *Scientometrics*, 54 (1): 75-89.
- Juzych, M.S., Shin, D.H., Coffey, J., Juzych, L., Shin, D. (1993). Whatever happened to abstracts from different sections of the association for research in vision and ophthalmology? *Investigative Ophthalmology and Visual Science*, 34 : 879-382.
- Koenig, M.E.D., Bookstein, A. (eds) (1995). Fifth international conference of the international society for scientometrics and informetrics: Proceedings 1995. Illinois: Medford, NJ, Learned Information.
- Krzyzanowska, M.K., Pintilie, M., Tannock, I.F. (2003). Factors associated with failure to publish large randomised trials presented at an oncology meeting. *JAMA*, 290 : 495-501.
- Marx, W.F., Cloft, H.J., Do, H.M., Kallmes, D.F. (1999). The fate of neuroradiologic abstracts presented at national meetings in 1993: rate of subsequent publication in peer-reviewed, indexed journals, *American Journal of Neuroradiology*, 20 : 1173-1177.
- Miguel-Dasit, A., Martí-Bonmatí, L., Aleixandre, R., Sanfeliu, P., Valderrama, JC. (2006a). Publications resulting from spanish radiology meeting abstracts : Which, Where and Who. *Scientometrics*, 66 (3) : 467-480.
- Miguel-Dasit, A., Martí-Bonmatí, L., Sanfeliu, P., Aleixandre, R. (2006b). Scientific papers presented at the European Congress of Radiology 2000: publication rates and characteristics during the period 2000-2004. *European Radiology*, 16 (2) : 445-450.
- Nguyen, V., Tornetta, P., Bkaric, M. (1998). Publication rates for the scientific sessions of the Orthopaedic Trauma Association (OTA). *Journal of Orthopedic Traumatology*, 12 : 457-459.
- Riordan, F.A.I. (2000). Do presenters to paediatric meetings get their work published? *Arch Dis Child*, 83 : 524-526.
- Sprague, S., Bhandari, M., Devereaux, P.J., Swiontkowski, M.F., Tornetta, P., Cook, D.J., Dirschl, D., Schemitsch, E.H., Guyatt, G.H. (2003). Barriers to full-text publication following presentation of abstracts at annual orthopaedic meetings, *Journal of Bone and Joint Surgery. American Volume*, 85 : 158-163.
- Timmer, A., Hildsen, R.J., Cole, J., Hailey, D., Sutherland, L.R. (2002). Publication bias in gastroenterological research – a retrospective cohort study based on abstracts submitted to a scientific meeting. *BMC Medical Research Methodology*, 2 : 7.
- Von Elm, E., Costanza, M.C., Walder, B., Tramer, M.R. (2003). More insight into the fate of biomedical meeting abstracts: a systematic review, *BMC Medical Research Methodology*, 3 : 12.
- Weber, E.J., Callaham, M.L., Wears, R.L., Barton, C., Young, G. (1998). Unpublished research from a medical specialty meeting: why investigators fail to publish. *JAMA*, 280 : 257-259.
- Yentis, S.M., Campbell, F.A., Lerman, J. (1993). Publication of abstracts presented at anaesthesia meetings. *Canadian Journal of Anaesthesia*, 40 : 632-634.

Origins of Measures of Journal Impact: Historical Contingencies and their Consequences on Current Use¹

Éric Archambault* and Vincent Larivière**

*eric.archambault@science-metrix.com

Science-Metrix, 4572 Avenue de Lorimier, Montréal, Québec H2H 2B5 (Canada) and
Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, Montréal, Québec (Canada)

**lariviere.vincent@uqam.ca

Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, CP 8888, Succursale Centre-ville, Montréal, Québec, H3C 3P8 (Canada) and

Graduate School of Library and Information Studies, McGill University, Montréal, Québec (Canada)

Abstract

This paper examines the genesis of journal impact measures and how their evolution culminated in the journal impact factor (JIF) produced by the Institute for Scientific Information. The paper shows how the form of the JIF, which is the result of historically contingent choices rather than a carefully chosen and tested set of features, affected its subsequent use, misuse, and manipulation by researchers, journal editors, and bibliometrists.

Keywords

Journal Impact Factor; history; contingencies.

Introduction

In the last three decades, librarians and bibliometrists have progressively come to rely on the journal impact factor (JIF). Particularly in the late 1990s, the indicator attracted a significant amount of attention in the scientific community. Many researchers have observed that this indicator is orienting the publishing strategies of scientists who want to maximize their impact factor and how, similarly, journal editors aspire to augment their journal's JIF. Consequently, bibliometrists increasingly try to "tame the beast" by suggesting numerous ways to improve the validity of the JIF as a quantitative measure. This growing interest is illustrated by an increase in the number of papers dealing with the indicator, rising from 23 papers indexed in Thomson Scientific's *Web of Science* in 1995 to 146 papers in 2005. Despite this growing interest, there is, apart from Eugene Garfield's own historical accounts, a real scarcity of contributions to the conceptual history of this important indicator. This paper provides an account of the history of the JIF and its subsequent use, misuse, and manipulation by researchers, journal editors, and bibliometrists. It would be beyond the purposes of this paper to attempt to capture every minute characteristic of this indicator as well as its origins and its effects on the evolution of the bibliometric field. Here, we will concentrate on five aspects that have received the largest share of interest: the fact that the indicator was developed to help the management of scientific journal collection, not the evaluation of scientific research; the field-specific nature of the scores produced by the indicator (scores are not readily comparable across scientific disciplines); the asymmetry between what is counted in the numerator and in the denominator; the two-year citation window; and how the scores are English language- and US-centred.

Origins of Measures of Journal Impact

The literature on the use of journal impact measures uniformly concludes that Gross and Gross (1927) were the first to develop this method (see, e.g., Allen, 1929; McNeely and Crosno, 1930; Gross and Woodford, 1931; Henkle, 1938; Brodman, 1944; Garfield, 1955; and Raisig, 1960). Gross and Gross sought to address the rising problems of small colleges at a time when one "of the biggest of these is the problem of adequate library facility." It is important to note that the first use of journal impact

¹ The authors wish to thank Jean-Pierre Robitaille and the two anonymous reviewers for their valuable comments and suggestions. They also wish to thank Julie Caruso and Johanna Kratz for their revisions of the manuscript.

calculation aimed to facilitate the task of journal selection, which is one core aspect in the marketing of the most visible commercial product that has emerged from this work—Thomson Scientific's Journal Citation Report (JCR).

The paper by Gross and Gross (1927) proved to be an inspiration for several US librarians and early information scientists. For instance, Brodman (1944) cites no less than 18 papers published after 1926 that used a method based on the Gross and Gross paper. It is not surprising that such an explosion occurred. During that period, the number of periodicals available to libraries was growing at an exponential rate, and the Great Depression was taking its toll on the budget of university libraries. Importantly, the problem is just as prominent today. This quote from Cunningham (1935) has all the elements of a librarian's pamphlet from the present day:

The tremendous number of journals being published and the continued increase in the cost of yearly subscriptions have made it increasingly difficult for libraries to maintain adequate subscription lists. At the same time, libraries have been facing a marked decrease in budgets, gifts and other forms of financial support.

What is immediately obvious when one examines the evolution of journal impact calculations that followed the method pioneered by Gross and Gross is the growing complexity and size of the compilations. In the early years of the method's development, studies were generally limited to a single field (e.g., chemistry, geology, or medicine) and references were often compiled from a single key journal (e.g., Allen, 1929) or key reference monograph (e.g., Hackh, 1936). The Gross and Gross method grew in size and complexity as it was adopted by other researchers. For instance, the Gross and Gross study carried out in 1927 used a single source journal and comprised a compilation of 3,633 references to 247 journals. By 1930, the method had gained two additional characteristics: several journals were used as sources and, although the practice was still predominantly centred on the English language, many non-US source journals were included. For example, McNeely and Crosno (1930) used seven source journals—"three American, one English, two German publications, and one French publication" (p. 82)—and compiled a total of 17,991 references. In a similar manner, Gregory (1937) produced a colossal study, considering the technical means available at the time, using the Gross and Gross method to identify key journals in 27 fields relevant to medicine and tabulating some 26,700 references from about 40 source journals or monographs. In 1956, Brown published a monograph entitled *Scientific Serials*, basing the approach, by then generalized, on collecting citations from several journals. Brown covered eight fields of science using 57 source journals and a compilation of close to 38,000 references.

Importantly, all of the aforementioned studies produced field-specific listings, and no author saw a need to adapt the method to enable comparisons across the field. There was, in fact, no need for such cross-field comparisons, since the purpose of the technique was to identify relevant journals for different fields. This characteristic produces some adverse effects when measures of journal impact are used to evaluate scientific production across fields; however, these difficulties could not have been foreseen because, at the time, this type of use was not driving the method's development.

In 1936, Hackh proposed the idea of dividing the number of references by the number of volumes, thus, for the first time, taking into account the extent of the citable material. However, this idea, with its added layer of complexity, was not taken up until 1960, when it reappeared in the work of Raisig (1960). By and large, the approach suggested by Raisig involved taking into consideration the "relationship of the number of articles quoted to the number of articles published," which was coined the RPR index or "index of research potential realized" (Raisig, 1960, p. 1418). Raisig's suggestion to use a ratio of citations to source articles was subsequently adopted by Garfield and Sher (1963b) for the calculation of the "journal impact factor". According to Garfield and Sher (1963b), their methodological changes to the existing literature involved the inclusion of multiple citations, in contrast to Raisig's (1960) suggestion, as well as self-citations, thus diverging from Westbrook's (1960) recommended method. Thus, Garfield and Sher's JIF was not a creation *ex nihilo*; it was

essentially a massive scale-up of existing techniques, permitted by the construction of the Science Citation Index in 1961 at the Institute for Scientific Information (ISI).

Interestingly, there is another important characteristic that appeared in Raisig (1960) that was eventually adopted in the construction of the impact factor commercialized by ISI in the 1970s. That characteristic was an asymmetry between the items that were considered valid counts for the numerator and for the denominator. Indeed, Raisig (*Ibid*) mentions that “[e]xcluded from the counts of original articles were letters, review articles, reports of patents, book reviews, abstracts, and purely biographical material.” Similarly, a few years later (but possibly without prior knowledge of Raisig’s approach), Martyn and Gilchrist (1967) also decided to exclude some source items, such as abstracts, obituaries, reviews, and bibliographies, from the counts. Clearly, the asymmetry between what is counted in the numerator (references to every type of material) and what is counted in the denominator (only the types of document that are deemed citable) predates ISI’s impact factor. The fact that Raisig, as well as Martyn and Gilchrist, were cited by Garfield in 1972, a turning point in the development of the ISI impact factor, strongly suggests that this characteristic was adopted from the prior art rather than invented by Garfield and his colleagues at ISI.

Another important attribute of ISI’s impact factor is the controversial two-year citation window that was developed by Martyn and Gilchrist (1967), who clearly exposed this characteristic when they wrote that “68,764 citations were made in 1965 of British items published in 1963 and 1964 to a total of 28,949 items” (p. vii). In a later study, Garfield (1972) mentioned his use of Martyn and Gilchrist’s method:

To calculate an impact factor for each journal, I divided the number of times 1967 and 1968 articles were cited in 1969 by the number of articles published in 1967 and 1968. Martyn and Gilchrist used a similar method in ranking British journals in an analysis of 1965 SCI data. (p. 476)

One can surmise that the use of this approach, based on considering only the previous two years of publications, was largely the result of an accidental choice rather than the result of an in-depth analysis of various solutions and the subsequent choice of optimal characteristics. Indeed, Garfield was aware that, overall, the vast majority of citations were older than two years. In 1963, he wrote that “over 50% of the cited references in the 1961 index are more than five years old” (Garfield and Sher, 1963a). This statement indicated that Garfield was aware that the half-life of the cited references was greater than five years and that going back two years certainly meant missing out on a very substantial part of the impact picture.

The evidence suggests that Martyn and Gilchrist are the creators of the impact factor as we know it. When Garfield (1972) adopted their method and furthered the work undertaken by librarians and information scientists over the course of the previous 40 years, work that had aimed to provide a way of acquiring adequate journals for libraries, his approach actually adopted most of the characteristics that had progressively taken shape in prior work. Like most of his predecessors, whose work had aimed to serve US scientific library users rather than to unambiguously determine the way the scientific system worked at the world level, a large number of Garfield’s source journals were no doubt US- and English-centric. Had this method been developed in a different country and had Garfield been, for instance, German, he would have started by using a vast majority of German-language journals as source items. This would have resulted in journal impact values that were different from those he obtained, as well as in the progressive inclusion of a set of journals that would most likely have resulted in a different source database than the one commercialized today by Thomson Scientific, and which incidentally serves to calculate today’s impact factor. Had this been the case, the impact factor values presented in the JCR would surely be substantially different.

Consequently, these measures cannot be considered objective measures of the worth of all journals published internationally. The JCR and its journal impact factor measures are the result of historically contingent events, and it is very important to consider the dire impact of these contingencies: the English-language and US bias, the presence of an asymmetrical numerator and denominator, and a

seemingly accidental two-year citation window are all characteristics that have a deep effect on the way research, journals, and even scientists are evaluated around the world today.

Use and misuse of measures of journal impact, and potential remedies

For the sake of convenience, one can categorise objections to the use of journal impact measures into three groups: 1) scientific activities should not be evaluated using bibliometric methods, particularly the impact factor; 2) if left in the wrong hands, indicators can easily be misused, but there are relatively simple normalizations that can, in general, make their use safer; 3) these indicators are technically flawed, but can be re-engineered in depth, and their flaws can be corrected.

The first type of objection arises mostly from epistemological reasoning. Examples of arguments used to unequivocally reject the use of citation analyses and journal impact measures include:

- Some scientific works are only recognised several years after their publication, while any citation analysis is limited to a predetermined citation window (Lindsey, 1989).
- Papers that are never cited do not necessarily have zero impact (Seglen, 1997).
- Negative citations are counted the same way as positive citations (Ophof, 1997).

Although bibliometrists will generally recognize some, if not all, of these limits, they will usually counter such arguments by stating that the strength of their indicators is conferred by the law of large numbers and that this is linked with the levels of aggregation. Glänzel and Moed (2002) distinguish between three levels of aggregation: 1) the micro level (individual scientist, research group); 2) the meso level (institutions, journals); and 3) the macro level (national and supra-national research, subject analyses). As noted by Seglen (1997), “[s]ince any large, random sample of journal articles will correlate well with the corresponding average of the JIF, the impact factor may seem reasonably representative. However, the correlation between journal impact and actual citation rate of articles from individual scientists or research groups is often poor”. For macro-level analyses, several of the weaknesses of the JIF tend to disappear, but there is at least one aspect that nearly everyone who carefully uses these indicators will say: they need to be modified somewhat to take into account the inter-field variations.

The second type of objection is that measures of journal impact are prone to be manipulated, misused, and even abused. The analysis of the genesis of journal impact measures in the first part of this paper made it apparent that these indicators were developed with a clear intent: to support the work of librarians in the management of their journal collections. Likewise, Thomson Scientific has a clear message about the intended uses of the Journal Citation Report²:

Enables a variety of information professionals to access and assess key journal data:

- *Librarians can manage and maintain journal collections and budget for subscriptions [...].*
- *Publishers can monitor their competitors, identify new publishing opportunities, and make decisions regarding current publications.*
- *Editors can assess the effectiveness of editorial policies and objectives and track the standing of their journals.*
- *Authors can identify journals in which to publish, confirm the status of journals in which they have published, and identify journals relevant to their research.*
- *Information Analysts can track bibliometric trends, study the sociology of scholarly and technical publications, and study citation patterns within and between disciplines.*

Aside from these intended uses, the JIF often serves in research evaluation. One of the most blatant abuses of this tool involves giving bonuses or making promotion decisions for researchers based on raw impact factor values. For example, Fuyono and Cyranoski (2006) mention that, in Pakistan, researchers can earn bonuses amounting to anywhere between \$1,000 and \$20,000 based on the cumulative one-year impact factor of their publications. The authors also provide the example of the Chinese Academy of Sciences' Institute of Biophysics, which has a scale tuned to the impact factor:

² <http://scientific.thomson.com/products/jcr/>

publications in journals with a JIF between 3 and 5 are worth 2,000 yuan per JIF point, and a publication in a journal with a score higher than 10 is worth 7,000 yuan per JIF point. For anyone who has worked with or read about the impact factor, it is a well-known fact that JIF scores vary tremendously between fields. To careful users of the JIF, it becomes clear that schemes such as these are helping researchers in the biomedical field become wealthier (because these fields have high citation rates and therefore high non-normalized impact factor values), while others, such as mathematicians or social scientists, are obtaining only small bonuses (because of the lower citation propensity in these fields), even if they manage to publish in the best journals in their respective fields. A simple normalization by scientific domain would create a more level playing field and certainly a fairer and more informed reward system. For example, Garfield suggested that:

Instead of directly comparing the citation count of, say, a mathematician against that of a biochemist, both should be ranked with their peers, and the comparison should be made between rankings. Using this method, a mathematician who ranked in the 70 percentile group of mathematicians would have an edge over a biochemist who ranked in the 40 percentile group of biochemists, even if the biochemist's citation count was higher³ (Garfield, 1979, as cited in Schubert and Braun, 1996, p. 312).

There are a very large number of papers with suggestions on how to normalize for differences across fields⁴. For many bibliometrists, this type of correction is mainly useful for performing studies at more macro levels. For instance, Seglen (1992) argues that as long as corrections are made to account for differences across fields, “citedness can be a useful indicator of scientific impact at the national level”. Even then, some bibliometrists would still advise against this type of usage, given the numerous flaws of the currently dominant JIF.

The third type of objection is primarily technically based and is in very large part aimed at the specific incarnation of the JIF commercialized by Thomson Scientific. Unlike the first type of objection, it traditionally came from researchers within the field of bibliometrics but, as can be seen in the editorials of biomedical and clinical research journals, a large number of researchers have something, mostly negative, to say about the limits of this indicator.

An item that is repeatedly criticized is the two-year citation window used in the JCR. The calculation of the impact factor based on a citation window of only two years is far too short in many fields. Glänzel and Moed (2002) cite the example of the comparison between the impact of *The Lancet* and the *American Sociological Review* (ASR). When a short citation window is used, *The Lancet* has a greater mean citation rate, but when using a window of four years or more, it is the ASR that has a higher mean citation rate.

Among the technical limits often cited, the asymmetry between the numerator and denominator and journal self-citations are among the most commonly mentioned. This asymmetry induces some strong distortions, more particularly for highly cited journals (Moed and Van Leeuwen, 1995). This can also lead to manipulation on the part of editors who multiply source items that are not considered “citable” but are in fact cited frequently. As a matter of fact, in several fields, the journals with the highest JIF are review journals, simply because review articles are often more cited than regular articles. This also presents an open invitation to distort the JIF scores by simply increasing the number of reviews. Another way editors can manipulate the JIF is by inducing authors to cite the journal in which they publish, and because journal self-citations are counted, it is also possible to influence the JIF by “encouraging” authors to cite papers from the journal in which they seek to be published. For instance, this letter was sent to authors wishing to publish in Leukemia (Smith, 1997):

Manuscripts that have been published in Leukemia are too frequently ignored in the reference list of newly submitted manuscripts, even though they may be extremely relevant. As we all know, the scientific community can suffer from selective memory when giving credit to colleagues. While we have little power over other

³ Following on this idea, Pudovkin and Garfield (2004) suggested using a rank-normalized impact factor using percentiles.

⁴ See e.g. Fassoulaki et al. (2002), Huth (2001), Ramirez, Garcia and Del Rio (2000), Schwartz and Lopez Hellin (1996), Sen and Shailendra (1992), Sombatsompop et al. (2005) and van Leeuwen and Moed (2001) to name only a few.

journals, we can at least start by giving you and others proper credit in Leukemia. We have noticed that you cite Leukemia [once in 42 references]. Consequently, we kindly ask you to add references of articles published in Leukemia to your present article. (p. 463)

This last form of objection presents a greater problem, because these shortcomings can often be solved only by having access to all of the source data, and very few bibliometrists have that kind of access. The few teams that have access to source data can certainly produce corrected measures of impact factors and use these corrected measures for their own research and contractual undertakings, but they would not be authorized to commercialize these indicators and compete against the JCR.

Conclusion

The JCR, which is the most well known descendant of the work undertaken by Gross and Gross in the 1920s, continues to cater to the need of university librarians who have to carefully select the most relevant journals for their clientele within limited budgets. Since the original work of Gross and Gross, measures of journal impact have grown in complexity and in size. There is no doubt that even though the presence of many of the characteristics of the JCR (such as the presence of an asymmetrical numerator and denominator, the use of a two-year citation window and the prevalence of English-speaking journals) are justified with the help of rationale arguments, they are at least in large part the result of historical contingencies. For instance, had the *Institute for Scientific Information* emerged as the “Institut für Forschungsinformation”, the JCR would undoubtedly have evolved in a substantially different form.

The JCR and its measures of journal impact have some significant shortcomings, but these clearly have different levels of consequence, depending on the use that is made of measures of journal impact. For the use intended by Thomson Scientific, such as selecting journals for a library, the weaknesses are certainly acceptable. When used for policy making at the national level, it becomes important to normalize by field to obtain an adequate picture. As one goes down the scale of applications, it becomes absolutely imperative to normalize data, and the deficiencies of the impact factor become increasingly worrisome, for the laws of large numbers decreasingly come into play to compensate for the shortcomings in the way JCR metrics are computed. While some improvements (e.g., field normalization), can be made “in-house” without a huge infrastructure, most other improvements can only be made by having access to the source data, which is not the case for most users, especially those outside of the bibliometric community. At the other extreme, the US- and English-language-centeredness may not be correctable through the use of Thomson Scientific databases, so it is not likely that the debate on the limits of these tools will cease anytime soon.

References

- Allen, E. S. (1929) Periodicals for mathematicians. *Science*, 70(1825), 592-594.
- Brodman, E. (1944) Choosing physiology journals. *Bull Med Libr Assoc*, 32(4), 479-483.
- Brown, C. H. (1956) Scientific serials: characteristics and lists of most cited publications in mathematics, physics, chemistry, geology, physiology, botany, zoology, and entomology. *ACRL Monograph no. 16*. Chicago: Association of College and Research Libraries.
- Cunningham, E. R. (1935) The present status of the publication of literature in the medical and biological sciences. *Bull Med Libr Assoc.*, 24(1), 64-81.
- Fassoulaki A., Papilas K., Paraskeva A. & Patris K (2002) Impact factor bias and proposed adjustments for its determination. *Acta Anaesthesiologica Scandinavica*, 46 (7), 902-905.
- Fuyuno, I., & Cyranoski, D. (2006) Cash for papers: Putting a premium on publication. *Nature*, 441(7095), 792.
- Garfield, E. (1955) Citation Indexes for Science. *Science*, 122(3159), 108-111.
- Garfield, E. (1972) Citation analysis as a tool in journal evaluation. *Science*, 178(4060), 471-479.
- Garfield, E. (2006) The history and meaning of the journal impact factor. *JAMA-Journal of the American Medical Association*, 295(1), 90-93.
- Garfield, E., & Sher, I. H. (1963a) *Genetics Citation Index*. Philadelphia: Institute for Scientific Information.
- Garfield, E., & Sher, I. H. (1963b) New factors in evaluation of scientific literature through citation indexing. *American Documentation*, 14(3), 195-201.
- Glänzel, W., & Moed, H. F. (2002) Journal impact measures in bibliometric research. *Scientometrics*, 53(2), 171-193.
- Gregory, J. (1937) An evaluation of medical periodicals. *Bull Med Libr Assoc.*, 25(3), 172-188.

- Gross, P. L. K., & Gross, E. M. (1927) College libraries and chemical education. *Science*, 66(1713), 385-389.
- Gross, P. L. K., & Woodford, A. O. (1931) Serial literature used by American geologists. *Science*, 73(1903), 660-664.
- Hackh, I. (1936) The periodicals useful in the dental library. *Bull Med Libr Assoc.*, 25(1-2), 109-112.
- Henkle, H. H. (1938) The periodical literature of biochemistry. *Bull Med Libr Assoc.*, 27(2), 139-147.
- Huth, E. J. (2001) Authors, editors, policy makers, and the impact factor. *Croatian Medical Journal*, 42(1), 14-17.
- Lindsey, D. (1989) Using citation counts as a measure of quality in science: measuring what's measurable rather than what's valid. *Scientometrics*, 15(3-4), 189-203.
- Martyn, J., & Gilchrist, A. (1968) *An evaluation of British Scientific Journals* (1 ed.): Aslib.
- McNeely, J. K., & Crosno, C. D. (1930) Periodicals for electrical engineers. *Science*, 72(1856), 81-84.
- Moed, H. F., & Van Leeuwen, T. N. (1995) Improving the accuracy of Institute for Scientific Informations journal impact factors. *Journal of the American Society for Information Science*, 46(6), 461-467.
- Ophof, T. (1997) Sense and nonsense about the impact factor. *Cardiovascular Research*, 33(1), 1-7.
- Pudovkin, A. I. & Garfield, E. (2004) Rank-normalized impact factor: A way to compare journal performance across subject categories. *Proceedings of the 67th ASIS&T Annual Meeting*, 41, 507-515.
- Raisig, L. M. (1960) Mathematical evaluation of the scientific serial. *Science*, 131(3411), 1417-1419.
- Ramirez, A. M., Garcia, E. O., Del Rio, J. A. (2000) Renormalized impact factor. *Scientometrics* 47(1), 3-9.
- Schubert, A., & Braun, T. (1996) Cross-field normalization of scientometric indicators. *Scientometrics*, 36(3), 311-324.
- Schwartz S & Hellin J. L. (1996) Measuring the impact of scientific publications. The case of the biomedical sciences. *Scientometrics*, 35(1), 119-132
- Sen B. K. & Shailendra, K. (1992) Evaluation of recent scientific research output by a bibliometric method. *Scientometrics*, 23(1), 31-46.
- Seglen, P. O. (1992) The skewness of science. *Journal of the American Society for Information Science*, 43(9), 628-638.
- Seglen, P. O. (1997) Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314, 497.
- Smith, R. (1997) Journal accused of manipulating impact factor. *British Medical Journal*, 314(7079), 463.
- Sombatsompop, N., Markpin, T., Yochai, W. & Saechiew, M. (2005) An evaluation of research performance for different subject categories using Impact Factor Point Average (IFPA) index: Thailand case study. *Scientometrics*, 65(3), 293-305.
- van Leeuwen, T. N. & Moed, H. F. (2001) Development and application of new journal impact measures. *Cortex*, 37(4), 607-610.
- Westbrook, J. H. (1960) Identifying Significant Research. *Science*, 132(3435), 1229-1234.

The Lifespan of “Informetrics” on the Web: an Eight Year Study (1998-2006)

Judit Bar-Ilan^{*} and Bluma C. Peritz^{**}

^{*}*barilaj@mail.biu.ac.il*

Department of Information Science, Bar-Ilan University, Ramat Gan, 52900 (Israel)

^{**}*bluer@cc.huji.ac.il*

Hebrew University of Jerusalem, Jerusalem, 91904 (Israel)

Abstract

The World Wide Web is growing at an enormous speed, and has become an indispensable source for information and research. New pages are constantly added, but there are additional processes as well: pages are moved or removed and/or their content changes. We report here the results of an eight year long project started in 1998, when multiple search engines were used to identify a set of pages containing the term *informetrics*. Data collection was repeated once a year for the last eight years (with the exception of 2000 and 2001) using both search engines and revisiting previously identified pages. The results show that the number of pages grew from 866 in 1998 to 28,914 in 2006 – a 33-fold growth. Besides the obvious growth of the topic on the Web, we observed both decay (pages disappearing from the Web) and modification. Even though most of the pages from 1998 either disappeared or ceased to contain the term *informetrics*, 165 pages (19.1%) still exist in 2006 and contain the search term. We followed the “fate” of these 165 pages: characterized the publishers, the contents and the changes that occurred the whole period. In recent years e-print servers and publishers’ sites became sources of large number of pages related to *informetrics*. Longitudinal studies following the evolution of a topic on the Web are very important, since they provide insights about content and the underlying Web processes.

Keywords

informetrics; longitudinal web study; web evolution; growth decay

Introduction

The World Wide Web is continuously growing at an incredible speed both in terms of its content and in terms of the number of users accessing it. The Web has become an indispensable source for information and research. Its growth patterns are of interest for theoretical, technical, social and economic reasons.

The present study examined the evolution of *informetrics* on the Web. To be more specific, we identified Web pages containing either the term *informetrics* or *informetric*. Two complimentary data collection techniques were utilized: retrieving data from multiple search engines and revisiting Web pages identified at previous data collection points. The combination of the two techniques allowed us to study several evolution patterns: creation of new pages, removal of previously existing ones and modifications. This is the first study that we are aware of that tracks the evolution of a topic on the Web for such a long period of time using multiple collection methods for data collection.

Literature review and background

Longitudinal studies of the Web

Several previous studies examined sites and pages for shorter periods of time, usually for several weeks or months (e.g. Bar-Ilan & Peritz, 1999; Brewington & Cybenko, 2000; Cho & Garcia-Molina, 2000; Fetterly et al., 2004; Ntoulas et al., 2004 or Kim & Lee, 2005). However in these shorter term studies, the data sets were usually huge and the monitored pages were visited more often (typically once a week).

There are only a few studies that report findings based on several years of data collection, but even these are for shorter length than the current study. One of the longest studies to this day was carried out by Koehler (2004). He observed a fixed set of pages 361 Web pages for 325 weeks (over six years). The original set was assumed to be “a random representation of the Web as a whole”. Although only

122 of the pages were still accessible after six years, the author concludes that pages tend to stabilize as they become older. Gomes and Silva (2006) had data on the Portuguese national Web for a period of three years (8 data collection points) and their conclusion was that the lifetimes of URLs and their content can be modelled as logarithmic functions. Baeza-Yates and Poblete (2003) based their results on three data collection points over a period of three years of the Chilean Web. Their conclusion is that although the Web keeps growing, a significant part of it disappears. Bar-Ilan and Peritz (2004) report the results of a five-year long study. Toyoda and Kitsuregawa (2006) had access to the Japanese Web archive which collects data about once a year, and based their results on data from 2003-4 (three data collection points). They wanted to identify whether newly discovered Web pages are actually new or they already existed on the Web, "waiting to be found". Ortega et al. (2006) crawled about a thousand sites twice, once in 1997 and once in 2004; their results show considerable growth of different types of Web elements (e.g., images and links) over time. Rousseau (1999) studied the changes in the number of search results reported by AltaVista and Northern Light on three queries for a period of 84 days.

All previous studies that we were able to locate used a single data collection method. They either monitored a fixed data set (e.g., Fetterly et al., 2004 or Koehler, 2004) or crawled in a pre-specified manner a fixed number of pages from given starting points (e.g. Cho & Garcia-Molina, 2000 or Kim & Lee, 2005), or attempts were made to download complete Websites (e.g. Ntoulas et al., 2004) and/or entire national Webs (e.g., Baeza-Yates & Poblete, 2003 or Toyoda & Kitsuregawa, 2006). Ke et al. (2006) survey a large number of studies in the area of Web dynamics.

Persistence of Web references

Web sources are being referenced in scholarly publications, both Web pages and sites and scholarly publications freely available on the Web. Web references, unlike references to printed sources can disappear or change. A number of studies discussed this issue and evaluated its extent.

One of the largest studies was carried out by Lawrence et al. (2001), in which they tried to locate 67,577 URLs referenced in articles indexed by the Citeseer database (publications dates between 1993 and 1999). On the one hand, they found considerable increase in the number of Web-references over the years, but on the other hand they emphasized the lack of persistence of the Web-references – 54% of the Web-references published in 1994 were not accessible by 2000. Spinellis (2003) studied the availability of URLs mentioned in two computer science journals: Computer and Communications of the ACM – 72% of the URLs were retrieved without problems. Sellitto (2005) analyzed the references of conference papers from the AusWeb conference series; on the average 45.8% of the references were not locatable by November 2003, even for the papers published on July 2003, 9% of the Web references were already missing.

Casserley and Bird (2003) studied 1,425 LIS research articles published in 1999 and 2000. They were able to find at the original URLs, 56.4% of the references of their sample of references. The percentage of accessible Web references was increased through searching Google and using the Internet Archive – altogether 89.4% of the Web references were located. Markwell and Brooks (2003) complain about "link rot" as a limiting factor of Web-based references. In two years 20% of the URLs disappeared, moved or changed their content. McCown et al. (2005) analyzed the references in the D-lib Magazine, and found that about 30% of the sample of Web-references published between 1995 and 2004 failed to resolve by February 2005. Tyler and McNeill (2003) studied the longevity of 2,729 URLs that appeared in College & Research Libraries News Web bibliographies. They conclude that the half-life of the URLs in these lists is about 5 years. They also located "undead" URLs – URLs that seemed to be "dead" at the initial check, but became "alive and well" when they rechecked these URLs six weeks later. In a most recent article (Goh & Ng, n.d.) examined the extent of "link rot" in three leading IS journals for a sample of articles published in 1997-2003. The authors were unable to access 31% of the 2,516 Web citations extracted from these publications.

Wren (2004) studied URL references in PubMed abstracts and found that about 63% of them were accessible. In a recent paper Wren et al. (2006) studied URL decay in dermatology journals and found a relatively high availability (81.7% -varying by publication year). However, most authors agreed that

the unavailable URL content was important for the publication. Nelson and Allen (2002) monitored the persistence of 1,000 digital library objects accessible through the Web between November 2000 and December 2001. These items came from digital libraries containing freely accessible collections of scientific materials. The observed objects were reports, e-prints or re-prints of scientific and technical information. Only 3% of the sample disappeared during the time span. Thus, compared with the stability of general Web pages, Web-references seem to be much more stable, possibly because these Web references refer to higher quality content produced by more authoritative sources than the “average” Web page. Presumably, pages containing the term *informetrics* (in the bibliometric-scientometric sense) will be higher quality pages as well (it was previously found that a large number of these pages contain bibliographic references – see Bar-Ilan, 2003), thus we can expect slower decay of these URLs as well.

Methods

Data collection

The experiment started in January 1998. In the first stage (until June 1998) data was collected from the then major search engines (AltaVista, Excite, Hotbot, InfoSeek, Lycos and Northern Light) by running the query *informetrics OR informetric*. Originally we intended to run the query *informetrics* only, but because of Northern Light’s automatic stemming the query had to be extended. Data was collected once a month and changes between the data collected in consecutive data collection points were observed. In June 1998, 866 URLs were identified through the collective effort of the above-mentioned search engines. The query was chosen because we were looking for information on the scientific field *informetrics* - quantitative analysis of documents in all forms. However, as can be expected, on the Web *informetrics* has additional meanings as well (e.g., names of companies).

Search results fluctuated considerably between the data collection points, thus when rerunning the experiment in June 1999, an additional data collection method was employed besides querying the search engines. The URLs that satisfied the query in June 1998 were revisited in 1999 even if they were not located by the search engines in 1999. No data was collected in 2000 and in 2001. However, in retrospect this has not been a shortcoming of the research, since the growth and modification patterns can be easily interpolated for the missing data collection points (see Fig 1).

Thus in June 1999, 2002, 2003, 2004, 2005 and 2006 two separate data collection procedures were employed

1. Submitting the query *informetrics OR informetric* to the largest search engines at the time
 - a. In 1999 the same search engines were used as in 1998, namely AltaVista, Excite, Hotbot, InfoSeek, Lycos and Northern Light
 - b. In 2002 and 2003 AllTheWeb, AltaVista, Google, HotBot, Teoma and Wisenut were employed. By 2002 search engines started to retrieve non-html pages as well (pdf, ps, doc, etc.)
 - c. In 2004, we queried AllTheWeb, AltaVista, Gigablast, Google, Hotbot, Teoma, Yahoo and Wisenut. Note that in June 2004, AllTheWeb and AltaVista still retrieved slightly different results from the then newly launched Yahoo search engine; and Hotbot served a different set of results as well.
 - d. In 2005 and 2006, Exalead, Google, MSN, Teoma (Ask) and Yahoo were queried.

Although the initial data set was rather small (less than 900 URLs), enormous growth was witnessed during the years, and in 2006 the search engines retrieved 24,272 different URLs (4,642 additional URLs were located through the “revisit” process in 2006).

Search engines limit the number of displayed result for a query (the limitations as of June 2006 were: 1000 for Google, Yahoo, 2000 for Exalead, 250 for MSN and 200 for Teoma). In order to try to overcome these limitations we used several techniques:

- a. Including/excluding additional search terms (e.g. *informetrics -scientometrics* and *informetrics scientometrics*)
- b. Limiting the query by site or filetype e.g. *informetrics site:.es* (pages from Spain only) – including/excluding sites or filetypes.
- c. Limiting the query by date (the betweendate feature of Ask)

In one case we included/excluded 22 additional terms in order to break down the query results into small enough chunks.

The whole set of searches on all the search engines were run within 1-2 hours to minimize the effect of time on the results. For each year the searches were carried out in June. The URLs were extracted from the search results pages and duplicates (usually the same URL retrieved by several search engines) were eliminated. The URLs were compared as text strings, thus, for example, *informetrics.com* and *www.informetrics.com* were considered two different URLs.

All the documents residing at the identified URLs were downloaded to our local computer within 0-2 days of the searches, in order to minimize the effect of the time elapsed between the search time and download time on the possible changes that the documents undergo over time. A second attempt was made to download inaccessible URLs. Finally, the entire set of html documents was tested for the presence of the string *informetric*.

2. All pages that contained either the term *informetrics* or the term *informetric* (i.e., satisfied the query) at least at the first time that they were identified by the search process were revisited at each of the later data collection points.

The combination of the two methods allowed us both to follow the "fate" of previously identified pages and to enrich the collection of pages with newly retrieved ones from the search engines. Note that newly retrieved pages are not necessarily newly created pages. It is possible that the page existed before and was indexed by some of the search engines, but it did not contain the search term; or because of the incomplete coverage of the Web by the search engines, it is quite plausible that the page existed for a long time and was relevant to the search but was only discovered at one of the later data collection points. We are not aware of any other study that utilized multiple data collection methods for studying the evolution of a topic on the Web.

Data analysis

We analyzed the longitudinal patterns of the data set in general and the changes in the distribution of the domains over time. Our basic notion, *technical relevance*, defines whether a URL satisfies the query at a certain time. All html, text and office documents and a sample of the other document types (pdf and postscript) were checked for the presence of at least one of the search terms. We decided to use the terminology *technical relevance* instead of the more widely used term *relevance* in order to avoid the complex issues of defining relevance (see for example (Saracevic, 1996) or (Mizzaro, 1998)).

- All the URLs were checked at each *data check point* (in June of 1998, 1999, 2002-2006) for technical relevance (*trel* in short). Note that it is quite possible that the URL is accessible at a data check point, however its content was *modified* and the page ceases to be technically relevant to the query.
- Some of the documents were not accessible at the data check points. This inaccessibility could be temporary (*intermittent* URLs caused mainly by communication or server problems) or permanent. If from some point onwards the URL was never accessible then we conclude that the URL *disappeared*. Note that it is possible that a URL defined as disappeared based on the available data, will become intermittent if the data monitoring continues for a longer time.

Results and discussion

Growth, disappearance and modification

During the whole period 36,282 different URLs were identified that satisfied the query at least at the first time they were located. Table 1 and Figure 1 decribe the overall growth of the topic over the years

as reflected by the number of *technically relevant URLs* identified at each of the data collection points. The growth over the years is considerable, 80.2% of the total unique URLs identified during the whole period located and satisfied the query at the last data check point (2006), while only 2.3% of the total were discovered by the search engines in 1998. When analyzing the data we have to take into account two processes: the growth of the Web as a whole, and changes in coverage of the search engines.

Table 1. Number of technically relevant (*trel*) URLs identified at the *data check points*

	Total <i>trel</i> (% out of total)	<i>Trel html or text</i>	<i>Trel pdf</i>	<i>Trel MS Office</i>	<i>Trel postscript</i>	<i>Trel xml</i>
1998	866 (2.4%)	866	0	0	0	0
1999	1,249 (3.3%)	1,249	0	0	0	0
2002	4,034 (11.1%)	3,705	272	31	26	0
2003	5,176 (14.3%)	4,399	625	92	60	0
2004	8,454 (23.3%)	7,225	1,027	140	62	0
2005	13,454 (37.1%)	11,594	1,577	210	73	0
2006	28,914 (80.2%)	25,358	3,349	310	63	18
<i>Total unique URLs during whole period</i>	36,282	31,999	3,839	360	84	18

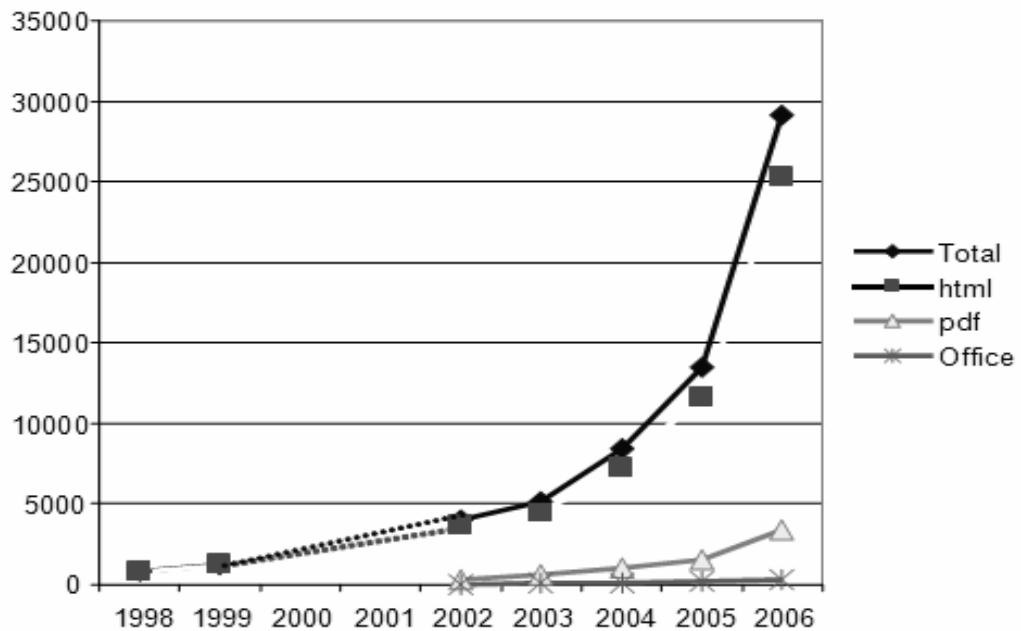


Figure 1. Growth curves for the different document types (growth curves interpolated for 2000 and 2001)

Growth is the definitely the strongest of all three processes: growth, decay and modification, however the other processes are considerable as well. Table 2 displays the changes that occurred to the original set of 866 URLs over time.

Table 2. The set of 866 URLs located in 1998 at the different data check points

	1999	2002	2003	2004	2005	2006
<i>trel</i>	648	291	242	216	176	156
<i>intermittent</i>	0	0	1	3	7	9
<i>inaccessible/disappeared</i>	183	495	551	575	615	629
<i>term not in document</i>	35	80	71	72	68	72

We observe that in 2006, 165 (155 *trel* + 9 *intermittent*) documents out of the original set of 866 documents are still accessible and still satisfy the query.

Page distribution on servers

In order to gain a better understanding of the document types, we tabulated the most “prolific” servers (i.e., servers with the largest numbers of *trel* pages that were located in each year) for the years 1998, 2002 and 2006, in Tables 3, 4 and 5 respectively.

Table 3. The most “prolific” servers in 1998 – number and percentage of html pages (N=866)

	Server	No. pages	% pages
1	<i>db.dk</i>	54	6.2%
2	<i>sfu.ca</i>	44	5.1%
3	<i>crrm.univ-mrs.fr</i>	24	2.8%
4	<i>bubl.ac.uk</i>	22	2.5%
5	<i>hu-berlin.de</i>	22	2.5%
6	<i>ust.hk</i>	21	2.4%
7	<i>informetric.com</i>	20	2.3%
8	<i>informatik.uni-trier.de</i>	18	2.1%
9	<i>informatik.rwth-aachen.de</i>	18	2.1%
10	<i>csic.es</i>	15	1.7%
	Total	258	29.8%

In 1998 the pages located by us were residing on 347 servers. The top ten servers covered 29.8% of the pages. Interesting to note that with the exception of *informetric.com* (a software company, not related to *informetrics* in the scientific sense) and *nasa.gov* (a WAIS server – see http://en.wikipedia.org/wiki/Wide_area_information_server for a reminder of what WAIS was - with huge lists of publications, a few of them related to *informetrics*); all the other servers are from outside the US, most of them in Europe. The Danish pages (from *db.dk*) are from the Royal School of Library and Information Science. The pages from Simon Fraser University (*sfu.ca*) are part of an early electronic library project. The French server (*crrm.univ-mrs.fr*) is not functional anymore, but it had information on the ISSI Society and its conferences. *bubl.ac.uk* served abstracts and tables of contents. The pages from Humboldt University (*hu-berlin.de*) mostly contained information on the curriculum for the Library Science studies. The pages from *ust.uk*, *uni-trier.de* and *rwth-aachen.de* are mirror pages of the DBLP server (Computer Science Bibliography) that contained the terms *informetric* or *informetrics*. The DBLP project provides bibliographic information on major computer science journals and proceedings (<http://www.informatik.uni-trier.de/~ley/db/welcome.html>). Finally, *csic.es* hosts the journal Cybermetrics – International Journal of Scientometrics, Informetrics and Bibliometrics (<http://www.cindoc.csic.es/cybermetrics/>).

Table 4. The most “prolific” servers in 2002 – number and percentage of html pages (N= 3,705)

	Server	No. pages	% pages
1	<i>unsw.edu.au</i>	136	3.7%
2	<i>db.dk</i>	128	3.5%
3	<i>csic.es</i>	74	2.0%
4	<i>uni-trier.de</i>	72	1.9%
5	<i>collnet</i>	71	1.9%
6	<i>citeSeer</i>	62	1.7%
7	<i>hu-berlin.de</i>	54	1.5%
8	<i>acm.org</i>	46	1.2%
9	<i>enssib.fr</i>	46	1.2%
10	<i>utk.edu</i>	45	1.2%
Total		734	19.2%

Comparing Tables 3 and 4 there are some striking differences. The top server, *unsw.edu.au* is a “newcomer” (only 4 pages from this server were located in 1998). Most of the captured pages relate to the 2001 ISSI Conference in Sydney, hosted by the University of New South Wales. Collnet (a global interdisciplinary research network for the study of all aspects of collaboration in science and technology - <http://www.collnet.de/>) had pages on several servers (collnet.de, collnet.org and collnet.net, as of November 2006 only collnet.de is active). It provides information on the group members, most of them ISSI members. Citeseer is a very extensive computer science digital library. ACM is the Association for Computer Machinery, the sites serves as a DBLP mirror site and the SIGIR newsletter that sometimes relates to *informetrics* is also published on this site. The site enssib.fr hosts the pages of SLISNET (Schools of Library and Information Science Network - <http://www.enssib.fr/autres-sites/SLISNET/>) that provides information on school curricula and on research interests of the research staff of the participating institutions. Discussions of the SIGMETRICS group are archived on listserv.utk.edu. This discussion list covers bibliometrics, scientometrics and informetrics and is a Virtual-SIG of ASIST (<http://web.utk.edu/~gwhitney/sigmetrics.html>).

Table 5. The most “prolific” servers in 2006 – number and percentage of html pages (N=25,358)

	Server	No. pages	%
1	<i>csic.es</i>	1,812	7.1%
2	<i>doclib.uhasselt.be</i>	1,351	5.3%
3	<i>acm.org</i>	1,102	4.3%
4	<i>garfield.library.upenn.edu</i>	1,038	4.1%
5	<i>doclib.luc.ac.be</i>	996	3.9%
6	<i>elsevier.com</i>	881	3.5%
7	<i>utk.edu</i>	758	3.0%
8	<i>eprints.rclis.org</i>	591	2.3%
9	<i>blogspot.com</i>	456	1.8%
10	<i>citeSeer</i>	333	1.3%
Total		9,318	36.7%

There is a huge increase in the number of pages from the *csic.es* site, which is mainly explained by the number of pages from the Cybermetrics journal that were located in 2006. The site *doclib.uhasselt.be* (and its earlier address *doclib.luc.ac.be*) is a document server for preprints, published articles and technical reports (<http://doclib.uhasselt.be/dspace/>), similarly to *eprint.rclis.org*, which is an open access archive for information and library science (<http://eprints.rclis.org/>). Google lately changed its

indexing policy – it currently indexes publishers' sites, even if access to the publication is fee or subscription based; if the publisher provides free access at least to the abstract of the article. This is the reason for the large number of pages from the ACM Digital Library (<http://portal.acm.org/dl.cfm>) and from Elsevier. The search engines also picked up the huge amount of full text and bibliographic information from Eugene Garfield's site, which is of course closely related to *informetrics*. The site blogspot.com appears in Table 5, as a result of the combined effort of 29 bloggers (<http://www.blogger.com/start>), the most prolific author is Francisco Javier Martínez Méndez (<http://irsweb.blogspot.com/>) who maintains a Spanish blog on Web information retrieval, and at the time of data collection had a link on the sidebar of his blog pages linking to the online copy of the Egghe & Rousseau book on Informetrics. As of November 2006 this link does not appear on the sidebar of the blog, but in June 2006, 252 technically relevant blog pages were located. The second most prolific author (with 135 pages located in June 2006) is Pedro Principe, who maintains the Portuguese blog Pedro Principe Rato de Biblioteca (<http://ratodebiblioteca.blogspot.com/>) and links from the sidebar of his blog to the Cybermetrics journal. It is interesting to note the major influence blogs have on the number of pages and links related to specific topics. Note that in 2006 the top ten servers (out of a total of 4784 servers; 0.21% of the servers) covered 36.8% of the pages; while in 2002 the top ten servers (out of a total of 1182 servers; 0.85% of the servers) covered on 19.2% of the pages). There seems to be a larger concentration of technically relevant pages on a smaller number of servers over time.

The persistent set of URLs from 1998

There were 165 URLs that were accessed and technically relevant at each of the data check points. We decided to analyze the contents of these pages (see Krippendorff, 2003 or Neudorf, 2001) and to characterize the modifications that occurred to these pages over time. However after a closer examination of these pages we discovered that a number of these pages were duplicates (from mirror sites and/or from alternative URLs). After the removal of the duplicates, the set was comprised of 97 URLs. We call this set the *persistent set* of URLs. For each URL we identified the publisher, the page type, the context in which *informetrics* was mentioned on the page and the type and frequency of modifications. In Table 6 the publishers with the largest number of pages in this set are displayed.

Table 6. The publishers with the largest number of pages in the *persistent set*

Publisher	No. pages (% of total)	Comments
<i>Cybermetrics</i>	12 (12.4%)	The electronic journal, Cybermetrics - International Journal for Scientometrics, Informetrics and Bibliometrics
<i>DBLP</i>	11 (11.3%)	Computer Science bibliography, publication lists of authors and TOCs of journals/proceedings
<i>Hebrew University</i>	8 (8.2%)	Hosted the 1997 ISSI Conference in Jerusalem
<i>Humboldt University</i>	8 (8.2%)	Study curriculum, pages of the Society for Science Studies, personal pages and list of events
<i>Simon Fraser U., Rob Cameron</i>	6 (6.2%)	An early project - electronic library in computer science
<i>SLISNET</i>	6 (6.2%)	Schools of Library and Information Science Networks, pages on participating institutions, study curriculum and research interests
<i>Danmarks Biblioteksskole</i>	5 (5.2%)	Newsletters and conference announcements from the Royal School of Library and Information Science, Denmark
<i>Informetric company</i>	4 (4.1%)	Pages of the Informetric System Inc. - a software company
<i>SIGIRLIST</i>	4 (4.1%)	Newsletters of the Special Interest Group on Information Retrieval
<i>CSNA</i>	3 (3.1%)	Newsletters of the Classification Society of North America
Total	67 (69.1%)	

Table 7 provides information on the different pages types, while Table 8 displays the distribution of contexts in which the term *informetrics* appeared on the pages of the *persistent set*. Note that the term may occur more than once on the page and it is quite possible that the different occurrences are categorized under different contexts. Thus the total for contexts is 127 and not 97 (the size of the *persistent set*).

For most of the pages in the *persistent set* (50 pages, 51.5%) no changes were observed at any of the data check points. Twenty seven pages (27.8%) were updated during the period: five pages were updated only once, six pages twice and the rest three or more times (out of the six data check points). For some pages (16 pages, 16.5%) only the formatting changed, such changes were mostly observed once or twice for this subset (for eleven out of the sixteen pages). Finally no content or format changes were observed on four pages, only the “data of last update” was updated. Thus the set of *persistent* pages were not modified considerably during the period, some of these pages were seemingly “forgotten” or “abandoned”.

Table 7. The distribution of page types in the *persistent set*

Page type	No.	%
<i>newsletter/news item</i>	13	13.4%
<i>publication list</i>	11	11.3%
<i>conference page</i>	10	10.3%
<i>content page</i>	8	8.2%
<i>journal list</i>	7	7.2%
<i>course list</i>	6	6.2%
<i>report/article</i>	6	6.2%
<i>TOC</i>	5	5.2%
<i>company page</i>	4	4.1%
<i>course page</i>	4	4.1%
<i>faculty list/ list of researchers</i>	4	4.1%
<i>abstract/list of abstracts</i>	3	3.1%
<i>homepage</i>	3	3.1%
<i>event list</i>	2	2.1%
<i>list of institutions</i>	2	2.1%
<i>other list</i>	8	8.2%
<i>other</i>	1	1.0%

Table 8. The distribution of contexts in which of the term *informetrics* used in the *persistent set*

Use of <i>informetrics</i> on page	No. occurrences	% occurrences (out of 127)
<i>publication/presentation title</i>	29	22.8%
<i>Cybermetrics</i>	21	16.5%
<i>ISSI</i>	15	11.8%
<i>topic discussed/expanded</i>	15	11.8%
<i>research interest/area</i>	12	9.4%
<i>conference topic</i>	8	6.3%
<i>ISSI proceedings</i>	8	6.3%
<i>name of institute/company</i>	8	6.3%
<i>other</i>	5	3.9%
<i>course topic</i>	4	3.1%
<i>affiliation</i>	2	1.6%

Conclusion

In this study we followed the evolution of a topic on the Web for a period of eight years. This is the longest systematic longitudinal study that we are aware of. The use to two complimentary data collection methods: retrieval from search engines and revisiting of previously existing pages allowed us to study the growth, decay and modification processes of pages that contain the term *informetrics*. A structured and detailed analysis of the content and page distribution allowed us to obtain a more qualitative understanding of the evolution of pages on the Web, in our case for the term and field of *informetrics*. Although such studies are extremely labor intensive, it is highly recommended to conduct additional ones. Since the Web has become during a relatively short period of time a major source not only for information but also for research and we rely more and more on this source, in depth longitudinal studies can provide vital information on its coverage and changes over time. They can help in understanding the role of the Internet on the overall development of a topic or a field.

References

- Bar-Ilan, J. (2003). The Web as information source on informetrics? - A content analysis. *Journal of the American Society for Information Science*, 51(5), 432-443.
- Bar-Ilan, J. & Peritz, B. C. (1999). The life-span of a specific topic on the Web – The case of “informetrics”: a quantitative analysis. *Scientometrics*, 46, 371-382.
- Bar-Ilan, J., & Peritz, B. C. (2004). Evolution, continuity and disappearance of documents on a specific topic on the Web - A longitudinal study of “informetrics”. *Journal of the American Society for Information Science and Technology*, 56, 980-990.
- Baeza-Yates, R. & Poblete, B. (2003). Evolution of the Chilean Web structure composition. In *Proceedings of the First Latin American Web Congress (LA-WEB 2003)*. Retrieved November 12, 2006 from: http://www.laweb.org/2003/stamped/02_baeza-yates-poblete.pdf
- Brewington, B. E., & Cybenko, G. (2000). Keeping up with the changing Web. *Computer*, 33(5), 52-58. Retrieved November 12, 2006 from: <http://ieeexplore.ieee.org/iel5/2/18198/00841784.pdf?arnumber=841784>
- Casserly, M. F. & Bird, J.E. (2003). Web citation availability: analysis and implications for scholarship, College and Research Libraries 64(7), 300–17.
- Cho, J. & Garcia-Molina, H. (2000). The Evolution of the Web and implications for an incremental crawler. In *Proceedings of 26th International Conference on Very Large Databases (VLDB)*, September 2000, (pp. 200-210).
- Fetterly, D., Manasse, M., Najork, M. & Wiener, J. L. (2004). A large scale study of the evolution of Web pages. *Software – Practice and Experience*, 34, 213-237.
- Goh, D. H. & Ng, P. K. (no date). Link decay in leading information science journals. To appear in *JASIST*. Retrieved November 17, 2006 from: <http://www3.interscience.wiley.com/cgi-bin/fulltext/113452914/HTMLSTART>
- Gomes, D. & Silva, M. J. (2006). Modeling information persistence on the Web. In *Proceedings of the 6th International Conference on Web Engineering (ICWE06)* (pp.193-200).
- Ke, Y., Deng, L., Ng, W. & Lee, D. L. (2006). Web dynamics and their ramifications for the development of Web search engines. *Computer Networks*, 50, 1430-1447.
- Kim, S. J. & Lee, S. H. (2005). An empirical study on the change of Web pages. In *Proceedings of APWeb 2005*, LNCS 3399 (pp. 632 – 642).
- Koehler, W. (2004). A longitudinal study of Web pages continued: A report after six years. *Information Research*, 9(2) paper 174. Retrieved November 12, 2006 from: <http://InformationR.net/ir/9-2/paper174.html>
- Krippendorff, K. (2003). *Content analysis: An introduction to its methodology*. 2nd edition. Sage Publications.
- Lawrence, S., Pennock, D. M., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. A. & Giles, L. E. (2001). Persistence of Web references in scientific research. *Computer*, 34(2), 26-31.
- Markwell, J. & Brooks, D. W. (2003). “Link rot” limits the usefulness of Web-based educational material in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education*, 31(1), 69-72.
- McCown, F., Chan, S., Nelson, M. L. & Bollen, J. (2005). The Availability and Persistence of Web References in D-Lib Magazine. *5th International Web Archiving Workshop (IWA05)*, Vienna, Austria. Retrieved November 12, 2006 from: <http://arxiv.org/ftp/cs/papers/0511/0511077.pdf>
- Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting With Computers*, 10(1998), 305-322. Retrieved November 12, 2006 from: <http://www.dimil.uniud.it/mizzaro/research/papers/IwC.pdf>
- Nelson, M. L., & Allen, B. D. (2002). Object persistence and availability in digital libraries. *D-Lib Magazine*, 8(1). November 12, 2006 from: <http://www.dlib.org/dlib/january02/nelson/01nelson.html>
- Neudorf, K. A. (2001). *The content analysis guidebook*. Sage Publications.
- Ntoulas, A., Cho, J. & Olston, C. (2004). What's new on the Web? The evolution of the Web from a search engine perspective. In *Proceedings of the World-Wide Web Conference (WWW)*, May 2004, (pp. 1-12).

- Ortega, J. L., Aguillo, I. & Prieto, J. (2006). A longitudinal study of content and elements in scientific Web environment. *Journal of Information Science*, 32, 344-351.
- Rousseau, R. Daily time series of common single word searches in AltaVista and Northern Light. *Cybermetrics*, 2/3(1), paper 2. (1999). Retrieved November 12, 2006 from:
<http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html>
- Saracevic, T. (1998). Relevance reconsidered. In *Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)*, Copenhagen, Denmark (pp. 201-218).
- Sellitto, C. (2005). The impact of impermanent Web-located citations: A study of 123 scholarly conference publications. *Journal of the American Society for Information Science and Technology*, 56(7), 695-703.
- Spinellis, D. (2003). The decay and failures of URL references. *Communications of the ACM*, 46(1):71-77.
- Toyoda, M. & Kitsuregawa, M. (2006) What's really new on the Web? Identifying new pages from a series of unstable web snapshots. In *Proceedings of WWW2006 (2006)*, (pp. 233-241).
- Tyler, D. C. & McNeil, B. (2003). Librarians and link rot: A comparative analysis with some methodological considerations. *portal: Libraries and the Academy*, 3(4), 615-632.
- Wren, J. D. (2004). 404 not found: The stability and persistence of URLs published in Medline. *Bioinformatics*, 20(5), 668-672.
- Wren, J. D., Johnson, K. R., Crockett, D. M., Heilig, L. F., Schilling, L. M. & Dellavalle, R. P. (2006). Uniform Resource Locator decay in dermatology journals. *Archives of Dermatology*, 142:1147-1152.

International Collaboration, Mobility and Team Diversity in the Life Sciences: Impact on Research Performance^{1,2}

Franz Barjak* and Simon Robinson**

*franz.barjak@fhnw.ch

School of Business, University of Applied Sciences Northwestern Switzerland, Rickenbachstr. 16,
CH-4600 Olten (Switzerland)

**Simon.Robinson@empirica.com

empirica, Gesellschaft für Kommunikations- und Technologieforschung mbH, Oxfordstr. 2,
D-53111 Bonn (Germany)

Abstract

The combination of knowledge and skills from different backgrounds or research cultures is often considered good for science. This paper describes the extent to which academic research teams in the life sciences draw on different national knowledge pools and how this is related to their research performance. We distinguish between international collaboration between research teams and international mobility leading to team diversity, where scientists with a background in another country work as members of a team over time. Our findings confirm previous results on the positive relationship between international collaboration and team performance. Moreover, we show that the most productive teams have a moderate level of diversity: maximizing diversity does not maximize performance. These results have implications for research team management and for research policy, in particular pointing out a need for adequate integration support to mobile scientists.

Key words

research team; team diversity; international collaboration; research productivity; mobility in science

Introduction

The level of international cooperation and communication in science has been shown to have increased significantly over the past twenty to thirty years (European Commission, 2003; Narin et al., 1991; National Science Board, 2004). Though the practice of global research is evident, it is more difficult to pin down the impact of research globalisation. It seems reasonable to speculate that science draws some benefit if excellence is “matched” globally: scientists who do not find the ideal complementary expertise in their country may increase their output (quantity and quality) working with peers abroad. Moreover, location-specific objects and conditions of research may be brought together (e.g. natural resources, specific political and social regulations etc.). When human and physical resources from several countries are combined a richer mixture of research inputs is generated and scientific knowledge is advanced more rapidly. This, presumably beneficial, international combination and cross-fertilization of expertise is brought about by mobility:

“[Mobility] permits the creation and operation of multi-national teams and networks of researchers, which enhance Europe’s competitiveness and prospective exploitation of results.” (European Commission, 2001: 4)

Indeed, increasing the mobility of researchers has become a prominent goal in European research policy (European Commission, 2001, 2005). In support of this policy, it is pointed out how geographical mobility is assumed to lead to productive combination of localised knowledge, to fertilise intellectual exchange (Jöns, 2003, 2006), to foster international research collaboration and

¹ This paper builds on data gathered and analytical approaches developed in the "Study on the role of networking in research activities" (NetReAct) for the Institute for Prospective Technological Studies of the European Commission under contract 22540-2004-12 F1ED SEV DE. We thank the EC and all colleagues working on NetReAct, in particular Alexander Mentrup, Wolfgang Glänzel, Xuemei Li and Mike Thelwall for their contributions.

² Due to the length restrictions of the ISSI 2007 conference proceedings the paper underwent numerous cuts and may seem incomplete in some locations. A more complete version can be obtained from the authors or in *Social Geography*, 3 (<http://www.soc-geogr.net>, accepted for publication).

disseminate research excellence (European Commission, 2005). Despite the extent of policy measures to promote the mobility of researchers, few studies have investigated the impact of mobility in science on research performance and output.

This study aims to reduce this gap. *International researcher mobility*, with the diversity of teams by origin it causes, is addressed in the study as one of two principal modes by which pools of knowledge and scientific expertise merge across national boundaries. The second mode is *international collaboration* among research teams, where scientists join forces across borders in their work but remain located with their teams in different countries.³ The study sets out to explore the effect of the diversity of teams caused by the international mobility of researchers and the international collaboration between researchers on the performance of research teams in a large and growing scientific domain: the life sciences.

Pooling knowledge for research

Evidence of the impact of diversity of geographic origin

Empirical studies of the influence of diversity of geographical origin of team members (*origin diversity*) on the performance of these research teams are scarce. However, they provide some support to the idea that the mobility of scientists is good for research output. An investigation into the scientific success of the Rockefeller Institute showed a positive contribution of foreign permanent staff as well as of visiting scientists (Hollingsworth & Hollingsworth, 2000). However, the authors point out that the impact found may well have been due to the greater scientific eminence of the visitors rather than their geographic origins *per se*. It has been shown for the U.S.A. that foreign-born and foreign-educated scientists make more exceptional contributions to scientific output than would be expected from their proportion of the scientific workforce (Stephan & Levin, 2001).

When more general findings relating to the impact of group diversity on performance are applied to the possible impact of origin diversity on research team performance, such positive conclusions are opened to question. Research on groups in various settings has shown that diversity may have negative as well as positive effects on group processes and performance (Williams & O'Reilly, 1998). Positive effects are attributed to a broader range of knowledge, skills and contacts in the group, whereas negative effects arise from reduced and less efficient communication, less cooperation and more conflict (Bunderson & Sutcliffe, 2002; Williams & O'Reilly, 1998). Jehn et al. (1999) found that whereas differences in knowledge have a positive impact on performance, the impact of a diversity of values is negative. In the area of science, Jöns (2003) agrees that impact is ambivalent, arguing that cross-border interactions might lead to inspiration and cooperation but also irritation and confrontation, depending on the degree of heterogeneity between the parties involved. Whether positive or negative effects prevail has been related to the origin, degree and type of diversity within a group or team. In terms of degree of diversity, Williams and O'Reilly (1998) suggest that positive effects may prevail at low levels of diversity, but that at higher levels, group cohesion may reduce to an extent which negates the positive effect.

Evidence of the impact of international collaboration

Origin diversity is only one possible channel for a team to combine knowledge and technical skills from different national backgrounds. Another channel is the acquisition of knowledge through research collaboration. Katz and Martin (1997) conclude in their review paper that the empirical evidence supports the idea of a positive relationship between collaboration and research productivity.

In principle, the benefits quoted for international research collaborations are of the same nature as those listed for collaboration in general, see e.g. the benefits listed by Georghiou (1998); but international collaboration clearly gives rise to additional costs, for instance due to the necessity of bridging language and cultural differences or finding suitable contractual arrangements. It is clear that

³ These two modes are a useful simplification of all possible interaction types along the dimensions of duration, distance and interaction intensity (see Fiol & O'Connor, 2005).

international collaboration must bring additional benefits which outweigh higher transaction costs; otherwise it would be hard to explain its impressive growth rate (see European Commission, 2003; Narin et al., 1991; National Science Board, 2004). Such benefits might be access to equipment, local resources, data or other objects of study, or to eminent scientists and groups (Georghiou, 1998; Thorsteinsdóttir, 2000; Wagner, 2005).

Empirical evidence of the impact of international research collaborations on research productivity is mainly positive. We cite some results for the life sciences, the domain to be investigated in this paper: a study of Spanish biomedical research showed that international collaboration increased the productivity of team leaders and the impact of the published work (Bordons et al., 1996). Italian studies found a positive effect of the number of research collaborations with foreign non-profit institutions on the productivity of molecular biology and genetics research groups (Arora et al., 1998; Cesaroni & Gambardella, 2003). According to Narin et al. (1991) their finding that biomedical papers with international co-authors have greater impact than single-author and nationally co-authored papers can be generalised to other disciplines. Other studies have shown that international collaboration generally has a more pronounced positive effect on citation impact than local or domestic collaboration (Adams et al., 2005; Persson et al., 2004).

It must not be concealed that some contrary evidence has also been presented, often from other scientific domains (Adams et al., 2005; Carayol & Matt, 2004; Gläenzel, 2001; Gläenzel & Schubert, 2001). Gläenzel (2001) discovered a number of "cool links" - country pairs for which co-authored papers attract fewer citations than expected on the basis of the corresponding domestic reference standards (Gläenzel & Schubert, 2001). Though in biomedical research the citation impact of co-authored papers is generally found to be higher than the domestic impact of at least one of the involved authors, the impact of joint papers in chemistry and mathematics in some pairs of countries was found to be consistently lower than domestic impact. In addition, methodological concerns have been voiced and only partially been refuted: it has been shown that the larger number of self-citations (Herbertz, 1995) is not the main explanation for more citations to internationally co-authored papers (Raan, 1998).⁴ However, self selection – only the best scientists collaborate at international level – might indeed play a role (Bordons & Gomez, 2000).

The empirical evidence to date on international research collaboration shows a mixture of positive and negative impacts, but the positive aspects in regard to scientific productivity and visibility of the results seem to prevail.

Concepts and methods

The research team as the unit of analysis

The main unit of analysis in this study is the research team or group recognisable from outside the university as a distinct entity and understood as a group of people, scientists and non-scientists, some or all of whom are employed by a university, who work at the same location for a significant period of time to produce new scientific knowledge. Our definition is a blend of an institutional approach, which relies on organisational affiliation (Cohen, 1981; Hagstrom, 1965) and a functional approach, based on the specification of joint research activities (Andrews, 1979).

Limiting team membership to those who work at the same location allows us to address the impact on research performance of collaborative work spanning multiple locations. "Virtual teams", whose emergence is facilitated by the internet and other networks, are thus analysed not as a type of team but as collaborative activity between teams. For similar reasons, visiting scientists and research workers are not regarded as members of a team unless they stay collocated longer than a minimum period of six months.

⁴ We want to thank one of the reviewers for pointing this out.

Survey sampling and response

Webometric techniques were used to build a representative sample of 1,773 university-affiliated research teams in the life sciences across 10 European countries: Through internet research a population of 7,732 teams was identified (see Table 1) working in the life sciences, defined as ISCED 1997 category 42, and teams were drawn from this population by stratified random sampling. The stratification variable was the number of hyperlinks pointing to the team's internet homepage (inlinks). As previous research has shown that for universities and departments the number of hyperlinks is related to research performance (see Thelwall, 2003), this was deemed a possible way of securing the intended focus on successful teams. However, subsequent analyses with the data have shown that hyperlinks are not a good proxy for research performance at team level (at least in the life sciences, Barjak & Thelwall, 2006). For the sample teams we identified the names and email addresses of the team leaders via the internet.

Questionnaires were provided to team leaders electronically - online and via email. The survey produced 468 usable questionnaires (26.4% of the sample, see Table 1). A comparison of number of inlinks, team size and gender of the team leader between responding and non-responding teams revealed little bias in response. The Italian teams that responded tended to have somewhat fewer hyperlinks than those which did not respond. Teams with female team leaders were slightly overrepresented in Germany and underrepresented in Spain.

Using the information obtained from the internet and the survey it was possible to retrieve bibliographic data for the responding teams from the Thomson ISI Web of Science. Publication data was collected for the year 2001 and citation data for the years 2001-2003.

Table 2. Dataset of life sciences research teams by country

Country	Research population	Sample	Usable questionnaires				
			Num-ber	In % of sample	Mean no. of inlinks	Mean team size	% of female heads
CZ	173	119	30	25.2%	1.6	12.8	23.3%
DE	1,447	271	60	22.1%	9.5	16.6	20.0%
ES	896	164	37	22.6%	1.9	14.8	8.1%
FR	1,384	225	56	24.9%	4.4	16.8	12.5%
HU	214	108	34	31.5%	5.8	22.3	17.6%
IT	952	186	52	28.0%	1.5	8.8	21.2%
NO	199	122	37	30.3%	7.8	11.0	18.9%
PT	229	123	44	35.8%	11.4	12.6	50.0%
SE	650	148	41	27.7%	7.3	12.3	26.8%
UK	1,588	307	77	25.1%	8.7	13.1	13.0%
Total	7,732	1,773	468	26.4%	6.4	14.3	20.5%

Metrics for key variables

Research performance

Team research performance was operationalised in three variables built from bibliographic data extracted from the Science Citation Index Expanded (SCIE) provided by Thomson ISI:

- TOTPAP (output volume): TOTPAP is the total number of papers recorded in the 2001 SCIE volume as article, letter, note, or review authored or co-authored by a member of the team;
- ZTOTPAP (team productivity): this variable is TOTPAP divided by team size;
- TOTMOCR (output quality): the number of citations received up to 2003 for a team's 2001 papers is divided by TOTPAP to obtain the Mean Observed Citation Rate per publication for that team.

These indicators and the SCIE database itself have several well documented weaknesses, for instance, not all co-authors of publications really contribute intellectually, the bias towards English language publications might inflate values for native speakers of that language, self-citation inflates citation scores, citations are sometimes created for other reasons than the quality of a paper, etc. (Borgman & Furner, 2002; Cronin, 1984; Herbertz, 1995; Leeuwen et al., 2001; Raan, 2003). Notwithstanding these weaknesses, publications and citations are viewed as good measures for comparing and analysing research performance.

Origin diversity

Using an approach similar to Carayol & Nguyen Thi (2004) a Shannon Diversity Index of country of origin was calculated for each group of young researchers in a team with the formula below. Two indices were calculated per team: one for PhD students (ODIVPHD) and one for post-docs (ODIVPDOC). Country of origin was the country in which they obtained their most recent degrees.⁵ These indices represent in one value the degree to which different national pools of knowledge are present in the team. The larger the index, the larger the variety of countries in which the PhD students (post-docs) obtained their last degree, and the larger the variety of national pools of knowledge in a team.

$$\text{ODIVPHD} = -\sum_{i=1}^C (p_i * \ln p_i)$$

with ODIVPHD Origin diversity index of PhD students

 C Total number of different countries i where the PhD students of a team obtained their last degrees

p_i Proportion of C made up of the i th country

International collaboration

International collaboration was measured in this analysis by using data retrieved from the Thomson ISI database. Three binary indicators were built on co-author fields:

- ICPAP01 takes the value of one if the team has published one or more papers with co-authors from any foreign country, otherwise it is zero;
- EUCPAP01 takes the value of one if the team has published one or more papers with co-authors from another EU member state, otherwise it is zero;
- USCPAP01 takes the value of one if the team has published one or more papers with co-authors affiliated to organisations in the USA, otherwise it is zero.

Though it is known that co-authorship often reflects an intense research interaction between the authors (Harsanyi, 1993; Laudel, 2001), there is at least anecdotal evidence that co-authors might be included in a publication for other reasons, e.g. because they secured the resources for a project. However, we do not expect that this introduces significant bias at international level.

Modelling approach

In the estimation models we provide first a baseline model. This incorporates factors apart from knowledge-pooling mechanisms which impact on research productivity such as country of location or simply team size. The impact on research productivity of international collaboration and origin diversity is then examined in an extension of the baseline model.

Further independent variables of theoretical relevance and expected to improve the explanatory power of the model were included in baseline models. One set consist of team characteristics found to be influential in previous work such as country of team location, principal research discipline, age (time since foundation) and team size. Characteristics of team leaders which might affect research

⁵ Control calculations for the country of birth did not lead to any difference of the results.

performance were also included, in particular the experience (number of years leading a research team) and recognition (specific acts of professional recognition received since 2000).⁶

The properties of the dependent variables for team productivity (ZTOTPAP) and output quality (TOTMORC) - non-negative metrics - allow use of ordinary least squares (OLS) estimation. Residuals were tested for heteroscedasticity on team size using the Goldfeld-Quandt test and adjusted using the White estimator or by including team size as weighting variable (Greene, 2000).

TOTPAP, the number of papers listed in the SCIE database in 2001, is a non-negative integer, for which a Poisson distribution is a better approximation than the Gaussian. Count data models are known to deal efficiently with such variables. If the dependent variable is subject to overdispersion – the variance exceeds the mean – the negative binomial regression model (NEGBIN) is preferable to the Poisson model (Cameron & Trivedi, 1998). We tested for overdispersion as described in Cameron and Trivedi and include the alpha values from the NEGBIN estimation in the results tables – significant alphas indicate overdispersion. The difference between the Log-L and restricted Log-L (NEGBIN versus Poisson) was used as indicator of goodness of fit - as a substitute for the role of R² in OLS. Also, the Vuong statistic was used to test for "zero inflation" (Greene, 2000). The result was negative, indicating no need to use corrective techniques such as Zero Inflated models or Hurdle models.

Modelling the performance of research teams

Table 2 and table 3 show some of the regression models we examined.⁷ The first pair in Table 2 (Models 1 and 2) model factors affecting research output volume (TOTPAP - total number of publications in 2001) and the other models 3-6 productivity (ZTOTPAP - output volume per team member). The models in Table 3 show the results for output quality (using the MOCR indicator). Models 1, 3, 4 and 7-9 are baseline models to examine influences on team performance other than knowledge pooling such as team size or characteristics of the team leader. Models 2, 5, 6 and 10-12 are extended or full models including, where possible, variables for both modes of international knowledge pooling - origin diversity (mobility) and international collaboration. Model 1 in Table 2 shows the results of the baseline estimation for research output. Significant positive relationships can be confirmed between output and the size of the research team (TEAMSIZE) and the recognition of the team leader (RECOG). As with team size, the experience of the team leaders – measured as the number of years since attaining leadership of a team for the first time – has a non-linear but positive effect on the team's publication output: the more experienced the team leader, the higher the output. Only for higher values of experience – team leaders with many years of leadership and probably close to the end of their careers, expressed in the squared experience variable (EXPRNCE2) – does the curve slope downward again, i.e. the publication output is smaller. In the first estimations we also included a control variable for the gender of the team leader that was generally not significant, neither for TOTPAP nor the other dependent variables (see also footnote 7).

In model 2 we added to the baseline model a set of variables reflecting the different dimensions of internal structures at international level. The variables for research collaborations with foreign partners – at global level, from the EU or from the US – had to be excluded due to estimation problems. We obtain a curvilinear relationship for the origin diversity of the PhD students in the teams (ODIVPHD). The magnitude of the coefficients for the upward and downward slopes (squared variable ODIVPHD2) is similar. Hence, we can conclude that the optimum level of origin diversity for maximised output is rather low.

⁶ This was assessed through five related questions: "Since 2000, has your work been recognised in any of the following ways? Have you (a) won a scientific award, (b) been invited to serve on a major professional committee, (c) been invited to serve on the editorial board of a scientific journal, (d) organised an international conference, (e) been invited to serve on a national or international advisory committee."

⁷ In addition to the variables shown the estimations included several control variables, dummy variables for country, life sciences discipline and – only the full models – further variables on the team composition and collaboration activities by research field and by sector (industry collaboration). The results for these variables are not shown, but can be obtained from the authors upon request.

Table 3. Regression models of research output and productivity on international collaboration and origin diversity with other team characteristics

Variable	Model 1 TOTPAP		Model 2 TOTPAP		Model 3 ZTOTPAP		Model 4 ZTOTPAP ^a		Model 5 ZTOTPAP		Model 6 ZTOTPAP	
	b	t-ratio	b	t-ratio	b	t-ratio	b	t-ratio	b	t-ratio	b	t-ratio
Constant	0.862	3.936**	0.895	3.186**	0.514	4.211**	0.529	5.301**	0.364	2.844**	0.400	2.995**
TEAMAGE	0.005	0.822	0.011	1.484	0.002	0.540	0.003	1.676+	-0.29E-03	-0.077	0.002	0.803
TEAMSIZE	0.023	3.469**	0.019	2.419*	-0.019	-4.874**	-0.012	-5.749**	-0.019	-5.175**	-0.013	-5.127**
TEAMSIZE2	-0.11E-03	-1.983*	-8.4E-05	-1.404	0.11E-03	3.269**	6.1E-05	4.932**	0.10E-03	3.542**	6.1E-05	4.478**
RECOG	0.106	3.256**	0.137	3.403**	0.044	2.254*	0.024	1.504	0.029	1.468	0.015	0.839
EXPRNCE	0.038	2.318*	0.014	0.699	0.016	1.694+	0.005	0.746	0.010	1.050	-4.9E-04	-0.071
EXPRNCE2	-0.001	-2.771**	-0.001	-1.238	-0.001	-1.961+	-2.7E-04	-1.529	-0.41E-03	-1.419	-1.0E-04	-0.491
EUCPAP01	-	-	-	-	-	-	-	-	0.331	6.312**	0.230	7.028**
USCPAP01	-	-	-	-	-	-	-	-	0.206	3.684**	0.190	3.768**
ODIVPHD	-	-	0.674	1.930+	-	-	-	-	0.314	1.896+	0.378	2.633**
ODIVPHD2	-	-	-0.581	-2.002*	-	-	-	-	-0.267	-1.935+	-0.307	-2.444*
ODIVPDOC	-	-	0.122	0.362	-	-	-	-	0.043	0.257	0.106	0.664
ODIVPDOC2	-	-	-0.127	-0.486	-	-	-	-	-0.053	-0.413	-0.123	-1.118
Model type	NEGBIN		NEGBIN		OLS		OLS, weighted by teamsize		OLS		OLS, weighted by teamsize	
Alpha	0.534	7.571**	0.461	4.560**	-	-	-	-	-	-	-	-
Log-L	-1037.516		-770.4513		-	-	-	-	-	-	-	-
Rest. Log-L	-1334.189		-939.0459		-	-	-	-	-	-	-	-
F	-	-	-	-	2.62**	-	4.11**	-	6.08**	-	6.40**	-
Adjusted R2	-	-	-	-	0.076	-	0.136	-	0.348	-	0.362	-
Cases	395		296		395	-	395	-	296	-	296	-

a The Goldfeld-Quandt test returns a test value of 1.682 significant at $p < 0.01$, indicating heteroscedasticity due to the team size variable. A weighted regression may control for this.

Note: b: estimated coefficient; t-ratio: quotient of estimated coefficients and standard errors; Significance levels ** < 0.01 , * < 0.05 , + < 0.1 .

Table 4. Regression models of research quality (TOTMOCR) on international collaboration and origin diversity with other team characteristics

Variable	Model 7, OLS unweighted ^a		Model 8, OLS weight=teamsize		Model 9, OLS weight= teamsize		Model 10, OLS unweighted ^a		Model 11, OLS weight= teamsize		Model 12, OLS, weight= teamsize	
	b	t-ratio	b	t-ratio	b	t-ratio	b	t-ratio	b	t-ratio	b	t-ratio
Constant	5.953	3.345**	5.516	3.489**	5.647	3.904**	7.382	3.364**	6.372	2.696**	5.789	2.920**
TEAMAGE	0.013	0.230	-0.012	-0.370	-0.015	-0.531	0.023	0.350	-0.001	-0.035	-0.007	-0.201
TEAMSIZE	0.007	0.134	-0.031	-0.928	-0.022	-2.815**	0.014	0.219	-0.022	-0.555	-0.012	-1.233
TEAMSIZE2	-1.88E-04	-0.405	0.61E-05	0.306	-	-	-1.36E-04	-0.273	-1.21E-04	0.488	-	-
RECOG	0.172	0.604	0.336	1.207	0.326	1.178	-0.287	-0.855	-0.029	-0.084	0.052	0.153
EXPRNCE	0.135	0.970	0.119	1.137	0.082	1.737+	0.101	0.613	0.172	1.262	0.074	1.259
EXPRNCE2	-0.004	-0.838	-0.001	-0.339	-	-	-0.003	-0.558	-0.003	-0.707	-	-
EUCPAP01	-	-	-	-	-	-	1.377	1.533	1.323	1.802+	1.395	1.868+
USCPAP01	-	-	-	-	-	-	2.477	2.584*	1.807	1.713+	1.684	1.672+
ODIVPHD	-	-	-	-	-	-	-2.990	-1.056	-3.938	-1.693+	-1.181	-1.299
ODIVPHD2	-	-	-	-	-	-	1.710	0.724	2.563	1.283	-	-
ODIVPDOC	-	-	-	-	-	-	-2.716	-0.958	-1.978	-0.626	1.596	1.726+
ODIVPDOC2	-	-	-	-	-	-	3.602	1.629	2.728	1.012	-	-
F	2.01**		3.58**		3.99**		2.00**		3.28**		3.65**	
Adjusted R2	0.054		0.128		0.133		0.096		0.193		0.189	
Cases	353		353		353		296		296		296	

a The Goldfeld-Quandt test returns a test value of 2.755 significant at $p < 0.01$, indicating heteroscedasticity due to the team size variable. A weighted regression may control for this.

Note: b: estimated coefficient; t-ratio: quotient of estimated coefficients and standard errors; Significance levels ** < 0.01 , * < 0.05 , + < 0.1 .

Models 3-6 used a different dependent variable, namely the number of papers published in 2001 and listed in the Thomson ISI database divided by the number of staff in the team (ZTOTPAP). The normalised publication output is a continuous variable and we therefore changed the estimation method to OLS regressions. The results consequently differ somewhat to those using models 1 and 2. Still, model 3 – without the variables on team diversity and collaboration – is similar to model 1 when it comes to the direction of the relationship (the signs of the coefficients) except for the two team size variables. Their effect is exactly the opposite of that found with models 1 and 2: the larger the team size, the smaller the number of published articles per team member. The squared variable (TEAMSIZE2) points to a curvilinear effect. Goldfeld-Quandt and Breusch-Pagan Tests point to heteroscedasticity of the residuals caused by the TEAMSIZE variable. Therefore we also estimated a weighted model. In this model 4 the coefficients for the recognition and experience variables are not significant.

If we add the variables on international orientation and further control variables, the quality of the model improves considerably: the adjusted R-squared increases to 0.35 (unweighted model 5) and 0.36 (weighted model 6). We also see that the effect of collaboration with scientists from both other EU countries (EUCPAP01) and the US (USCPAP01) is positive and highly significant (Models 5 and 6). As for the non-normalised total number of papers, the relationship between normalised output and the origin diversity of PhD students is curvilinear with a low level optimum (ODIVPHD and ODIVPHD2 are of similar magnitude). This indicates that low origin diversity of PhD students is conducive to research output.

The models shown in table 3 use the Mean Observed Citation Rate MOCR as the explained variable. The MOCR can be considered as an indicator of the quality of publications. Again the tests point to heteroscedasticity of the residuals and we show the results of weighted estimations (Models 8, 9, 11, 12) in addition to the standard OLS models. The estimation with the restricted variable set (Models 7-9) show a negative effect of the team size; in this case it is linear, as the comparison between models 7/8 and 9 reveals. Papers of large teams are less often cited than papers of smaller teams. A slight positive effect of the team leader's experience can also be shown, but only if the squared variable that intends to measure non-linear effects is excluded (Model 9). In the full models (Models 10-12), we obtain positive coefficients for the variables assessing international collaborations (EUCPAP01 and USCPAP01). The results for the diversity indexes are not consistent in the models of table 3.

The estimated models provide a variety of results which corroborate some of the previous findings in the bivariate analyses and specify others:

- We find that only the origin diversity of PhD students exhibits a quantifiable relationship to the number of publications. The relationship is curvilinear as already suggested in Figure 1. with other factors kept constant, the publication output is highest for teams with moderate PhD student origin diversity and lower for teams with high, low or zero diversity.
- As with previous studies (Arora et al., 1998; Bordons et al., 1996; Cesaroni & Gambardella, 2003), we find a positive relationship between international collaborations and research productivity. We also confirmed other work in regard to the positive effect of international collaborations on the impact of research papers (Adams et al., 2005; Glänzel, 2001; Narin et al., 1991; Persson et al., 2004).

A further remark on team size is appropriate: this was included as a control variable and has been discussed controversially in previous research (see the reviews in Bonacorsi & Daraio, 2005; von Tunzelmann et al., 2003). We found an inverse relationship between size and performance and an optimum team size of only a few team members (the maximum average publication per capita is reached for teams with 7 members). This contradicts the expectation voiced by Bonacorsi and Daraio (2005) that increasing returns of size might apply at the team level, as they themselves could not find them at the level of institutes.

Summary and conclusions

Those life sciences teams appear to be most successful which have a strong domestic base but collaborate actively enough outside the country to ensure a moderate amount of external involvement in the team. Non-zero but small proportions of PhD students from domestic origins, from the EU and from further abroad are linked to the highest rates of publication. The message to research managers and team leaders is that team composition matters, and that it is indeed beneficial to integrate researchers from another country. A well-balanced team is characterised by some heterogeneity, but this should not be excessive.

Diversity provides a team with different skills, experience and cognitive frameworks which is believed to underlie the enhanced productivity we have found. At the same time diversity gives rise to additional costs, as people from different cultural backgrounds may speak different languages, attach different significance to concepts and research questions, and have been taught different norms for research procedures etc., placing burdens on communication and consensus formation. The negative effects might be counteracted by improvements in research management. The integration of scientists from abroad could be improved, e.g. by reinforcing mentoring schemes and allocating specific responsibility for integration of new team members.

A message to research policy-makers is that further increases in the mobility of scientists between countries are not necessarily beneficial to research performance, unless flanked by other measures. The requirement to improve integration is already well recognised, for instance, the EC Mobility Strategy specifies as an objective “*to encourage host organisations to take more responsibility for their foreign staff and visiting researchers.*” (European Commission, 2001: 11). However, if appropriate improvements in diversity accommodation cannot be made, it seems that financial support to mobility of scientists should be limited to a certain proportion per team of guest scientists, PhD students or post-docs, or otherwise spread as widely as possible over recipient teams.

Some words of caution are appropriate here in respect of specific recommendations. Our analysis uses publications and citations as metrics for the performance of research teams. However, scientific work has a number of other valuable outputs, such as new methods and tools, well-trained graduates, and knowledge or other products of use to private enterprise, the public sector and the general public (Larédo & Mustar, 2000). There may well be a trade-off between optimising levels of publication-based research output and achieving other valuable results. Clearly, our analysis of the link between diversity in a research team and publication-based research output is only valid for the output we have chosen to study.

Our recommendations also assume that some mechanism of pooling knowledge and expertise across countries underlies the relationship found here between the presence of researchers of different geographical origin in a research team and research output. However, what these underlying mechanisms are is not yet entirely clear. Education systems in different countries may give rise to ideas and perceptions, behaviours and practices etc. which are complementary in some general way. In addition, the fact that research programmes tend to be relatively homogeneous within countries may mean that mixing research staff exposed to different programmes might be a route to improving research productivity. Methods learned elsewhere for other problems may be found to be useful in the research task at hand.

Pooling knowledge is not the only imaginable link between team diversity and performance. Diversity may instead improve job satisfaction and promote stability in team composition (Williams & O'Reilly, 1998). It is also possible that mechanisms work in the reverse direction, for instance, high research productivity and visible success might well attract researchers from other countries into a team. Though these alternative hypotheses are plausible, is difficult to imagine that they are responsible for the magnitude of the effects found.

The possible underlying mechanisms for the impact of knowledge-pooling appear plausible and we therefore believe we have provided substantial evidence of the impact of pooling knowledge

internationally on research performance. However, until the underlying mechanisms are better understood, there remains some uncertainty attached to specific recommendations to research and policy decision-makers on the optimal levels of team origin diversity and international collaboration they should strive for.

References

- Adams, J. D., Black, G. C., Clemons, J. R., & Stephan: E. (2005). Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981-1999. *Research Policy*, 34(3), 259-285.
- Andrews, F. M. (Ed.). (1979). *Scientific productivity, the effectiveness of research groups in six countries*. Cambridge u.a.: Cambridge University Press, Unesco.
- Arora, A., David:, & Gambardella, A. (1998). Reputation and competence in publicly funded science: estimating the effects on research group productivity. *Annales d'Economie et de Statistique*, 49/50, 163-198.
- Barjak, F., & Thelwall, M. (2006). A statistical analysis of the web presences of European life sciences research teams. Paper presented at the 9th International Conference on S&T Indicators "New Challenges in Quantitative Science and Technology Research", Leuven, Belgium, 7-9 September 2006.
- Bonacorsi, A., & Daraio, C. (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics*, 63(1), 87-120.
- Bordons, M., & Gomez, I. (2000). Collaboration networks in science. In B. Cronin & H. B. Atkins (Eds.), *The Web of Knowledge A Festschrift in Honor of Eugene Garfield* (1 ed., pp. 197-213).
- Bordons, M., Gomez, I., Fernandez, M., Zulueta, M. A., & Mendez, A. (1996). Local, domestic and international scientific collaboration in biomedical research. *Scientometrics*, 37(2), 279-295.
- Borgman, C., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, 36, 3-72.
- Bunderson, J. S., & Sutcliffe, K. M. (2002). Comparing alternative conceptualizations of functional diversity in management teams: Process and performance effects. *Academy of Management Journal*, 45(5), 875-893.
- Cameron, A. C., & Trivedi: K. (1998). *Regression analysis of count data*. Cambridge: Cambridge University Press.
- Carayol, N., & Matt, M. (2004). Does research organization influence academic production? Laboratory level evidence from a large European university. *Research Policy*, 33(8), 1081-1102.
- Carayol, N., & Nguyen Thi, T. U. (2004). Why do Academic Scientists Engage in Interdisciplinary Research? Retrieved 7 June, 2005, from <http://cournot2.u-strasbg.fr/users/beta/publications/2004/2004-17.pdf>
- Cesaroni, F., & Gambardella, A. (2003). Research productivity and the allocation of resources in publicly funded research programmes. In A. Geuna, A. Salter & W. E. Steinmueller (Eds.), *Science and innovation rethinking the rationales for funding and governance* (1 ed., pp. 202-232). Cheltenham, UK, & Northampton, MA, USA: Edward Elgar.
- Cohen, J. E. (1981). Publication rate as a function of laboratory size in 3 biomedical-research institutions. *Scientometrics*, 3(6), 467-487.
- Cronin, B. (1984). *The citation process. The role and significance of citations in scientific communication*. (1 ed.). London: Taylor Graham.
- Earley, C., & Mosakowski, E. (2000). Creating hybrid team cultures: An empirical test of transnational team functioning. *Academy of Management Journal*, 43(1), 26-49.
- European Commission. (2001). *A mobility strategy for the European Research Area. COM(2001) 331 final*.
- European Commission. (2003). *Third European Report on Science & Technology Indicators 2003 - Towards a knowledge-based economy*. Brussels: European Commission.
- European Commission. (2005). *COMMISSION RECOMMENDATION of 11 March 2005 on the European Charter for Researchers and on a Code of Conduct for the Recruitment of Researchers*. Retrieved. from [http://europa.eu.int/eracareers/pdf/C\(2005\)576%20EN.pdf](http://europa.eu.int/eracareers/pdf/C(2005)576%20EN.pdf).
- Fiol, C. M., & O'Connor, E. J. (2005). Identification in Face-to-Face, Hybrid, and Pure Virtual Teams: Untangling the Contradictions. *Organization Science*, 16(1), 19-32.
- Georghiou, L. (1998). Global cooperation in research. *Research Policy*, 27(6), 611-626.
- Glänzel, W. (2001). National Characteristics in International Scientific Co-authorship Relations. *Scientometrics*, 51(1), 69-115.
- Glänzel, W. (2006). *Co-authorship links of life sciences institutes. Bibliometric measures of networking activities and of their impact. (NetReAct Deliverable 2.2)*. Report to the Institute for Prospective Technological Studies, Commission of the European Communities. Retrieved August 14, 2006 from: <http://www.netreact-eu.org/documents/NetreactDeliverable2.2.pdf>.
- Glänzel, W., & Schubert, A. (2001). Double Effort = Double Impact? A Critical View at International Co-authorship in Chemistry. *Scientometrics*, 50(2), 199-214.
- Greene, W. H. (2000). *Econometric Analysis* (4 ed.). Upper Saddle River, NJ: Prentice Hall.
- Hagstrom, W. (1965). *The scientific community* (1 ed.). New York: Basic books.

- Harsanyi, M. A. (1993). Multiple authors, multiple problems - bibliometrics and the study of scholarly collaboration: a literature review. *Library & Information Science Research*, 15, 325-354.
- Herbertz, H. (1995). Does it pay to collaborate? A bibliometric case study in molecular biology. *Scientometrics*, 33(1), 117-122.
- Hollingsworth, J. R., & Hollingsworth, E. (2000). Major Discoveries and Biomedical Research Organizations: Perspectives on Interdisciplinarity, Nurturing Leadership, and Integrated Structure and Cultures. Retrieved 08 March, 2005, from <http://www.umu.se/inforrk/universitetsligan/hollingsworth.html>
- Jeoh, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why differences make a difference: A field study of diversity, conflict, and performance in workgroups. *Administrative Science Quarterly*, 44(4), 741-763.
- Jöns, H. (2003). *Grenzüberschreitende Mobilität und Kooperation in den Wissenschaften: Deutschlandaufenthalte US-amerikanischer Humboldt-Forschungspreisträger aus einer erweiterten Akteursnetzwerkperspektive*. Heidelberg: Department of Geography.
- Jöns, H. (2006). Internationale Mobilität von Wissen und Wissensproduzenten. In *Tagungsband 55. Deutscher Geographentag Trier 2005*.
- Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1), 1-18.
- Larédo:, & Mustar: (2000). Laboratory Activity Profiles: An Exploratory Approach. *Scientometrics*, 47(3), 515 - 539.
- Laudel, G. (2001). Collaboration, creativity and rewards: why and how scientists collaborate. *International Journal of Technology Management*, 22(7/8), 762-781.
- Leeuwen, T. N. v., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & Raan, A. F. J. v. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics*, 51(1), 335-346.
- Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology, & Human Values*, 17(1), 101-126.
- Narin, F., Stevens, K., & Whitlow, E. S. (1991). Scientific collaboration in Europe and the citation of multinationally authored papers. *Scientometrics*, 21, 313-323.
- National Science Board. (2004). *Science and Engineering Indicators 2004* (1 ed.). Arlington, VA: National Science Foundation.
- Persson, O., Glänzel, W., & Dannell, R. (2004). A relational charting approach to the world of basic research in twelve science fields at the end of the second millennium. *Scientometrics*, 55(3), 335-348.
- Raan, A. v. (1998). The influence of international collaboration on the impact of research results. *Scientometrics*, 42(3), 423-428.
- Raan, A. v. (2003). The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments. *Technikfolgenabschätzung - Theorie und Praxis*, 12(1), 20-29.
- Stephan: E., & Levin, S. G. (2001). Exceptional contributions to US science by the foreign-born and foreign-educated. *Population Research and Policy Review*, 20(1), 59-79.
- Thelwall, M. (2003). Web use and peer interconnectivity metrics for academic web sites. *Journal of Information Science*, 29(1), 1-10.
- Thorsteinsdóttir, O. H. (2000). External Research Collaboration in Two Small Science Systems. *Scientometrics*, 49(1), 145-160.
- von Tunzelmann, N., Ranga, M., Martin, B., & Geuna, A. (2003). *The Effects of Size on Research Performance: A SPRU Review*.
- Wagner, C. S. (2005). Six case studies of international collaboration in science. *Scientometrics*, 62(1), 3-26.
- Williams, K. Y., & O'Reilly, C. A. (1998). Demography and diversity in organizations: A review of 40 years of research. *Research in Organizational Behaviour*, 20, 77-140.

A Modular Sequence of Retrieval Procedures to Delineate a Scientific Field: from Vocabulary to Citations and Back

Elise Bassecoulard*, Alain Lelu**,** and Michel Zitt*,****

*^{***}*bassecou@nantes.inra.fr*

*INRA, Lereco, F-44316, (France), **INRA, Crebi, F-78352, Jouy-en-Josas (France)

^{*}*alain.lelu@jouy.inra.fr*

Université de Franche-Comté, LASELDI, F-25000 Besançon (France)

****^{***}*zitt@nantes.inra.fr*

Observatoire des Sciences et des Techniques (OST), F-75015 Paris, (France)

Abstract

This communication presents a modular arrangement of lexical and citation operations to achieve a satisfactory delineation of a scientific field. Three querying methods are considered: on journals, on articles vocabulary, and on citation network. Rather than associating these querying modes, we consider possible sequences that make the best use of each method. At any stage of iteration, the retrieved set can be enhanced by either a citation-based analysis, or a vocabulary analysis (relative dictionaries), in order to reduce silence or noise. General noise-reduction techniques, such as clustering, are applicable at various points of the procedure. A particular sequence on a complex field (nanosciences) is described, starting with journal and lexical query, then applying a citation expansion with a final lexical adjustment. Another sequence is sketched, starting with an additional journal-based procedure.

Keywords

bibliographic coupling; nanosciences; citation flows; language processing; term extraction

Introduction

Analysis of leading edge, emerging or complex fields for science policy purposes is conditioned by the quality of field delineation (van Leeuwen et al., 2001). For example, the value of bibliometric indicators depends on the recall and precision qualities of the considered set. Macro-level delineation, based on big molecules represented by whole journals, is not costly but limited by Bradfordian conditions (see for example Rinia, 1993). Things may be particularly unfavorable in emerging fields when concentration of sources distribution is unusually low.

To get a fine-grain delineation, one has to work at the article-level, in other words to turn to information retrieval and data-mining techniques that can achieve satisfactory recall and precision. The rich structure of scientific articles or similar codified items (patents to a certain extent) creates networks of different kinds -- lexical, citation, authoring, journals -- each of them offering a target to information retrieval techniques. Garfield citation indexing (Garfield, 1967) was a revolution in this respect, breaking the monopoly of lexical forms of retrieval. Progress in data-mining techniques, especially mapping and clustering, provide additional opportunities, especially for ex post noise reduction. Besides, iterative procedures are the rule, from spontaneous web-surfer's behavior on interactive engines, to learning algorithms developed for specific purposes.

With hybridization, iteration is not limited to operations in a single network, such as vocabulary linkages. Google, the most famous retrieval system world-wide, is heterogeneous, with a lexical entry and a probabilistic navigation amongst web linkages (Brin&Page, 1998), formally analogous with citation linkages. In more conventional bibliometrics, hybridization opens up a wide space to design efficient retrieval processes.

We here propose an example of operations chaining for the delineation of scientific fields, based on lexical and citation procedures; this, within a general scheme of association of techniques. In section I,

methods, we will describe both the general framework and the particular chaining adopted in a recent study. In section II, we report some quantitative results, before discussion and conclusion.

Methods

General framework

The elements of a scientific article (and the corresponding bibliographical notice) inscribe it in a collection of bibliometric networks, especially: actors (authors/institutions) giving access to collaboration and mobility networks; linguistic contents, usually processed in the form of terms lists (title, abstract, full text terms); list of references, opening towards citation networks of many kinds; and, as metadata, bibliographic notice with journal identification (or other media), date, position in journal.

Delineation of scientific fields can be seen as a large-scale information retrieval exercise on scientific databases. Delineation protocols can iteratively mobilize procedures belonging to the four families above, starting with some initial input. In Fig. 1A the external circle figures input or supervision by experts, possibly active in each procedure. The initial input, often based on experts' advice, may be a selection of journals, key-words, key-authors or institutions, or key cited works. Bibliometric teams with an experience of micro-level studies have reported a variety of combinations (for example Debruin, 1993; Aksnes, 2000). Typically, each procedure takes the retrieved set in its existing state, analyses its contents comparatively to the universe of reference (all science in this study, with the Web of Science as a proxy), and formulates a new query, in a wide sense, to enrich or restrict the set.

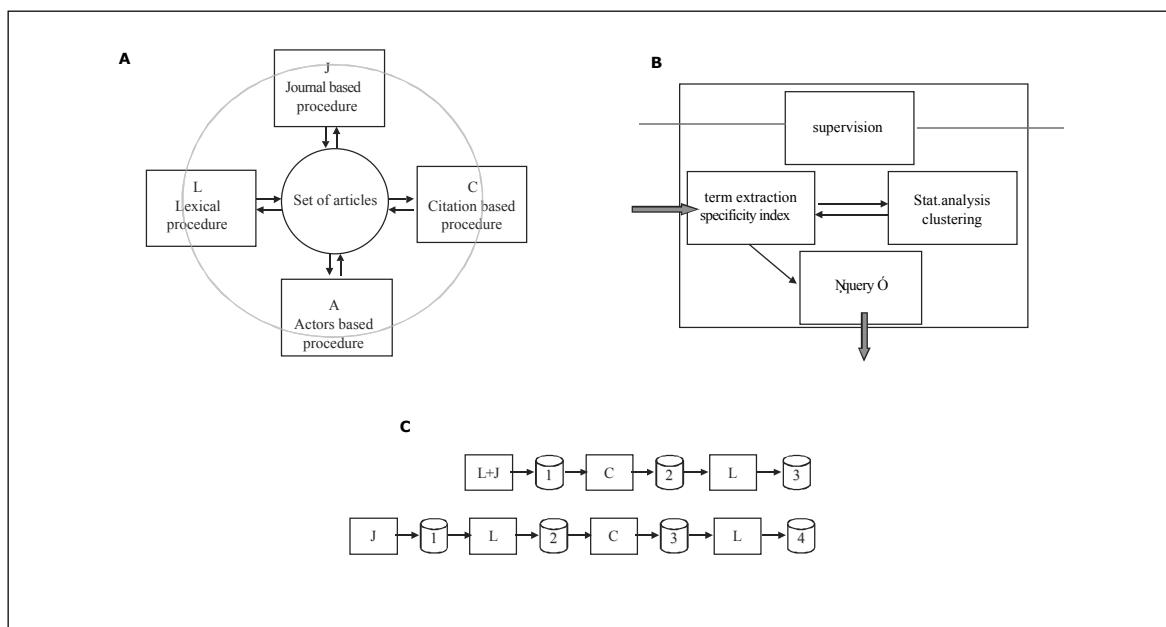


Figure 1. Retrieval procedures for field delineation

The chaining of procedures is variable, depending on the type and budget of the study, the existence and type of previous "seed" datasets and the features of the field. Here we will consider a particular instantiation in a study of Nanosciences and mention some other possibilities.

Within a particular procedure, the chaining of operations is quite similar (Fig. 1B). The entry consists of some statistical analysis of the current set of articles in the family considered, e.g. vocabulary. The set of articles contains richer information than the one contained in the query: for example some terms, specific of the set (compared to all science) that appear in collected articles, were not included in the

query. The list of such terms, with a possible weighting scheme among IR classic solutions, gives a foundation to improve or reformulate the previous query, in order to extend and/or restrict the set.

Noise contained in the set can be traced by several means, using experts' evaluation: assessment of articles ranking by IR score, suggesting thresholds; data analysis, especially mapping and clustering of the articles and/or structuring item (e.g. terms), which highlight themes and border or out-of-scope areas.

The principle of gradual delineation is quite old in IR, the hybridization of techniques is now topical, benefiting from the soar of data-mining techniques. Some association of lexical and citation indexing was already present in early ISI developments ("keyword+"), and a few combinations/comparisons of structuring techniques are reported in literature (Braam et al., 1991). Bibliometric networks, as investigated by many scholars (Kostoff, 2001), are an appealing object for data-mining. In this study however, we address bulk processing of datasets more than data navigation.

In the Nanosciences project, the chaining of procedures is as follows (see also Fig.1C):

- *Step 1.* Starting point: lexical procedure. The starting point was a lexical query, based on experts' advice, by CWTS and FhG-ISI (Noyons et al., 2003). We slightly modified the query by a complementary work and some other inputs (Meyer et al., 2001). The formulation has to comply with the possibilities of retrieval softwares. The query leads to a first "seed" set. We will not expand on this classic procedure.
- *Step 2.* Citation-based procedure. We recall shortly the principle of the protocol in the next subsection. Citation indexing and matching is much easier. Longer lists are usually required, but with no need for supervision. These features rather designate citation approach as an excellent complementary method. The citation procedure is first meant to enrich the set, but can be used in addition to discard marginal items from the initial set.
- *Step 3.* Final smoothing: lexical procedure, on which we focus later on.

Step 1 and 2 are detailed in Zitt & Bassecoulard (2006). Some results of the global chaining are reported in section II

Citation-based procedure

1) principle.

As stated above, the gradual delineation process supposes a contents analysis of the set of literature in its current state, in order to collect a "sister literature" with some proximity to this set. In a citation based procedure the proximity to the current set of articles may be defined in several ways by some path selection in the network, if we limit to paths of length 1 and 2:

- 1 - current set in citing position --> articles cited by the set
- 2 - current set in citing position --> articles citing those cited by the set
- 3 - current set in cited position --> articles citing the set
- 4 - current set in cited position --> articles cited by those citing the set

The way 2 (which implies the stage 1) is the more appropriate one for our purpose. In a Mertonian interpretation, we try and collect the "sister literature" which refers to the same "intellectual basis" (cited articles). To achieve it, we identify the literature cited by the current set, and apply a selection implying three parameters, detailed below.

The global rationale is bibliographic coupling, but instead of computing coupling links at the document level, a rather heavy task in terms of computer requirements, we considered the current set of literature as a macro-document. A particular advantage is that articles may be retrieved without having much proximity with any particular article in the current set, but because they contain other combinations of structuring items (here cited articles of the current set.).

The procedure involves three parameters, two on the cited side, one on the citing side.

- y , *genericness*, defined for each article of the cited set, is the "local" citation score, i.e. the citations retrieved from the current set of literature. The higher Y (the threshold on y) is, the most selective we want to be, discarding marginal cited items and thematic areas.

- u , *specificity*, defined for each article i of the cited set, is the ratio of local citations y and global citations y' , namely citations received from the whole Web of Science (WoS). U denotes the threshold on u . If U is set to low value, non specific cited articles, such as general physics or chemistry in the "Nano" application, are kept, likely to produce noise. Limiting to high U does not allow escaping from the limitations of the current set. A probabilistic variant of u , u^* , is obtained by dividing u by its expected value, which is the ratio of the sizes of the local citing set (the current literature) and the whole WoS.

- x , *relevance*, defined for each article j of the citing set, is the number of references that this particular article places in the subset of cited articles with y equal or superior to threshold Y and u equal or superior to threshold U . x (correspond threshold: X) is the simplest proxy for *relevance*.

To sum up, the principle is to recall relevant citing articles with at least X references to the "intellectual basis", i.e. the core of cited articles having a given level of genericness Y and specificity U .

A first analysis of interplay of Y and X was studied in (Zitt et al., 2003), linked to the vulnerability of the citation network to the removal of less connected nodes (Zitt & Bassecoulard, 2006). This stylized process can be adjusted with proper combination of the three parameters in retrieval scores.

2 retrieval score.

The list of cited articles from the "intellectual basis" can be seen as a query of many terms (cited articles), applied to the global set (WoS). We observe a loose connection with a TF-IDF scheme, considered at the current set level, TF being the frequency in the current set (in-set frequency and not in-doc frequency). For cited references, in-doc frequency is conventionally reduced to a Boolean in the WoS, so only presence counts are considered.

Table1. Vector-space analogy:

Citation process	IR analogy		Terms (here cited documents)			
			.	o		
<i>cited document k</i>	term					
<i>core of selected cited articles</i>	List of OR terms: query		1	k		n
<i>citing articles</i>	Documents to retrieve	in-doc frequency (Boolean)* = presence of terms	P(1)	..	P(k)	..
Weighting elements					P(n)	
# citations in the current set (y)	Term frequency in current set	in-set frequency (integer)	y(1)		y(k)	y(n)
# citations in the global set (y')	Term frequency in global set	in-set frequency (integer)	y'(1)		y'(k)	y'(n)

* In lexical applications, in-documents term-frequency can be used instead of presence count.

The parameters may be combined into a retrieval score for citing articles

$$s1 = \sum_{i=1,n} u_i z_i$$

where n is the number of bibliographic references in the citing article, u_i the specificity of the reference i , and $z_i=1$ for any reference present in the article.

If no weighting is practiced:

$$s2 = \sum_{i=1,n} z_i = x$$

Retrievable or retrieved articles may be ranked by s_1 or s_2 . The parameter y was not used here in the retrieval score, it could however be considered. It can be noted that cosine distance would involve the product of norms as a divisor. We did not include a rationale of fractionation of references in our exercise.

Retrieval scores could also be used to discard articles from the initial set: not meeting the threshold of retrieval score in terms of citations (X or X weighted).

In practice, a threshold must be set *ex ante* on Y to limit noise and keep the cited set within a reasonable size (a complementary pre-selection on a low value for U may be necessary for computing constraints). Several strategies can then be chosen:

- fixing thresholds U , X *ex ante*
- fixing U *ex ante* and use the non-weighted ranking s_2
- use the weighted ranking s_1

3 mapping

A map of the field was conducted, based on a rationale of bibliographic coupling. The map and clustering made use of the Neuronav software (see below). Details can be found in Bassecoulard et al., 2007). The main purpose was to analyze the differentiation of Nanosciences research themes, but this also allowed us to enlighten some borders of the field, such as the *Mesoporous structures* area. Outcomes are discussed in section II.

Mapping at various scales is a particularly efficient means to detect "clustered noise" that can arise from the many limitations of lexical queries. The same type of map would also have identified such a noisy area as *Nanoplankton*, had it not been removed at the first lexical stage.

Lexical procedure

We do not discuss here the use of the lexical procedure as a starting point, but as a procedure taking an existing set and aiming at improving it. In the implementation studied here, it took place after completion of the previous citation-based stage.

1 principle and retrieval scores

The principle is quite similar to the citation case: analyzing the contents of the current set in order to collect a sister literature sharing the same contents. Again, the general idea is that this set contains richer information than the query that built it: for example terms co-occurring in seed articles may be considered for addition in a next version of the query; or articles with some degree of lexical proximity to seed articles may be brought in the set. Although calculating proximities for each pair of words or articles is quite feasible nowadays on large volumes by graph analysis and data-mining software, we considered the global approach, in line with the method described above for citations.

The equivalent of the cited set is here the vocabulary of the current set. Although some language processing techniques tend to favor low-frequencies terms (for example textual applications of correspondence analysis, based on chi-square distance), it is generally considered that low-frequency terms carry various flaws, spelling errors, and moreover have a low structuring effect. It seems wise to apply a threshold Y on term frequency.

Similarly, the specificity of terms is of great importance, measured by the same type of index u as above: ratio of local frequency y (in the current set) and global frequency y' (in the WoS). Building this list is quite critical, as we will see below.

The equivalent of x is the number of terms in a candidate article, above thresholds on y and u or particular combinations. Any IR weighted formula may be applied, involving x and u , and possibly y . A difference with the citation approach is that in-document frequencies are integers and not Booleans,

hence two definitions of frequencies are applicable at the document and set levels: “presence” (Boolean count) and “frequency” (integer count). In this study we mostly used the presence count.

2 unification of vocabulary

The main difference with the citation way is the issue of relevant term extraction and vocabulary unification. Citations do not need more than a careful matching, once admitted that a complete recall is out of range (misspellings, etc.). Except when one limits oneself to controlled language and corresponding databases (MESH or CAS thesaurus for example), these issues are critical. The frequencies and especially ratios mentioned above only make sense if terms are sufficiently unified. Bibliometrists are confronted with the numerous issues of natural language: flexion, misspelling, homonyms, synonyms and acronyms. Natural language processing has inspired a vast literature boosted by the necessity to address very large datasets on the web. The scientific vocabulary adds particular difficulties: inflation of terms, for example in natural classifications; prevalence of accretion processes leading to complex multi-terms, uniterms often lacking precise signification; multiplication of acronyms and coded forms, difficulties in automatic filtering of irrelevant multi-terms, useless for subsequent human or automatic processing.

We used here the commercial software Neuronav, initially designed as an interactive term-cleaning and clustering environment (Lelu, 1994; Lelu & François, 1992). Diatopie (S. Aubin, <www.diatopie.com>) enhanced it with a basic, but robust and efficient term extracting sequence, and developed it up to industrial standards. Term extraction first identifies non-trivial uniterms, removes “frozen phrases” (such as “*all things considered*”) and tags each remaining character string in one of five grammatical categories, noun, verb, etc. This sequence relies on a statistical simplification¹ that allows the processing of large datasets with a good tagging precision, as shown by Vergne (Vergne, 2001). In a second step, phrase syntactic patterns specific to noun-phrases are extracted to form multi-terms. Only nouns and noun-phrases have been kept as content-markers.

The treatment was complemented by procedures developed by Lereco lab (dedicated PERL and SAS programmes), namely: further unification of spelling (case, hyphens etc), unification of English and American forms, further unification of singular and plural forms. The outcome of the unification stage is a correspondence list between unified terms and variants forms, as they were extracted and processed from the local file (“Nano” current dataset).

Then, possible variants in natural language have been generated. Some examples are shown on Table 2. On this base, the frequency of unified forms both on the local “Nano” and the global file “WoS” can be calculated. The practical implementation depends on the RDBMS and related querying language. This allowed us to calculate the specificity u of each unified form in a quite similar way as above for citations: ratio of local to global frequency. Presence count was privileged. The probabilistic index u^* is again derived by dividing u by its expected value, the ratio of the local (current literature) and global datasets (all WoS).

Specificity is a very powerful means to obtain a relevant list of terms, candidates for addition in the query if intended. These lists suggest some scrutiny and supervision. The integration in the final query of low frequency words, even with high values of u , is questionable. The threshold Y may be modulated to avoid managing too long a list, and some final selection could perhaps be envisaged before improving the query. All things equal, new words with high frequency promise a good recall, and u promises a good precision, at the expense of the recall.

¹ “one string, one tag” statistical simplification:: attributing as default the most frequently observed syntactic tags to homographs may lead to a tagging precision, in the context of the GRACE taggers evaluation (Vergne & Giguet 1998), as high as 92.1% for gross 9-categories tags, and 82.5% for 311-tokens tags.

The intermediary outcome is the list of unified terms, ranked by specificity, possibly weighted and qualified for expert supervision. An experiment has been done to try weighting schemes that place the words of the initial query on top-rankings: u not weighted, u.y, u.sqrt(y), u.log(y). The latter scheme, moderately weighted by frequency y, has clearly given the best result in this respect. (see results section).

Table 2. Examples of variants

Initial forms	Extracted forms	Unified form	Possible variants
nanometer-scale	nanometer-scale	Nanometer_Scale	nanometer-scale nanometer-scales Nanometer-scale Nanometer-scales
nanometer scales	nanometer_scales	Nanometer_Scale	nanometer scale nanometer scales Nanometer scale Nanometer scales
nanometre scale	nanometre_scale	Nanometer_Scale	nanometre scale nanometre scales Nanometre scale Nanometre scales
Nanometre-scale	Nanometre-scale	Nanometer_Scale	nanometre-scale nanometre-scales Nanometre-scale Nanometre-scales
nanometerscale	nanometerscale	Nanometer_Scale	nanometerscale nanometerscales Nanometerscale Nanometerscales
nano-meter scale	nano-meter_scale	Nanometer_Scale	nano-meter scale nano-meter scales Nano-meter scale nano-meter scale

The contents of the final outcome, the set of articles, depends on the retrieval strategy: a short list of terms may be extracted from the vocabulary on u or weighted u criteria, and then a Boolean OR query is set up to retrieve articles (x is then implicitly set to 1); instead, a retrieval formula combining x and u as done for citations, possibly with also with y, may be used.

3 mapping

Lexical mapping may be helpful for the query improvement that follows. It was not carried on in the present implementation.

Results

The application described here relied on a particular chaining: (a) start with lexical query; (b) expand with citation proximity; (c) finally enhance with lexical procedure. (b) and (c) are mainly meant to reduce silence, but as already stated, retrieval scores set for the expansion on citations can also be used to reduce noise in the initial dataset. However, the main tool to reduce noise is, at various stages, the clustering and mapping process, which can detect areas of noise in clusters, with respect to the level of observation.

Some arguments for this chaining are developed in Zitt & Bassecoulard (2006). Lexical queries are taken as a starting point, along with selection of specialized journals, because they are a good basis for experts' input and discussion. Moreover, due to the strong concentration of lexical distributions they possess an apparent efficiency of short lists in terms of retrieval. The shortcoming is the natural

language issues with heavy unification problems and related necessity of supervision. In the same article, the outcomes of the citation process were presented. The extent of the retrieval was observed for various combinations of Y, U, X., showing that the extension process follows classic laws of information concentration. The quasi-Bradfordian behavior does not suggest a formal optimum, unless cost functions are applied. Despite the quality of the (initial) lexical query, a large amount of relevant literature could be recovered, for example we pushed extension through 31,2% with little noise problems. The first mapping exercise reported in Bassecoulard et al. (2007), studies a breakdown in 50 themes, aggregated in 7 super-themes. The core is formed of 3 super-themes (*Magnetism, Quantum dots, Microscopy; Optical & Electro Applications; Nanomaterials*) and a satellite, *Bionano*. *Theoretical studies & Calculations* appear in central position. Opposed to the core, we find *Nano-objects* on the first factor, *Mesoporous structures* on the second one. The rate of extension was rather contrasted among themes, from 4% for particular carbon nanotubes (*Nano-objects*) to 51% for dendrimers (*Nanomaterials*).

Then why go back to a lexical procedure after the citation extension? First, citation extension brings topics that were hardly contained in the initial query: This is the case for isolate domains like *Mesoporous structures* where more than 45% of articles come from the extension process, but also for clusters in the core domains (42% of extension rate for *Optical & Electro Applications*) or in central position, like *Theoretical studies* (3 clusters out of 4 show an extension rate of 40% and more). There is no guarantee that these added areas are completely covered and a scrutiny of vocabulary is anyway helpful to describe the extension. The second reason is that vocabulary can be used to transfer queries on other databases, usually deprived from citations linkages. Last reason, IP rights rule the use of data (Thomson license for the WoS). This is not the case for exchanges of queries, the legal issues being dealt with by each user.

However, the new lexical procedure does not escape natural language issues. The choice of an efficient software for term extraction and unification is critical. In particular, the calculation of specificity would be jeopardized by a poor unification. The specificity index is the key to a reliable query formulation. Sensible choices on specificity thresholds will achieve a controllable recall, without facing noise explosion.

We will limit ourselves here to the intermediary result, namely the list of candidate terms for the query. The final outcome depends on the IR formula applied. Table 3A presents the 50 first terms – those present in the initial query are not shown -- ranked by decreasing u, Table 3B the first 50 terms ranked by u.logy. Depending on u values in the range considered, candidate terms could bring up to 20 % new articles.

The effects of expansion, either by citations or by word proximities, on the volume of retrieved articles depends on the network structure of the field and on the strategy employed, namely the particular combination of y, u, x. Beyond regular distributions à la Bradford sustaining the gradual extension process, the rate of extension is not uniform. We reported above that the percentage of extension was very contrasted amongst themes, and we can now see whether the final lexical enhancement is also necessary for particular themes.

Clearly, a high rate of extension by the citation process for a particular theme makes the next lexical stage all the more necessary, since the citation process may not entirely recover the appeared satellite themes, if any. Conversely, a low rate of extension of a theme indicates that no satellite appears, but it does not mean that the lexical enhancement will be useless, although it will probably be much less productive.

Table 3. 50 first candidate terms

A-Ranked by decreasing u Specificity	B-Ranked by decreasing u.logy Weighted Specificity
AAO_Templates	STM_Images
Alkanethiol_Self_Assembled_Monolayers	AFM_Image
Alkanethiolate_Monolayers	AFM_Tip
Armchair_Tubes	Gold_Surface
Buried_Islands	Reflection_High_Energy_Electron_Diffraction
Carbon_Onions	Wetting_Layer
Catalytic_Growth	STM_Tip
CdSe_Dots	Single_Electron_Transistor
Coherent_Islands	Mica_Surfaces
Compact_Islands	InAs_QD
Conductance_Histogram	Microcontact_Printing
Gas_Source_MBE	Dot_Size
High_Resolution_STM_Images	Quantum_Confinement_Effect
Hut_Cluster	AFM_Measurements
InAs_Coverages	Field_Emission_Properties
InAs_Deposition	Scanning_Tunneling_Spectroscopy
InGaAs_QDs	Near_Field_Optical_Microscopy
Interdot_Coupling	Scanning_Near_Field_Optical_Microscope
Interdot_Distance	Quantum_Size_Effect
Monolayer_Island	Dimer_Row
Mucoadhesive_Properties	AFM_Imaging
Ordered_Adlayer	Vicinal_Surface
Periodic_Mesoporous_Organosilicas	Gold_Substrates
Phonon_Confinement_Model	Highly_Oriented_Pyrolytic_Graphite
Photon_Scanning	Scanning_Near_Field_Optical_Microscopy
Polyelectrolyte_Multilayer_Films	STM_Observations
QD_Ground_State	Dye_Sensitized_Solar_Cells
Reflection_High_Energy_Electron_Diffraction_	Polyelectrolyte_Multilayers
Intensity_Oscillation	
Resolved_STM_Images	AFM_Observation
Root_3_Ag_Surface	Surface_Enhanced_Raman_Scattering
Self_Assembled_Alkanethiol_Monolayers	Dot_Layers
Semiconducting_Tube	InAs_Island
Simulated_STM_Images	AFM_Cantilever
Solid_Source_MBE	Near_Field_Scanning_Optical_Microscopy
Spin_Blockade	Near_Field_Optical_Microscope
Strained_Islands	Island_Density
Stranski_Krastanow_Mode	NF_Membrane
Submolecular_Resolution	Tip_Sample_Interaction
Terrylene_Molecules	Turn_On_Field
Thiol_Monolayer	QD_Structure
Tip_Sample_Contact	STM_Imaging
Tip_Surface_Interaction	Organic_Monolayer
True_Atomic_Resolution	Single_Dot
Vacancy_Islands	Layer_By_Layer_Growth
Dot_Layers	Three_Dimensional_Islands
InAs_Island	TiO2_Electrodes
SWNT_Bundles	Alumina_Template
Alkanethiol_Monolayers	InAs_Dot
Alumina_Template	Melt_Intercalation
Colloidal_Carriers	Dot_Structures

Let us take a few typical examples:

- For the super-theme *Nano-objects* such as various Carbon Nanotubes, the lexical description used in the initial query reveals itself to be, apparently, very efficient, judging by the negligible amount of extension brought by the citation process, except for the fullerene cluster (37,8% of extension). Two possible reasons: the strength of terms such as *Nanotubes* used as labels of the field; the relative

isolation of the super-theme of *Nano-objects*, and perhaps of the particular themes within the family. If so, we can even imagine that other Nano-objects areas are missed if not captured by the initial query and especially the “Nano” tokens. For *Nano-objects*, the lexical enhancement of the query is not expected to bring much improvement; We can await some quite marginal gain by including particular objects appearing in the ranked lists of vocabulary, such as *Armchair_Tubes* on top of the non-weighted list in Table 3. Composite terms from the Fullerene family should also be checked.

- The situation is quite different for *Mesoporous structures*, another super-theme, on the borderline of Nanoscience. As mentioned above, the term was not present in the initial query and *Mesoporous* terms are present among candidate terms ...
- For biological terms, the query had been restricted to avoid explosion. The *Bionano* super-theme shows an average extension rate (30,2%) Some further gain can be expected from specific terms like *DNA films*, *DNA_Sensors* etc

Another striking feature is the abundance of instrument-related terms among candidates in Table 3B. Some of them (like the various *STM_Images*) include acronym forms of the terms of the initial query and probably co-occur with them. Their contribution to the expansion is dubious, despite their high specificity. Other terms like *Scanning_Tunneling_Spectroscopy* shoud bring a further gain.

Table 4 Example on the global sequence

Super-theme	Initial lexical query	Citation process	Expected Lexical Enhancement
<i>Nano-objects</i>	Present and complete*	Little use*	Little use*
<i>Mesoporous structures</i>	Absent	Important	Important
<i>Bionano</i>	Present but restricted	Average	Average
<i>Magnetism, Quantum dots, Microscopy</i>	Present but incomplete	Average	Moderate

Except for fullerenes (restrictions)

Discussion and conclusion

We stressed in this study the powerful combination of proximity analysis in two bibliometric networks, citations and words, to achieve a satisfactory coverage of a scientific field. The network of actors, either authors or institutions, could also be mobilized, with some additional hypotheses on thematic specialization of actors, and further name unification issues.

This particular chaining of modules, starting with vocabulary -- plus a few journals -- and followed by citation with a final lexical smoothing is likely to be the most appropriate in many cases. The starting point may be different however. In the Nano study, the starting point was mainly lexical, with some specialized journals added (they were not many in the WoS for the period under study, as the field was still emerging). In an on-going study on another, more established, field, the collection of specialized journals (forming the relevant Thomson subject category codes), or any other quasi-Bradfordian collection, is much more representative of the field. The lexical content of these journals appears to be a good starting point. Using the methodology described in 1.3, we are able to translate these contents into a ranked list of specific terms, which can be used to build an efficient lexical query. Lexical mapping in the space of words or articles is recommended, in order to remove the noise coming from any cut off on the journal distribution. As to silences, the lexical query deduced, possibly enhanced by a citation enrichment stage, will efficiently contribute to remove it.

Beyond general statistical regularities of information processes, mapping may detect clusters with local particularities in their density or external connections. These local conditions rule the outcomes of expanding processes. We mentioned the case of *Nano-objects* (mainly carbon *Nanotubes*), where a sister literature, in terms of bibliographic coupling, hardly appears. Generally speaking, the higher the rate of citation extension (with the probable appearance of thematic satellites), the more productive the next lexical stage- then able to complete the delineation of these borders, with the good level of security brought about by the specificity index setting.

We addressed only bulk processing (taking whole datasets), and envisaged only one pass for each method. Adding iterations can be envisaged, but as soon as mapping and expert supervision are integrated to the steps, the multiplication of operations can be costly and difficult to manage. Similarly, we favored whole-set treatment rather than aggregating individual neighborhoods, a choice that could be advocated in some situations.

Last remark, the validation process can be envisaged from several points of view. One would be the evaluation of the method (or at least of one possible chaining) on a controlled set of documents (preferably WoS documents, as citations are compulsory) in order to measure IR performances for various threshold settings. To our knowledge, there is no such standard collection, and the issue is open. Another point is the expert validation of retrieval scores for set expansion or query expansion. We are working on the practical implementation of the process in two on-going projects that will be reported later on.

References

- Aksnes, D. W., Olsen, T. B., & Seglen, P. O. (2000). Validation of bibliometric indicators in the field of microbiology: A Norwegian case study. *Scientometrics*, 49(1), 7-22.
- Bassecoulard, E., Lelu, A. & Zitt, M. (2007). Mapping nanosciences by citation flows: a preliminary analysis, *Scientometrics*, 70(3), 859-880.
- Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991). Mapping of science by combined cocitation and word analysis. I Structural aspects. *Journal of the American Society of Information Science*, 42(4), 233-251.
- Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and Isdn Systems*, 30(1-7), 107-117.
- Debruin, R. E., & Moed, H. F. (1993). Delimitation of Scientific Subfields Using Cognitive Words from Corporate Addresses in Scientific Publications. *Scientometrics*, 26(1), 65-80.
- Garfield, E. (1967). Primordial Concepts, Citation Indexing and Historio-bibliography. *Journal Library History*, 2, 235-249.
- Kostoff, R. N., Toothman, D. R., Eberhart, H. J., & Humenik, J. A. (2001). Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change*, 68, 223-253
- Lelu, A. (1994). Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In E. Diday & Y. Lechevallier (Eds.), *New Approaches in Classification and Data Analysis* (pp. 241-248). Berlin: Springer-Verlag.
- Lelu, A., & François, C. (1992). Automatic generation of hypertext links in information retrieval systems. In D. L. al. (Ed.), *Proceedings of ECIT'92 (Milano)* (pp. 112-121). New York: ACM Press.
- Meyer, M., Persson, O., & Power, Y. (2001). *Nanotechnology Expert Group and Eurotech Data. Mapping Excellence in Nanotechnologies* (Preparatory study). Brussels: EC, DG-Research.
- Noyons, E. C., Buter, R. K., Hinze, S., van Raan, A. F. J., Schmoch, U., Heinze, T., et al. (2003). *Mapping excellence in science and technology across Europe: nanoscience and nanotechnology* (Draft Report No. EC-PPN CT 2002-0001): EC.
- Rinia, E. J., Delange, C., & Moed, H. F. (1993). Measuring National Output in Physics - Delimitation Problems. *Scientometrics*, 28(1), 89-110.
- van Leeuwen, T. N., van der Wurff, L. J., & van Raan, A. F. J. (2001). The use of combined bibliometric methods in research funding policy. *Research Evaluation*, 10, 195-201.
- Vergne, J. (2001). Parsing natural languages: from "combinatorial" to "deterministic" parsing. In D. Maurel (Ed.), *Actes de TALN 2001 (Traitement automatique des langues naturelles)* (pp. 15-29). Tours: ATALA & Université de Tours.
- Vergne, J., & Giguet, E. (1998). Regards Théoriques sur le "Tagging". In *Actes de la cinquième conference Le Traitement Automatique des Langues Naturelles (TALN)* (pp. 22-31). Paris.
- Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2003). Bridging citation and reference distributions: Part I- The referencing-structure function and its application to co-citation and co-item studies. *Scientometrics*, 57(1), 93-118.
- Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing & Management*, 42(6), 1513-1531.

Variety in Web Spheres between Research Fields: Content and Function¹

Peter Van den Besselaar*, Gaston Heimeriks**, Koen Frenken***

*p.vandenbesselaar@rathenau.nl

Netherlands Center for Science System Assessment, Rathenau Instituut & Amsterdam School of Communication Research, Universiteit van Amsterdam (The Netherlands)

**gaston.heimeriks@cec.eu.int

European Commission – Joint Research Centre IPTS, Seville (Spain)

***k.frenken@geo.uu.nl

Urban and Regional Research Centre Utrecht, Utrecht University (The Netherlands)

Abstract

In this paper we investigate the different ways the Internet and the WWW are used in different research fields. The question we address is: is the variation in use related to the type of research field – especially with the difference between basic research and more application oriented research? We compare fields from sciences, life sciences, social sciences and humanities, and among these some are more application oriented than others. The results indicate that the observed differences between the disciplinary websites are not systematically related to application orientation. We discuss other differences between research fields that may explain the use and nature of the websites.

Keywords

webometrics; indicators; research fields.

Introduction: background and research questions

Scientific change is often based on the availability of new tools and techniques, which enable researchers to explore new answers for existing questions, and to explore new questions that were not easily researched without these new *instrumentalities* (Price 1986). Information and communication technologies typically are such an innovation in research tools and data, on different levels (Van den Besselaar 2007). First of all, it changes the nature of research data: increasingly electronic data become available about everything – research becomes data intensive ever more than before. Secondly, the access to scientific data changes. The online data repositories and internet *archives* provide new ways of doing research. This type of sharing is not only restricted to data. Also repositories for papers, and other collaboration technologies become increasingly available. Thirdly, new tools for analyzing these data become available. These new research practices are increasingly based on *monitoring*, *modeling* and *mapping* (De Jong and Rip 1997; Nentwich 2003). Fourthly – and related – contemporary research focuses more on the properties and behavior of artifacts (such as computers) and artificial systems rather than on natural phenomena in the real world (Gibbons et al. 1994). These developments are not a finished project; and they are subject of science policies (ESFRI 2006). Of course the level of use of ICTs differs between fields, as some were much earlier taking it up than others (Nentwich 2003).

Parallel to this changing nature of research, also the contract between science and society has changed (Rip 2002), and the emphasis of research – in an increasing number of research fields – is increasingly use oriented. A variety of concepts have been suggested for this, such as *Pasteur's quadrant* (Stokes 1997), *mode-2 knowledge production* (Gibbons et al), or the *triple helix* (Etzkowitz & Leydesdorff 1997). In these more use oriented fields, the relation with the non-academic environment is also supported with electronic media – something that is visible in the content of the websites as in the hyperlink networks (Van den Besselaar & Heimeriks, forthcoming). In other words, the internet and

¹ The research has been partially funded by the European Commission, the EICSTES project (IST-1999-20350). Partners were CINDOC (Spain), ARCS Vienna (Austria), DTI (Denmark), INIST/CNRS (France), NIWI-KNAW (Netherlands), and the University of Surrey (UK). Eleftheria Vasileiadou and Arie Rip provided useful comments on earlier drafts of this paper, as did two anonymous reviewers.

the WWW are used in the internal academic communication and collaboration, as well as in the communication with the external network.

We do not focus on the motives for establishing hyperlinks (e.g., Wilkinson et al. 2003) but on the way hyperlinks can be used to map the communication networks within research fields and between research fields and their academic and non-academic environments (Heimeriks, Hoerlesberger, Van den Besselaar 2003; Heimeriks & Van den Besselaar 2006; Vasileiadou & Van den Besselaar 2006; Van den Besselaar & Heimeriks, forthcoming. See also Harries et al, 2004). In this perspective, the following research questions arise: Do we find empirical evidence for the existence of distinct online communication patterns across fields? What do these differences relate to? Are the web based communications related to field specific use of the Web or can we identify more general patterns? In particular, is the level of ‘use orientation’ of research fields important to understanding the use of the web by research fields? Our main hypotheses are:

- 1 use oriented ('mode-2') sciences make more extensive use of internet and Web than the fields that are only aiming at fundamental understanding without any considerations of use ('mode-1' fields);
- 2 use oriented sciences are characterized by a greater variety of outputs disseminated through the web;
- 3 use oriented sciences address a greater variety of audiences through the web.

Data and Methods

The study is based on web data about the size, content, and outlinks of the websites of universities and departments from fifteen ('old') EU member states.² Once web sites were identified and selected, some basic information was collected using software tools called 'mappers': the name of the department, the institution they belong to, and the URL that identifies them. These tools simply 'crawl' the web starting from a certain site and following the trace of its embedded links and registering the objects found in this process.³ The software program used to construct the database of European universities is Microsoft Site Analyst⁴. All URLs were classified in three ways: an institutional code that classifies the type of entity based on a survey of the higher education systems in the European Union; a geographical code using the NUTS classification (Nomenclature of Territorial Units for Statistics) of EUROSTAT; and a thematic code according to the UNESCO classification of science and technology domains. The UNESCO codes have a 3-level structure. The first two digits refer to the discipline, the third and fourth digits refer to fields, and the last two digits refer to subfields. In this study, the first 4 digits are used for the delineation of the fields.

Table 5. The distribution of the departments over the 15 EU countries

Field	AU	BE	DE	DK	ES	FI	FR	GR	IE	IT	LU	NL	PT	SE	UK	TOTAL
<i>Comp</i>	68	41	383	14	205	70	108	17	39	55	0	95	26	34	334	1489
<i>Astro</i>	3	1	2	0	7	1	1	0	1	1	0	4	4	1	15	41
<i>Bio</i>	3	3	19	5	17	4	10	1	1	4	0	9	1	5	21	103
<i>Gen</i>	7	5	34	2	18	2	5	0	4	7	0	10	2	9	39	144
<i>Hep</i>	0	2	3	0	1	1	0	1	0	1	0	6	0	0	1	16
<i>Info</i>	11	8	43	6	32	8	11	3	4	21	1	7	4	7	51	217
<i>Lit</i>	3	8	11	5	16	12	7	0	0	16	1	17	0	3	37	136
<i>Psy</i>	1	3	14	0	16	1	0	0	2	3	0	6	0	0	7	53
Total	96	71	509	32	312	99	142	22	51	108	2	154	37	59	505	2199

Astro = Astrophysics; Bio = biotechnology; Comp = computer science; Gen = genetics;
Hep = High energy physics; Inf = information science; Lit = literature studies; Psy = psychology

² Data were collected in 2003.

³ Details about the data collection and the further processing of the data can be found in Arroyo et al. (2003).

⁴ Shareware (Back Office Pack) developed from Webmapper.

The data consists of website characteristics and outlinks that enable us to construct disciplinary hyperlink networks between departments. The fields differ in terms of the relevant dimensions of knowledge production: we selected fields with more and with less use orientation (mode-2 versus mode-1) and we included fields from the humanities, the social sciences, and the natural sciences: High Energy Physics, Astrophysics, Genetics, Biotechnology, Computer Science, Information Science, Literature Studies and Psychology. Table 1 shows the distribution of the departments over the 15 countries. The distribution over countries is reasonable (but we do not claim representativity).⁵

This study consists of several steps. We will start (section 4) by analyzing the outlink patterns of the websites, and focus on shared outlinks in the fields, using only unique outlinks.⁶ We analyze the composition of the outlink network, and compare the international and domestic networks. Then (section 5) we analyze the content of the websites. Finally (6) we analyze the relation between the size and content of a website, and the number of inlinks a website receives from other departments in the field: the academic impact of websites. Do the websites indicate the nature of knowledge production, the context of application, and the importance of web based communications to relevant audiences? Can we establish general patterns in web-based communications of scholarly departments or are field-specific patterns visible?⁷

Disciplinary outlink patterns

For each field, the 100 most frequently hyperlinked organizations are classified in the categories university, *publishers and journals*, *governmental organizations*, *companies*, *professional organizations*, *research organizations*, *data repositories* and *archives*. Differences exist in the frequency distributions, but it is also clear that the different types of organization are visible in all fields. It can therefore be argued that the internet maintains similar *networks* in all fields, albeit with rather *different compositions*, as large differences in the share of the different groups are visible (table 2).

Table 2. The distribution of different types of linked organizations

% outlinks to (by field)	Comp	Astro	Bio	Gen	Hep	Info	Lit	Psy
<i>Companies</i>	42	15	33	21	25	30	27	34
<i>Publishers</i>	10	13	14	24	15	18	8	13
<i>Universities</i>	29	36	29	28	23	28	46	43
<i>Research organizations</i>	2	5	0	5	11	1	0	0
<i>Professional organizations</i>	8	9	9	4	9	7	2	3
<i>Governmental organizations</i>	4	16	13	3	7	13	10	4
<i>Archives, Data repositories</i>	3	4	2	13	5	3	4	3
<i>Other</i>	2	2	0	2	5	0	3	0

Cells: percentage of the links

Not unexpectedly, in most of the mode-1 fields, universities are largest category in the outlink environment, whereas in most of the mode-2 fields companies are the largest category. In all fields we find links to software companies and to internet providers, but links to companies are most visible within Computer Science and Biotechnology. Publishers and journals are especially well represented in Genetics and Information Science and under-represented in Computer Science and – unexpectedly – in Literature Studies. Governmental organizations are most occurring in Astrophysics, Information Science and Biotechnology, while data-repositories are most important in Genetics. Archives only occur in Computer Science, Genetics, Astrophysics and High Energy Physics. Data repositories occur in all fields, but in Genetics they are a rather large category. Also the intensity of the hyperlink

⁵ Because of lack of space we do not describe the different fields in detail.

⁶ Also if a department has more links to an organization, this only counts for one relation. In this way, links represent (unvalued) relations between organizations

⁷ The dataset contains *all outlinks* from, but only the *academic inlinks* to the departments. We therefore cannot analyze non-academic audiences, nor whether size and content relate to visibility for non-academic audiences.

relations differs considerably between the fields (Table 3). The intensity of outlinks is largest in Astrophysics and High Energy Physics. If we look in the other direction, we find that link relations are the strongest with universities, companies, and publishers.

Table 3. Average number of outlinks per department by different types of organizations

Links per department to:	Comp	Astro	Bio	Gen	Hep	Info	Lit	Psy
<i>Companies</i>	4,6	3,6	1,5	1,5	5,0	2,6	2,4	2,6
<i>Publishers</i>	1,0	3,5	0,7	1,9	3,3	1,7	0,6	0,8
<i>Universities</i>	2,3	8,8	1,1	1,7	4,3	2,1	3,5	2,5
<i>Research organizations</i>	0,1	1,1	0,0	0,3	2,8	0,1	0,0	0,0
<i>Professional organizations</i>	1,1	2,4	0,4	0,3	1,9	0,6	0,2	0,2
<i>Governmental organizations</i>	0,5	4,0	0,6	0,2	1,5	1,0	0,9	0,4
<i>Archives, Data repositories</i>	0,2	1,0	0,1	1,0	0,6	0,2	0,5	0,2
<i>Average</i>	1,2	3,1	0,6	0,9	2,4	1,1	1,0	0,8

The next question is whether *individual organizations* are prominent in disciplinary hyperlink environments ('preferential attachment' or 'codification')? Despite the lack of clear mechanisms structuring hyperlink behavior, differences exist. Links to internet-related companies such as Google are excluded in order to focus on field-specific outlinks. Figure 1 shows the distribution of the share of the departments in a field (on the y-axis) linking to the same organization (on the x-axis). Astrophysics shows the highest level of 'preferential attachment': Ten organizations in the environment of the field receive links from more than 30% of the European departments in Astrophysics included in this study. On the other side of the spectrum we find Information Science, Literature Studies, Genetics, Psychology and Biotechnology where no organization exists that receives links from more than 15% of the departments.

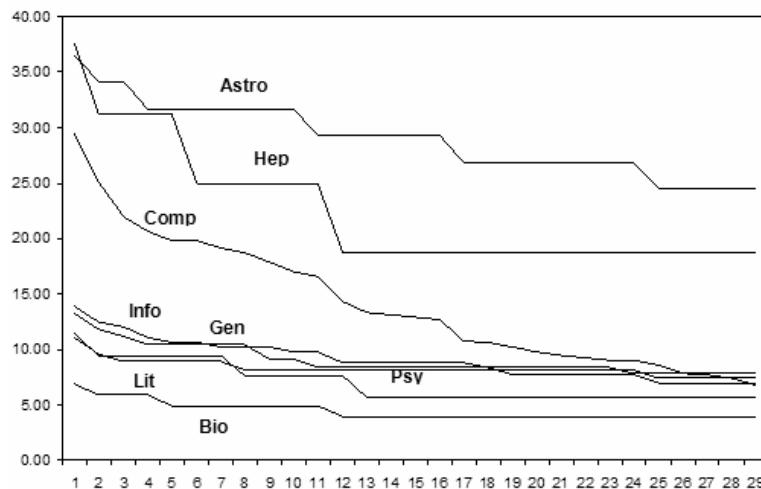


Figure 1. The distribution of most frequently linked websites in selected fields.

What kinds of websites receive many links? The 'codification' seems to depend on data sharing, on the role of governmental and professional organizations and of scientific publishers in the environment, and in some fields on links to important university departments. More specifically, we find that links to peer departments are important within Computer Science, Astrophysics, and High Energy Physics. In the latter two, and also in Genetics, global data repositories receive links from a large number of departments, and important global information repositories (NCBI, NASA and CERN respectively) clearly are the most frequently linked organizations. Shared links to companies are relatively

important in Computer Science, in Information Science and in Biotechnology. Publishers (Springer) and journals (Nature and Science) are relatively important in Genetics.

Table 4. The outlink pattern of the random samples compared to the top-100*

Relative in random sample:	Comp	Astro	Bio	Gen	Hep	Info	Lit	Psy
<i>Companies</i>	1,2	2,1	1,5	1,5	1,2	1,4	1,6	1,2
<i>Publishers</i>	1,1	0,8	0,8	1,2	1,1	0,8	0,7	0,9
<i>Universities</i>	0,6	0,8	0,7	0,6	0,9	0,8	0,8	0,8
<i>Research organizations</i>	1,6	1,0	1,2	1,3	1,4	1,5	1,1	1,3
<i>Professional organizations</i>	2,8	1,5	-	1,2	0,8	8,0	-	-
<i>Governmental organizations</i>	0,5	0,6	0,5	1,0	0,7	0,2	1,8	2,0
<i>Data repositories</i>	2,1	0,6	0,6	0,8	1,5	0,8	0,8	0,5
<i>Archive</i>	1,0	0,4	-	-	0,6	-	-	-

* Index: the top 100 links = 1

Is the the top 100 outlinks different from outlinks in general? We randomly selected some 100 unique outlinks per field, and compared these with the top 100 lists (table 4). Generally, we find in the tail a higher share of links to universities and companies, but a lower to publishers, professional organizations, and data repositories. Interestingly, in Literature and Psychology we find a higher share of (national) professional organizations in the tail. Possibly in these fields, the focus of research is more local in orientation, which may be reflected by the presence of local professional organizations in the outlink environment.

Does local orientation differ between the fields? The more a field is oriented on a national and local context, the more it relies on local resources and dissemination outlets. We analyzed this for the Netherlands. Computer Science and Information Science have the lowest share national outlinks (19% and 20%) and Astrophysics and High Energy Physics slightly more (22% and 26%). Psychology and Literature are more locally oriented (28% and 34%) and in Biotechnology 56% of the outlinks are domestic organizations, whereas this much lower in the related field of Genetics (28%).⁸

On the country level, the ‘preferential attachment’ of the departmental outlinks is much stronger than on the European level. This is true for all fields, and ranges from 45% in Literature Studies to 100% in Astrophysics and Information Science. In all fields, the most occurring links are generally domestic. And, the links indicate the relevant local audiences and local resources. In addition to the national research council (funding) and the academy of science (policy), local universities are well represented in all fields.

Differences in website characteristics

Since the target audiences and types of output vary across fields, we expect that the field differences in We operationalize this in terms of the number of webpages, outlinks, images, video-files, audio files, web-maps, applications (java, docs, pdf, etc.) and total number of objects (Table 5)

The average size of the sites show enormous differences (in the year of observation) ranging from 76 pages per site in the field of Psychology to 1665 pages per site in Computer Science. In some fields departmental websites are obviously a less important medium for communication of data and output (Genetics, Biotechnology, Psychology and Literature Studies). At the same time, Astrophysics and Computer Science research groups use their websites to a great extent, and Information Science and High Energy Physics in between⁹. Not surprisingly, the large websites have the most outlinks ($r = 0.84$), and the size of websites is also positively related to the level of codification.

⁸ Netherlands' figures. National orientation correlates with size of countries: organizations in large countries have more local linking opportunities.

⁹ The differences in distribution between these groups are also significant in the pairwise analyses.

Very pronounced differences between the fields are visible in the number of images and video files that the websites contain. The websites in the fields of Astrophysics and High Energy Physics contain a large amount of digital visualizations, which indeed are important in these fields (OECD 1998; Gooding 2002). It also shows that the web is used for data sharing. Computer Science websites also contain relatively large numbers of images and videos. Audio files were equally present in all fields.

Table 5. List of site characteristics in selected fields (averages per site by field).

Number of:		Web Pages	Out Links	Images	Gate ways	Applications	Audio Files	Video Files	Text Files	Web Maps
Fields	N									
Comp	1489	1666	639	876	51	254	10	3,68	41	51
Astro	41	1321	1073	1082	29	183	7,7	14	41	4,4
Bio	103	413	147	371	10	47	0,2	0,7	0,4	0,2
Gen	144	313	261	273	11	37	0,4	0,9	2	0,1
Hep	16	519	445	1857	22	168	1,5	0,7	4,7	0,1
Info	217	678	290	482	25	268	8,8	1,2	14	34
Lit	136	350	279	220	18	27	4,2	0,2	0,8	0
Psy	53	76	47	63	2	17	0,17	0	0,1	0
Anova (sign)	,000	,004	,000	,118	,000	,865	,000	,038	,989	

The number of applications (like *.doc and *.pdf files) and of text (*.txt) files indicates the digital content available on the sites. Two groups exist. Biotechnology, Genetics, Literature and Psychology have in average less than 47 application files on their websites, showing a low web use for information exchange. In contrast with this, the average site in Computer Science, Astrophysics, High Energy Physics and Information Science is much bigger, and contains more than 168 files.¹⁰ Also here, the size of the websites is decisive, also for the number of files and applications. This distribution is not unexpected. Only for Genetics is, as in this field large online databases play a crucial role – which was also visible in the analysis of the outlink pattern. Obviously, this function is not necessarily reflected in a rich content of websites. Finally, the size of the website is the main determining factor for the content, as the variation of *content per webpage* is relatively low between fields (Table 6)

Academic impact of the websites

In the previous sections we showed that the characteristics of the websites differ largely, as did the hyperlink patterns. In the analysis in this section we investigate the relationship between the academic impact of a departmental website and the characteristics of its website. The academic impact (or importance) of a website is measured here in terms in the number of inlinks it receives from *other departments in the field*. The question to be answered is whether this academic web impact is based on (correlates with) characteristics of the website. We only use the important characteristics, such as numbers of pages, outlinks, images and content (documents, databases, programs, and so on). The analysis is done per research field, and the results are shown in Table 7.

All correlations are positive, indicating that in general large websites with a lot of content (documents, databases, spreadsheets, etc) and outlinks are more popular and seem to have a larger academic impact than smaller sites with less outlinks and content. Inspecting the results in more detail, we do not find a systematic difference between the so-called mode-1 and mode-2 fields, nor between the fields with large, medium size and small websites. The only ‘systematic’ difference seems between the sciences and the social sciences and humanities: in the latter fields, the correlation between website characteristics and academic inlinks seems somewhat lower than in the sciences. Overall, the academic status of websites seems to be discipline specific – or even more department specific – and not so much related to mode-1 versus mode-2 fields.

¹⁰ Significant between the two groups - not between fields within the same group.

Table 6. List of average page characteristics in selected fields

Field	Out links	Images	Gate ways	Applications	Audio files	Video Files	Text Files	Web maps
<i>Comp</i>	0.38	0.53	0.03	0.15			0.02	0.03
<i>Astro</i>	0.81	0.82	0.02	0.14	0.01	0.01	0.03	
<i>Bio</i>	0.36	0.90	0.02	0.11				
<i>Gen</i>	0.83	0.87	0.04	0.12			0.01	
<i>Hep</i>	0.86	3.58	0.04	0.32			0.01	
<i>Info</i>	0.43	0.71	0.04	0.40	0.01		0.02	0.05
<i>Lit</i>	0.80	0.63	0.05	0.08	0.01			
<i>Psy</i>	0.62	0.83	0.03	0.22				
Coeff of variation	0.34	0.91	0.31	0.59			0.46	0.35
Empty cell: <0.01								

Table 7. Correlation between academic inlinks and various website characteristics.

	Comp	Astro	Bio	Gen	Hep	Info	Lit	Psy
<i>Size</i>	0.47	0.67	0.78	0.61	0.40	0.20	0.63	0.20
<i>Outlinks</i>	0.34	0.43	0.66	0.51	0.29	0.46	0.31	0.46
<i>Images</i>	0.47	0.45	0.72	0.43	0.28	0.25	0.59	0.25
<i>Content</i>	0.51	0.81	0.45	0.44	0.69	0.35	0.44	0.35

Italic: not significant; Size = nr of pages; Content = nr applications (text, data, programs)

Conclusion and discussion

The previous sections analyzed the web-spheres of eight scientific fields, specifically their 1) linked environments, 2) content, and 3) academic reputation. We now firstly summarize the most salient findings per research fields. Then we reflect on the differences and similarities, and answer questions whether the differences relate to differences between sciences and humanities, or between mode-1 and mode-2 research fields.

In *Astrophysics* ICTs play an important role, as found in the website characteristics: an exceptionally large number of video files, for example. The outlinks suggested a well-defined academic audience with a large set of shared outlinks, many to universities. Additionally, the high number of outlinks to governmental organizations indicate the role of government support.

High Energy Physics departments link almost all to CERN, and for the rest to other academic institutions. In term of content, the number of images on the websites is exceptionally large, but also for the rest, the sites have much content. In this discipline, the websites seem to be an important medium of communication content to a predominant academic audience.

Biotechnology, a clear example of mode-2 knowledge production, has a focus on applications, is subject to policy involvement and has a heterogeneity in producers and users of knowledge. The websites are small and have not much content, suggesting that the role of the web is small in this field. The (small number of) outlinks are local and have a strong commercial orientation. This latter orientation explains the low level of web use (Nentwich 2003).

In *Genetics*, websites are small, as is the number of outlinks. However, within the outlinks, those to international data-repositories have a prominent role, as expected. For the rest, outlinks seem domestically oriented, apart from links to publishers and scientific journals. This confirms that in Genetics researchers typically circulate information only within smaller groups and broader access

depends upon publication in journals (Kling and McKim 2000). Here the distinction between fields with a restricted flow of information (like Biotechnology and Genetics) and fields with an open flow of information (like Astrophysics and High Energy Physics) becomes relevant – and this relates of course to the economic potential of genetic and biotech data.

Computer Science websites show a relatively high number of shared outlinks ('codification') and contain a large number of files and outlinks. The number of applications (content) is among the highest of the fields studied here. Furthermore, the outlinks have a more commercial orientation than other fields, suggesting the relevance of non-academic audiences in the field of Computer Science.

Like in Computer Science, in *Information Science* the web plays an important role, suggesting that the field has an 'open information flow'. Sites have a bigger content (number of applications) than in any other field. The outlinks go to a variety of audiences (apart from other academic departments). A relatively large number of outlinks are directed to governmental organizations and companies, underlining a stronger application orientation than most other fields. Therefore, big websites are not necessarily full with academic output, and this explains the comparably low correlation of the number academic inlinks with website size and content.

The departmental websites in *Literature Studies* are generally very small, and contain small numbers of files. The few – and generally local oriented - outlinks indicate a mainly academic audience. In this example of a traditional mode-1 field, scholars have a strong tradition in book publishing, a factor that Nentwich (2003) identified as having a negative impact on the level of 'cyberness'. Our analyses confirms this: in the hyperlink environment we find relatively many book publishers.

Finally, *Psychology* represents a mode-1 field in the social sciences. Websites are very small and maintain a small number of outlinks, showing that the Web plays a minor role in the field (Barjak, 2004). Furthermore, there is little common orientation in the set of shared outlinks, and a big percentage of these outlinks were local.

The sample seems to group into two categories. In Astrophysics, High Energy Physics, Computer Science and Information Science the web is used intensively, the number of shared outlinks is relatively high, the outlinks show an international orientation, and the number of webpages, outlinks, and content on the websites are large. A difference is that in the two physics specialties data sharing is an important issue (NASA; CERN as the most linked organizations), whereas in the two other fields it is not. And in Computer Science and Information Science, the relation with the non-academic environment seems stronger.

On the other hand, in Biotechnology, Genetics, Literature Studies and Psychology, websites are in average small, have a modest content, hardly share outlinks which are more often local. In some of the latter fields this may indicate that the WWW is not very important yet, in others, such as Genetics, the size may be more a reflection of the restricted access to the data, and not that data are not shared – as they are through (NCBI).

In light of these results, we now turn to the three hypotheses formulated in the introduction about the relation between 'cyberscience' and changes in the knowledge production system:

- (1) mode-2 sciences make more extensive use of Internet applications than mode-1 sciences.
- (2) mode-2 sciences disseminate a greater variety of outputs through the web compared to mode-1 sciences.
- (3) mode-2 science address a greater variety of audiences through the web compared to mode-1 sciences.

Firstly, the size of the websites is obviously not related to the difference between mode-2 and mode-1. Secondly, the same holds for the content of the websites, in terms of applications, of images, video and audio, and in numbers of outlinks¹¹. In other words, *hypotheses 1 and 2 are not supported*. If there is a relationship, we find it more between open information fields (like physics, computer science and

¹¹ If we compare the *content per webpage* between the fields, differences disappear.

information science) and the fields with restricted information flows (like the life sciences). This relates more to the type of valorization of knowledge than to the question of whether application contexts play a role or not. The position of the social sciences and humanities in this context needs further exploration. Another finding is that the early adopters of ICTs have the bigger sites (physics, computer/information science). The question is whether this is an issue of being behind (social sciences, humanities) or of variation, of heterogeneous developments.

Thirdly, outlink patterns were rather different, in terms of the codification, the type of linked organizations, and in terms of the shares of international links. Codification differed, and was mainly related to the size of the websites, and not to the mode-2/mode-1 distinction. The linked environments differed between the disciplines, and could sometimes be related to specific mode-2 characteristics of the field – but certainly not always. For example, disciplines like computer science, biotechnology and information science have many commercial outlinks, as one would expect given the economic role of these fields, but why this also is the case for High Energy Physics is less clear. Astrophysics and Literature had significantly more academic outlinks. On the other hand, it was not clear why one would expect more governmental outlinks in fields like Astrophysics, Biotechnology, Information science or Literature. Also the size of the outlink environment, its diversity, and its (inter)national orientation does differ, but not related to ‘mode-2-ness’. In other words, outlink patterns were different between disciplines, but not systematically related to the mode-1 versus mode-2 distinction. Summarizing, also *hypothesis 3 is not supported*. Finally, in all fields we found that the size of sites (in terms of pages, content and outlinks) correlates relatively strong with the academic impact of the site, but also here the strength of the correlations did not systematically differ between mode-1 and mode-2 fields.

As a general conclusion, the web does play an important role in facilitating the mode-2 characteristics of knowledge production: in sharing data and information, in showing the network of the research organization, in supporting the interaction with non-academic partners, and in the dissemination of output. However, these characteristics of mode-2 can be observed in each of the fields to a different extent. The distinction between mode-1 and mode-2 sciences therefore seems less a dichotomy. Rather, it is better to speak of mode-1 aspects and mode-2 aspects of knowledge production, with each scientific field being characterized by a mix of both types of aspects. If such nuances are forgotten, terminologies quickly start to live a life on their own, and such lives tend to replicate extremely fast in academic and policy circles alike. May be we need more subtle differences, in more dimensions. For example, low levels of codification, and related, high shares of local outlinks, may reflect heterogeneity of research fields, reflecting uncertainty (Whitley 2000) and diverging search regimes (Bonacorssi 2005), or low levels of dependency between researchers in the field (Whitley, 2000). And, as already suggested, the use of the public web for sharing networks, knowledge, data and information may depend on the way the (economic) value of science is appropriated (Nelson 2004; Dasgupta & David 2004, David & Foray 2002). This of course, needs further exploration.

References

- Arroyo, N., V. M. Pareja, et al. (2003). *Description of web data*. Eicstes deliverable D3.2. Madrid, CINDOC.
- Barjak, F. (2004). *On the integration of the Internet into informal science communication*. Solothurn, University of Applied Sciences, Northwestern Switzerland.
- Bonaccorsi, Andrea (2005). *Better policies versus better institutions in European science*. Paper presented at the PRIME conference, Manchester, January 2005.
- Dasgupta, P. & P.A. David (1994). Toward a new economics of science. In: *Research Policy* 23, pp. 487-521.
- David, P.A. & D. Foray (2002). An introduction to economy of the knowledge society.' In: *International Social Science Journal* (March), pp. 9-23.
- Etzkowitz, H. & L.A. Leydesdorff (eds.) (1997). *Universities and the global knowledge economy. Science, technology and the international political economy*. London/New York, Continuum.
- De Jong, H. and A. Rip (1997). The computer revolution in science: steps towards the realization of computer-supported discovery environments. In: *Artificial Intelligence* 91, 225-256.
- ESFRI (2006), *First european roadmap for new, large-scale research infrastructures*. Luxembourg, EC.
- Gibbons, M. C., H. Limoges, et al. (1994). *The new production of knowledge*. London, Sage.

- Gooding, D. C. (2002). Narrowing the cognitive span: experimentation, visualization and digitalization. In: H. Radder, *Scientific experimentation and its philosophical significance*. Pittsburg, University of Pittsburgh Press.
- Harries, G., D. Wilkinson, L. Price, R. Fairclough, M. Thelwall (2004). Hyperlinks as data source for science mapping. In: *Journal of Information Science* 30, 436-447.
- Heimeriks, G., M. Hoörlesberger, P. Van den Besselaar (2003). Mapping communication and collaboration in heterogeneous research networks. In: *Scientometrics* 58, 391-413.
- Heimeriks, G. & P. Van den Besselaar (2006). Analyzing hyperlink networks: the meaning of hyperlink-based indicators of knowledge production. In: *Cybermetrics* 10. Accessible at: www.cindoc.csis.es/cybermetrics
- Kling, R. and G. McKim (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. In: *Journal of the American Society for Information Science* 51 (14): 1306-1320.
- Nelson R.R. (2004). The market economy, and the scientific commons. In: *Research Policy* 33, pp. 455-471.
- Nentwich, M. (2003). *Cyberscience, research in the age of the internet*, Austrian Academy of Sciences.
- OECD (1998). *The global research village: how information and communication technologies affect the science system*. Paris, OECD.
- Price, D., de Solla (1984). The science-technology relationship. In: *Research Policy* 13, pp. 3-20.
- Rip, A. (2002). Science for the 21st century. In: P. Tindemans, A. Verrijn-Stuart and R. Visser, *The future of the sciences and humanities*. Amsterdam, Amsterdam University Press.
- Stokes, D. (1997). *Pasteurs quadrant. Basic science and technological innovation*. Washington: Brookings Institution Press.
- Van den Besselaar, P. (2007). *Knowledge networks*. Inaugural lecture University of Amsterdam, December 8, 2005. Amsterdam, Vossius Pers.
- Van den Besselaar, P., and Heimeriks, G. (forthcoming). New media and communication networks in knowledge production: A case study.
- Vasileiadou & Van den Besselaar (2006). Linking shallow, linking deep; how scientific intermediaries use the web for their network of collaborators. In: *Cybermetrics* 10. Accessible at www.cindoc.csis.es/cybermetrics
- Whitley, R. (2000). *The intellectual and social organization of the sciences*. Oxford, Oxford University Press.
- Wilkinson, D., G. Harries, M. Thelwall, L. Price (2003). Motivations for academic website interlinking: evidence for the web as a novel source of information on informal scholarly communication. In: *Journal of Information Science* 29, 49-56.

Impact of Indian Patents: Assessment through Citation Analysis¹

Sujit Bhattacharya

Sujit_academic@yahoo.com

NISTADS (National Institute of Science, Technology and Development Studies), Pusa Gate K. S. Krishnan Marg, New Delhi-110012 (India)

Abstract

The present study examined the impact of patents granted to Indian organisations by the USPTO during the period 1990-2002. The impact analysis was based on the patents cited by other patents and journal articles. The citing and cited data was disaggregated at different levels to bring out the various characteristics. The impact of patents of 'prolific' patenting organisation, impact of patents in different sectors/sub-sectors, highly cited patents, scientific fields/sub-fields that were citing Indian patents, and organisations that were noticing these patents, etc were uncovered. The importance of this exercise for policy and strategic purpose are discussed.

Keywords

citation analysis; forward citation; impact assessment

Introduction

Empirical studies have pointed out the highly skewed distribution of patents' value. Along with Forward-Patent-Citation (*number of citations a patent receives from other patents*), different approaches have been undertaken to determine patent value (Chen and Hicks, 2004; Harhoff, Narin, Scherer & Vopel, 1999; Harhoff, Scherer & Vopel, 2003; Lanjouw & Schankerman, 1997; Rosenkopf & Nerkar, 2001; Lanjouw & Schankerman, 1999; Lerner, 1994; Gullec & Pottelsberghe, 2000; Sherry & Teece, 2004; Shane, 2001). Sapsalis and Pottelsberghe (2005) review on patent value have found that despite strong heterogeneity across studies, some similarities emerge. *The most important is probably the fact that the number of Forward-Patent-Citation (FPC) is closely associated with the value of a patent. All studies using FPC reach this conclusion (see for example Trajtenberg, 1990; Hall, Jaffe & Trajtenberg, 2000).* Drawing lessons from empirical studies, Forward-Citations were used in this study to determine the impact of patents granted to Indian organisations by the USPTO during the period 1990-2002. The study was not restricted to citations from patents only (FPC) but also included citations by journal articles (*thus we use the term Forward-Citations in this study*).

Methodology

Cited patents, for this study, comprised patents granted to resident Indian organisations (organisations with legal address in India) for the period 1990-2002 in the USPTO. Citing patents i.e. all patents that cited the above patents during this period was accessed from the online USPTO database. Further validation of this data was undertaken through INPADOC database. Only Front-Page citations (Examiner Citations) were examined. Citation by journal articles to the above patents was culled out from the Web-of-Science. Citation-impact-analysis was thus restricted to journals covered by this database.

Results

Overview of Indian Patenting Activity in the USPTO

669 patents were granted to Indian organisations during the period 1990-2002 by the USPTO (Bhattacharya, Garg, Sharma & Dutt, 2005). The maximum growth in patenting activity was observed in the later period (1999-2002). This period accounted for 74% of total patents (i.e. 492 patents). There were 93 organisations involved in patenting activity. However, patenting activity was highly skewed with 8 organisations (termed as 'prolific' organisations) accounting for approx. 78% (522 patents) of the overall patents. Pharmaceutical (284 patents) and Chemical (232 patents) were the major areas of

¹ This work was supported by the Office of the Principal Scientific Advisor to the Government of India.

patenting activity. The patents granted in other sectors were: Machinery (28 patents), Instruments (18 patents), Electronics (9 patent), Transport (6 patents), and Electrical Equipment (1 patent).

Indian Patents Cited by Other Patents

Overall Citation Pattern

Table 1 exhibits the cited profile of patents granted to Indian organisations.

Table 1. Cited details of patents granted to indian organisations

Year	Total Patents	Patents Cited	Citation in Different Time Periods			Times-Cited (1990-02)	Citation-Per-patent
			1990-94	1995-98	1999-2002		
1990- 94	50	36 (72%)	12	70	93	175	3.5
1995- 98	127	77 (61%)		24	293	317	2.5
1999- 02	492	149 (30%)			328	328	0.7
1990- 02	669	262 ($\approx 39\%$)				820	0.8

It can be observed that almost 40% of the patents were ‘noticed’. It corroborates the fact that it takes time for patents to attract citations, i.e. to get noticed (this is similar to the research paper, i.e. on an average there is some time gap before a research paper is noticed).

Figure 1 exhibits the citation distribution of the cited patents using a Lorentz curve distribution.

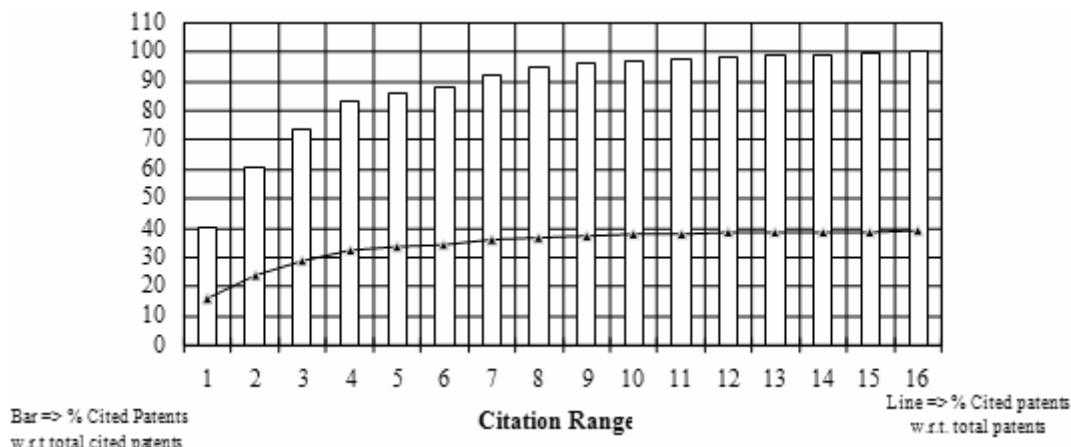


Figure 1. Distribution of Cited Patents

The cited distribution follows the established pattern of these types of distribution i.e. highly skewed with a long tail (Scale-free-network) which can be described by a power law (Chen & Hicks, 2004). There were 10 highly cited patents attracting a total of 172 citations. These patents can be termed as ‘Key’ patents’ or using the analogy of highly cited paper (Price, 1965) can be called as *technology fronts*.

Impact Assessment Organisation-wise

A significant correlation (0.9 significant at 0.01 levels) was observed between the intensity of patents being cited with the number of patents. Very high and significant correlation (0.8, significant at .01 levels) was observed between organizations own patents being cited and total citation it had received. Examiner’s citing the patents of a same organisation may indicate the technological continuity in the organisation. Current-Impact-Index (CII) (Narin, 1993) defined as the number of times an

organisation's most recent five years of patents are cited in the current year, relative to the entire patent database was calculated to uncover the current technological relevance of the patents to the prolific institutions. It is possible to calculate another index from CII, the Technology- Strength-Index (TS) (Narin, 1993) defined as:

$TS = \text{Number of patents} \times \text{Current-Impact-Index}$

In the present study, the impact of the five-year patents, (1998-2002) was seen in the year 2003. The number of patents in 2003 was taken for calculating TS. The CII and TS was calculated for all the eight prolific organisations, exhibited by Table 2.

Table 2. Current impact and technology strength for patents of prolific organisations

Organizations/ Industries	CII	Patents in 2003	Technology Index
CSIR	0.52	142	71
Ranbaxy Laboratories	0.07	8	1
Dr. Reddy's Research Foundation	0.43	8	3
Dabur Research Foundation	0.42	9	4
IOCL	0	5	0
NII	0	-	0
Panacea Biotech	0.17	1	0.2
Lupin Laboratories	0	-	0

CSIR had highest CII value of 0.52 among the prolific organisations. This indicates that 52% of its patents were expected to have impact in 2003. In- spite of the highest citation-per-patent of Panacea Biotech, its patents exhibited a lower impact in 2003, i.e., impact of 17 % only. IOCL, NII and Lupin have value of CII as zero as none of their patents have attracted any citation in 2003. CSIR had 142 patents in 2003. TS indicate that its patent portfolio strength is approximately 71 in 2003. TS imply that out of 142 patents of CSIR, 71 patents were expected to be noticed in 2003. Similarly, the other values can be interpreted from the above Table.

Impact Assessment: Sector/Sub-sector wise

The number of patents in a sector had direct bearing on the number of times they were cited. However, Electronics sector with 9 patents had higher impact then Instrument sector that had 18 patents. Further examination was done by normalising the data to minimise the size effect by taking ratio between total patents cited in a sector to total patents. It was observed that 'Transport', 'Electronics' and 'Instrumentation' sectors were attracting 66%, 55% and 50% citations respectively. On the other hand, considerably less number of patents 36% and 35% respectively were attracting citations in the sectors 'Pharmaceuticals', and 'Chemicals'.

Pattern of citation received

In some major MNCs granted patents, examiners have cited Indian patents. Examiners have cited CSIR, 41 times in patents granted to Eli Lilly, 24 times patents granted to GEC, and 10 times patents granted to Conoco Inc. This plausibly points out to some technology link between CSIR and the above MNCs. Similar technology link can be interpreted between Gramercy Jewellery (foreign firm) and Fine Jewellery; 3 M Innovative Prop Co and Carborundum, etc. However, this type of technology link is absent among cited patents of Indian firms.

Indian Patents Cited by Journals

Journal articles cited 95 patents out of total 669 patents granted to Indian organisations. Thus, approx. 14% of patents received citations from journal articles. These 95 patents received 167 citations from journal articles. *This is an encouraging result indicating that Indian patents were beginning to be noticed by journal articles and thus have scientific significance.* Similar to patent citation, citations by

journal articles were extremely skewed. 5 patents attracted 36 citations implying their significant scientific impact. Journals prominently citing Indian patents were Expert Opinion on Therapeutic Patents (7 citations), Journal of Micro encapsulation (6 citations) and Journal of Molecular Catalysis A- Chemical (6 citations). Chemistry (67 citations) followed by Clinical Medicine (14 citations), Biomedical Research (11 citations), and Engineering & Technology (10 citations) were the prominent scientific fields from where citations were received. Table 3 indicates the major subfields that contributed to the majority of the citations.

Table 3. Major sub-field of citing journals

Sub-field Name	Cited activity
<i>Physical Chemistry</i>	27
<i>Organic Chemistry</i>	12
<i>General Chemistry</i>	10
<i>Pharmacology</i>	8
<i>Materials Science</i>	4
<i>Analytical Chemistry</i>	3

Relationship of Forward Citations with Backward Citations

Quantitative relationship among patent cited w.r.t journal cited was done. A negative correlation was obtained implying patents that were highly cited by other patents were not cited by journal articles and *vice-versa*. *In other-words, technologically significant patents were not scientifically significant and vice-versa.*

Both patent cited and journal cited had negative correlation with backward citations. A very weak correlation (.01) was observed between patents cited by other patents (Forward-patent-citation) and citations to patents given by the examiners in those patents (backward patent citations). A weak negative correlation was observed between patents cited by journal articles and citations given by the examiners to journal articles in those patents. These results were not expected. It could only be concluded that technologically as well as scientifically significant patents require much less ‘prior art’ to satisfy the claims. However, further examination would be required to come to any conclusion.

Conclusions and Discussion

The present study by examining citations received by patents granted to Indian organisations by the USPTO was able to provide some assessment of their impact. This type of extensive analysis is important as it can play a key role in important policy and strategic level decisions at the country/organisation level. Some of the questions that can be answered through this analysis are: Who are noticing our patents? Which are the key patents? Is scientific research leading to invention?, etc. Further extension of this analysis can be undertaken by examining the full text of the citing patents to identify in what context the citations were made. On the specific sub-set of highly cited patents, multiple indicators such as family size, claims, and backward- citations can be applied. These inputs can help in future licensing of the said patent and deriving other commercial appropriation from it. The weak relationship between forward and backward citations requires more in depth examination to find out the causality behind these results. A more informed judgement of impact of India’s patent can be obtained by benchmarking it with some other countries. Future research would address this.

References

- Chen, C. & Hicks, D. (2004). Tracing knowledge diffusion. *Scientometrics*, 59(2), 199-211.
- Harhoff, D., Narin, F., Scherer, F. M., Vopel K. (1999). Citation frequency and the value of patented innovation. *Review of Economics and Statistics*, 81(3), 511-515.
- Harhoff, D., Scherer, F.M. & Vopel, K. (2003). Citations, Family size, Opposition and Value of patent rights. *Research Policy*, 32(8), 1343-1363.
- Lanjouw J. & Schankerman M. (1997). Stylised facts of patent litigation: Value, scope and ownership. *NBER Working Paper* (6297).

- Rosenkopf, L. & Nerkar, A. (2001). Beyond local search boundary- spinning, exploration, and impact in the optical disc industry. *Strategic Management Journal*, 22, 287-306.
- Lanjouw, J. & Schankerman M. (1999). The quality of ideas: measuring innovation with multiple indicators, *NBER Working Paper* (7345).
- Lerner, J. (1994). The importance of patent scope: an empirical analysis. *RAND Journal of Economics*. 25(2), 319-332.
- Gulledge, D. & van Pottelsberghe, B. (2000). Applications, grants and the value of patent. *Economic Letters*, 69(1), 109-114.
- Sherry, E.F. & Teece, D.J. (2004). Royalties, evolving patent rights, and the value of innovation. *Research Policy* 33, 179-191.
- Shane, S. (2001). Technological opportunities and new firm creation. *Management Science* 47(2), 205-220.
- Sapsalis, E. & van Pottelsberghe B. (2005). The institutional sources of knowledge and the value of Academic patents. *Proceedings of the 5th Triple Helix Conference on The Capitalisation of Knowledge: Cognitive, Economic, Social and Cultural Aspects*, 18-21 May, Turin (Italy).
- Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations, *The Rand Journal of Economics*, 21(1), 172-187.
- Hall, B., Jaffe, A. & Trajtenberg, M. (2000). Market value and patent citations: A first look: *NBER Working Paper* (7741).
- Lanjouw, J. & Schankerman M. (1999). The quality of ideas: measuring innovation with multiple indicators, *NBER Working Paper* (7345).
- Bhattacharya, B., Garg, K.C., Sharma, S.C., Dutt, B. (2005). Indian Patenting Activity in International and Domestic Patent System: Contemporary Scenario.
- Price, D.D. (1965). Network of scientific papers. *Science* 149, 510-515.
- Narin, F. (1993). Tech-Line Background Paper. Version 19. CHI Corporation. <http://www.chiresearch.com>

Is HIV/AIDS in Africa Distinct? What Can we Learn from an Analysis of the Literature?

Omwoyo Bosire Onyancha * and Dennis N. Ocholla **

*b_onyancha@yahoo.com

University of Eastern Africa, Baraton, P.O. Box 2500 Eldoret (Kenya)

**docholla@pan.uzulu.ac.za

University of Zululand, Department of Library and Information Science,
Private Bag x1001 KwaDlangezwa 3886 (Kenya)

Abstract

This paper investigates the notion held by several people that HIV/AIDS in Africa is unique by use of the published literature. Using co-word and factor analyses of MEDLINE-extracted HIV/AIDS records, this study used five lists of terms to investigate the relatedness of various factors and diseases to HIV/AIDS. The lists consisted of risk factors, sexually transmitted diseases, tropical diseases, opportunistic diseases, and pre-disposing factors. Data (i.e. words.txt – consisting of keywords/phrases describing the aforementioned factors and diseases; and text.txt – containing HIV/AIDS papers' titles) were analyzed using TI computer-aided application software, developed by Prof. Loet Leydesdorff. Results revealed that several factors and diseases that are pre-dominant in Sub-Saharan Africa exhibited strong and high pattern of co-occurrences with HIV/AIDS, implying close associatedness with the epidemic in the region. Further areas of research, whose results will be used to make conclusive observations and arguments concerning the uniqueness of HIV/AIDS in Sub-Saharan Africa, are recommended.

Keywords

AIDS; Africa

Introduction

In a letter written by Thabo Mbeki in 2000 to world leaders, the South African president, observed that “*it is obvious that whatever lessons we have to, and may draw from, the West about the grave issue of HIV/AIDS, a simple superimposition of Western experience on African reality would be absurd and illogical*” (as cited in Cohen, 2000: para 1). Not only do the manifestations of the AIDS disease in Africa differ from those in the West but, as Cohen (2000) observes, AIDS-related diseases, and possibly disease progression itself, differ in the continent (i.e from region to region) that is home to about 71% of the global population infected with HIV. In turn, this difference is said to be clinical. Cohen reports that while tuberculosis amongst AIDS patients is rare in the west – especially, the USA and Europe – it is the most common disease afflicting HIV-positive people in Africa. He further notes that Kaposi’s Sarcoma, a cancer that causes purple skin blotching, commonly afflicts both HIV uninfected and infected persons in Africa, while in industrialized nations, the disease is largely restricted to HIV-infected, gay men. The same applies to *pneumocystis carinii*, a strain of pneumonia predominant in HIV-infected persons in developed countries. These arguments are based on clinical diagnoses of various diseases in HIV infected persons. Further observations point to how various factors aggravate the spread of HIV/AIDS in developing countries, hence the argument that the impact of HIV/AIDS in these countries is different from that felt in developed countries.

But what can we learn from an analysis of published literature on HIV/AIDS? Which of these known diseases/infections and factors are most commonly associated with HIV/AIDS in Africa? Given that scientific research is often mirrored in published literature, is the above description of the uniqueness of HIV/AIDS in Africa reflected in published literature? These questions reflect the aim of this study, namely, to find out whether or not HIV/AIDS in Africa is a distinct disease by identifying the opportunistic infections, pre-disposing factors, risk factors, sexually transmitted diseases, and other tropical diseases most commonly associated with HIV/AIDS in Africa as a whole, and Eastern and Southern Africa in particular. At this stage, the study provides the preliminary findings of a broader content analysis study of HIV/AIDS literature as produced in or about Africa.

Methods and Procedures

The Method

Co-word analysis was employed to examine the relatedness of HIV/AIDS-specific terms to five groups of terms (i.e. opportunistic infections, risk factors, pre-disposing factors, sexually transmitted diseases and other tropical diseases). The intention was to reveal whether or not HIV/AIDS in Africa is distinct. Co-Word analysis is a content analysis technique that “reveals patterns and trends in technical discourse by measuring the association strengths of terms representative of relevant publications or other texts produced in a technical field” (Coulter, Monarch & Konda, 1998:1206). The method is meant to identify associations between publication descriptors in order to determine themes and trends in a discipline (Kostoff, 2001). Co-word analysis provides a set of terms or descriptors that not only regularly occur together in a text or record, but also [may be used to] measure the regularity with which events occur (Jacobs, 2002). Thus, the process “measures the strength of association between two or more documents by the co-occurrence of the same ‘words’ (phrases, descriptors, classification codes, etc) in a chosen field”. Contextually, the term ‘documents’ refers to the title, abstract, and/or descriptor fields (Callon et al in Schneider & Borlund, 2004:537).

This method has been extensively used, as illustrated and exemplified in its published literature (Callon, Law & Rip, 1986; Leydesdorff, 1988; Turner, Chartron, Laville, & Michelet, 1988; Courtial & Law, 1989; Whittaker, 1989; Callon, Courtial & Laville, 1991; Law & Whittaker, 1992; Courtial, 1994; Coulter, Monarch & Konda, 1998; Kopesa & Schiebel, 1998; Bookstein & Raita, 2001; Ding, Chowdhury & Foo, 2001; Jacobs, 2002; Krsul, 2002; Aizawa & Kageura, 2003; Baldwin, Hughes, Hope, Jacoby & Ziebland, 2003; Bookstein, Kulyukin, Raita, Nicholson, 2003; Schneider & Borlund, 2003; Hui & Fong, 2004; and Onyancha & Ocholla, 2005). The different approaches and ways that co-word analysis has been applied in a variety of studies confirms Leysdedorff's (1988:209) observation that “since most science studies and nearly all science policy studies use institutionally defined sets of documents, this instrument [co-word analysis] could have a wide range of applications”.

The Co-word analysis technique has been most commonly utilized in mapping, or tracing patterns and trends in term associated-ness. Most of the aforementioned studies fall in this category. We briefly provide a glimpse of the applicability of co-word analysis by reviewing a few of the studies that have used the method, beginning with Aizawa & Kageura (2003), who used the technique to calculate the association between technical terms based on co-occurrences in keyword lists of academic papers. The technique was also employed by Baldwin, Hughes, Hope, Jacoby & Ziebland (2003), who mapped ethics and dementia literature in order to identify dominating ethical issues, new and emerging areas of interest and those areas triggered by external events such as legal cases. Onyancha & Ocholla (2005) used co-word analysis to measure the relatedness of opportunistic infections to HIV/AIDS. Further examples of applications include: Kostoff (2001), who used the method to identify research themes in software engineering that (1) remained constant (2) matured and diminished as major research topics and (3) emerged as predominant research topics throughout the period of study; Jacobs (2002), who employed co-word analysis to study the use of particular words to describe respondents' job functions and the citation of information sources; and Schneider & Borlund (2004), who considered the applicability of co-word analysis in the construction and maintenance of thesauri. Citing several authors, Schneider & Borlund (2004:537) noted that the “units of analysis connected to co-word analysis (i.e. words, phrases, and descriptors) may illustrate cognitive structures of a field when displayed in so-called ‘semantic maps’”.

Table 1. List of countries and regions used in downloading papers from MEDLINE, SCI and SSCI

Angola	Botswana	Djibouti	Eritrea	Ethiopia
Kenya	Lesotho	Malawi	Mozambique	Namibia
Somalia	South Africa	Sudan	Swaziland	Tanzania
Uganda	Zambia	Zimbabwe	Eastern Africa	Africa, East*
Southern Africa	Africa, South			

Data Analysis and Presentation Procedures

The MEDLINE database was used to extract relevant data on HIV/AIDS research in Eastern and Southern Africa. Two sets of terms (i.e. HIV/AIDS-specific terms and country terms) were generated using several published sources. Tables 1 and 2 provide the terms used to download HIV/AIDS papers from the MEDLINE database. An advanced search mode was used to search and download data using the Title (TI), Author's Address (AF), Subject (SU) and Abstract (AB) fields. A total of 6176 records were downloaded, and upon screening (removal of irrelevant and duplicate records), 6178 records were obtained and analyzed. In order to find out the uniqueness of HIV/AIDS in Africa, five aspects were considered, i.e. HIV/AIDS' relatedness with opportunistic infections; pre-disposing factors; risk factors; sexually transmitted diseases; and other diseases (especially tropical diseases)

Table 2. List of terms used to identify HIV/AIDS papers from MEDLINE, SCI and SSCI

Acquired Immunodeficiency Syndrome	Immunodeficiency syndrome, Acquired	Immunologic Deficiency Syndrome, Acquired	Acquired Immune Deficiency Syndrome	Pneumonia, Pneumocystis Carinii
AIDS Arteritis, Central Nervous System	AIDS Dementia Complex	AIDS Seropositivity	HIV Seroprevalence	Immunologic Deficiency Syndromes
HIV*	HTLV-III	LAV-HTLV-III	Receptors, HIV	mmunoblastic Lymphadenopathy
Human T-Cell Lymphotropic Virus Type III	Sarcoma, Kaposi's	Human Immunodeficiency Virus	AIDS related complex	Human T Lymphotropic Virus Type III
Cytomegalic Inclusion Disease	Immunodeficiency Virus, Human	Virus, Human Immunodeficiency	Viruses, Human Immunodeficiency	Reverse Transcriptase Inhibitors
Human T-Cell Leukemia Virus				

Five lists of these diseases and factors were initially drawn from our personal experience with their usage in literature. Several sources (e.g. Nordberg, 2001; Conlon & Snydman, 2004) were thereafter consulted to refine the lists. Finally, expert advice was sought from a resident medical doctor and lecturers in the Departments of Nursing (University of Eastern Africa, Baraton and University of Zululand, respectively) who advised us on the terms that needed to be dropped from, or added to the lists. Extreme caution was taken to ensure that the lists were as exhaustive as possible. Two computer files were prepared, namely, words.txt (containing the words/names in Appendix A) and text.txt (containing titles of HIV/AIDS records) for analysis. Various authors (e.g. Luhn, Feinberg, Buxton, Manten, and Tocatlian, all as cited by Yitzhaki, 2001:759) have noted that titles are very important components of any scientific or scholarly article as they form part of the access points in search and retrieval processes. According to Yitzhaki (2001:759), many information retrieval systems "*depend heavily on indexing by automated, computerized selection of words from article titles*". Perhaps this is why great importance is placed on highly informative titles and it was on this basis that we considered the title words for a co-word analysis.

Data (i.e. words.txt and text.txt) were analyzed using TI.exe computer application software, developed by Prof. Leydesdorff, University of Sweden. The co-occurrence files thus generated (i.e. COOCC.DBF and COSINE.DBF) were exported to UCINET for the preparation of computer files that could be used by Pajek Software to construct social networks of the associated-ness of HIV/AIDS with each of the variables (i.e. words/names). Leydesdorff (2004) explains that *coocc.dbf* contains a co-occurrence matrix of the words found in the texts. In turn, this matrix is symmetrical and contains the words both as variables and as labels in the first field. The main diagonal is set to zero. The number of co-occurrences is equal to the multiplication of occurrences in each of the texts. *Cosine.dbf*

contains a normalized co-occurrence matrix of the words from the same data. Normalization is based on the cosine between the variables conceptualized as vectors (Salton & McGill as cited by Leysderdorff, 2004). Both files (*coocc.dbf* and *cosine.dat*) contain the information in DL-format. Whereas the file *coocc.dbf* consists of co-occurrence frequencies, *cosine.dat* contains the strengths of ties between two or more words in the text, in which case the value ranges between zero and one whereby the higher the value, the stronger the relationship between the words.

Finally, the findings from the above three analyses were compared with results from previously conducted international and foreign studies in order to determine whether or not there were differences that would warrant a generalized conclusion illustrating that HIV/AIDS in Africa is a distinct disease.

Results and discussions

This section provides an analysis of the co-occurrence of HIV/AIDS' most used acronyms (AIDS, HIV and HTLV) with selected terms such as opportunistic infections, pre-disposing factors, risk factors, sexually transmitted diseases, and other diseases in an attempt to find out the relatedness of these factors and diseases to HIV/AIDS in Africa at large, and E&S in particular. It also provides the normalized co-occurrence of words as a measure of the strength of the network (link) ties (whereby the strength S ranges between 0 and 1). The vertices in the normalized co-occurrence networks are not labeled since by so doing, the strengths of word ties could have been obscured. The labels, however, are provided in the preceding figures.

Co-occurrence of HIV/AIDS with Opportunistic Diseases

Figures 1 and 2 present the co-occurrence and relatedness of opportunistic infections and/to HIV/AIDS. The networks represent a large network that consists of AIDS, HIV and HTLV and their inter-linkages with other terms. Outside the networks are terms that were not associated with any of the terms in the network. These include Toxoplasmosis, Isosporiasis, Encephalopathy, Immunoblastic Lymphoma, and Coccidiomycosis, and although all the terms in the network seem to be associated with each other, some are not directly linked to AIDS, HIV or HTLV.

It was observed that AIDS co-occurred with 16 OIs as follows: Kaposi's sarcoma (16, $S=0.04$), Tuberculosis (16, $S=0.02$), Cancer (7, $S=0.03$), Mycobacterium Avium Complex (3, $S=0.01$), Pneumocystis Carinii (2, $S=0.01$), Pneumonia (2, $S=0.01$), Salmonella (2, $S=0.02$), Cryptococcosis (1, $S=0.02$), Cytomegalovirus (1, $S=0.01$), Leukoencephalopathy (1, $S=0.02$), Lymphoma (1, $S=0.01$), PML (1, $S=0.01$), Streptococcus pneumoniae (1, $S=0.01$), and Varicella Zoster (1, $S=0.02$). HIV co-occurred with 19 terms, with the highest co-occurrences originating from Tuberculosis (198, $S=0.17$), Pneumonia (23, $S=0.06$), Mycobacterium Avium Complex (18, $S=0.05$), Candidiasis (17, $S=0.03$), Kaposi's sarcoma (16, $S=0.03$), and Herpes Simplex (10, $S=0.03$). Others were Pneumocystis carinii (9, $S=0.03$), Carcinoma (4, $S=0.02$), Lymphoma (3, $S=0.01$), Salmonella (3, $S=0.02$), Streptococcus pneumoniae (3, $S=0.01$), Kansasii (2, $S=0.02$), Cryptosporidiosis (1, $S=0.01$), Cytomegalovirus (1, $S=0.01$), Histoplasmosis (1, $S=0.01$), Staphylococcus pneumoniae (1, $S=0.02$), and Varicella Zoster (1, $S=0.01$).

Co-occurrence with Pre-Disposing Factors

Figs 3 and 4 reveal that there are several inter-linkages between AIDS, HIV and HTLV and most of the pre-disposing factors, implying that some of these factors may be playing a role in the spread of HIV/AIDS. These include Drug Abuse, which co-occurred with AIDS in 51 titles and produced a normalized co-occurrence of $S=0.10$, followed by Rural-related factors (41, $S=0.06$), Orphans (27, $S=0.10$), Gender (9, $S=0.04$), Poverty (8, $S=0.05$), and War (8, $S=0.06$). Other terms that co-occurred with AIDS are Culture (3, $S=0.02$), Refugees (3, $S=0.02$), Violence (3, $S=0.02$), Discrimination (2, $S=0.02$), Labor Migration (2, $S=0.03$), and Rape (2, $S=0.01$). The rest produced one co-occurrence each.

HIV co-occurred with rural-related issues 213 ($S=0.21$) times, followed by drug abuse (51, $S=0.07$), gender (20, $S=0.07$), violence (11, $S=0.05$), and socioeconomic factors (7, $S=0.05$), while orphans, poverty, rape and refugees produced 6 co-occurrences each. Of the 32 pre-disposing factors, 11 terms

did not have any links with any other term. These were: Primitivity, Illiteracy, Unemployment, Sanitation, Handicapped, Uneducated, Disability, Urbanization, Conflict, Under-development (or underdeveloped), and Marginalization.

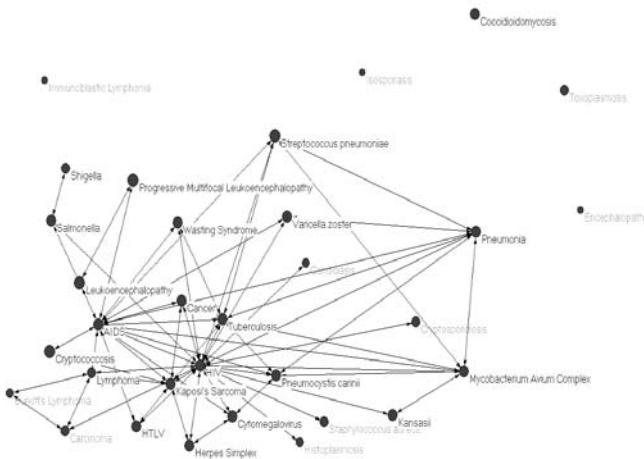


Figure 1. Co-occurrence of HIV/AIDS and Opportunistic Infections

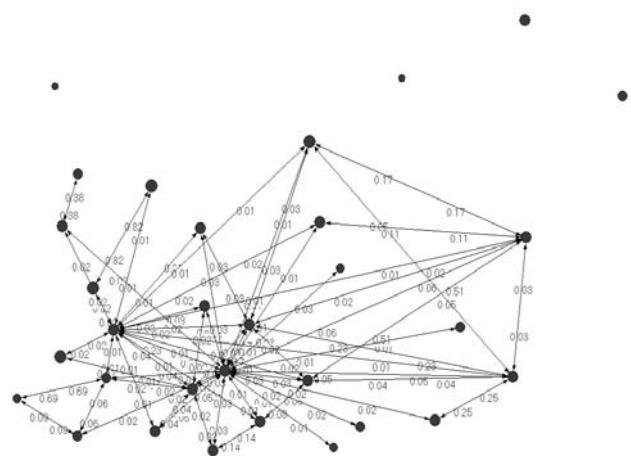


Figure 2. Normalized Co-occurrence of HIV/AIDS and OIS

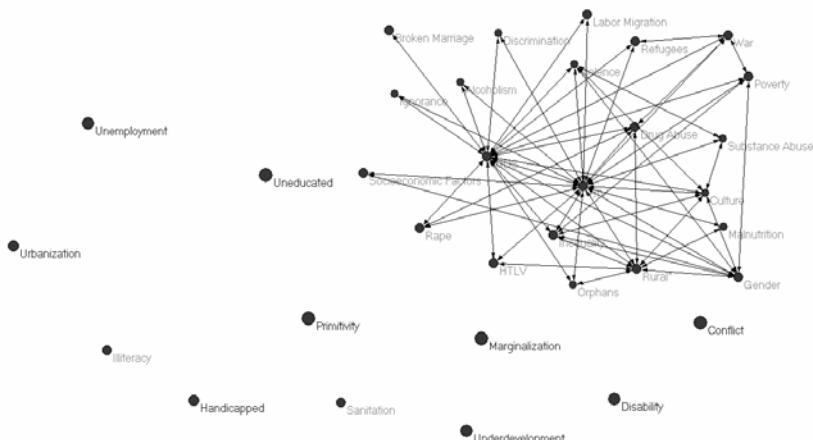


Figure 3. Co-occurrence of HIV/AIDS and Pre-Disposing factors

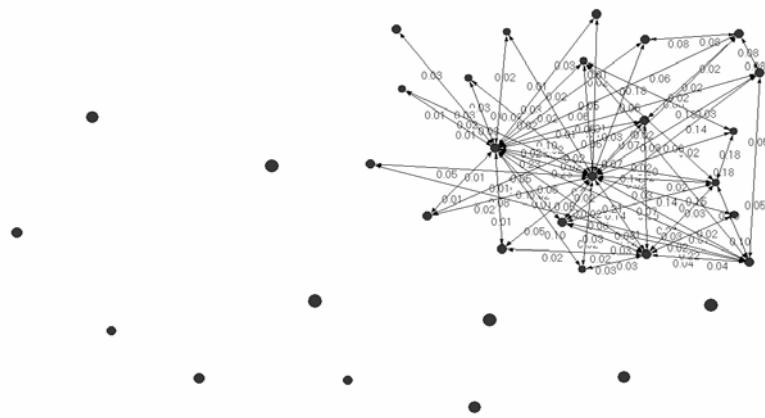


Figure 4. Normalized Co-occurrence of HIV/AIDS and Pre-Disposing factors

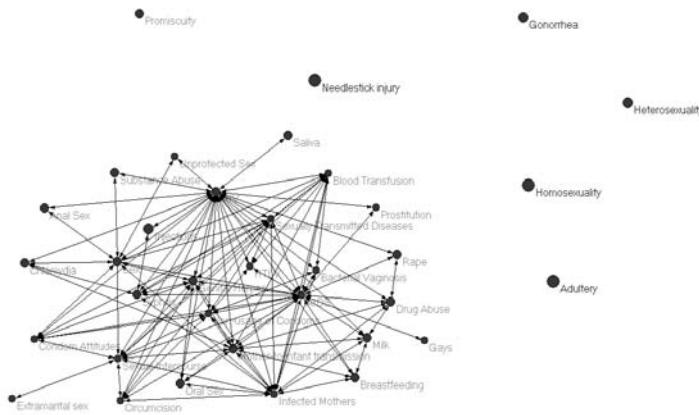


Figure 5. Co-occurrence of HIV/AIDS and Risk Factors

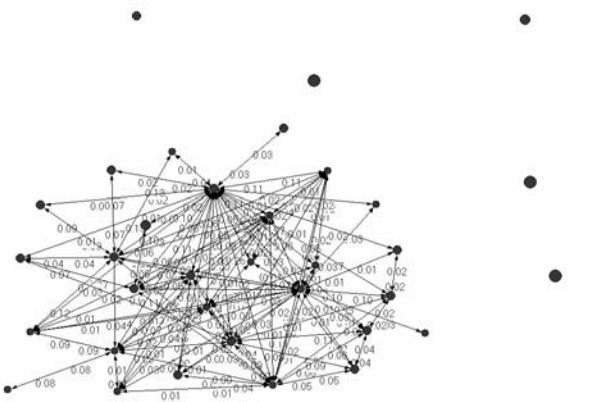


Figure 6. Normalized co-occurrence of HIV/AIDS and Risk Factors

Co-occurrence with Risk Factors

Five descriptors did not have any inter-linkages with HIV/AIDS terms. These consisted of Adultery, Heterosexuality, Gonorrhea, Needlestick injury, and Promiscuity. Figures 5 and 6 show that AIDS had links and co-occurred with 20 terms, in the descending order (measured by co-occurrence frequencies): Sexual Intercourse (40, $S=0.08$), Drug Abuse (51, $S=0.10$), Condom Attitudes (19, $S=0.04$), Infected

Mothers (10, $S=0.01$), Sexually Transmitted Diseases (9, $S=0.02$), and Non-usage of Condoms (8, $S=0.03$), to mention a few. HIV had its highest frequency of co-occurrence with Infected Mothers (303, $S=0.26$), followed by Mother-to-Infant Transmission (128, $S=0.17$), Sexual Intercourse (80, $S=0.11$), Sexually Transmitted Diseases (78, $S=0.11$), Blood Transfusion (75, $S=0.11$), and Drug Abuse (51, $S=0.07$). Others that recorded high frequencies of co-occurrence with HIV were Oral Sex (34, $S=0.08$), Breastfeeding (28, $S=0.08$), Genital Herpes (26, $S=0.06$), Circumcision (23, $S=0.07$), Non-Usage of Condoms (22, $S=0.06$), Condom Attitudes (21, $S=0.03$), Syphilis (21, $S=0.06$), Bacterial Vaginosis (15, $S=0.06$) and Milk (11, $S=0.04$). HTLV co-occurred once [each] with Breastfeeding, Homosexuality, Non-Usage of Condoms, and Sexually Transmitted Diseases.

Co-occurrence with Other Sexually Transmitted Diseases (STDs)

Figures 7 and 8 provide visual networks of the STDs and their inter-relationships with HIV/AIDS. Stand-alone terms (i.e. terms that are not linked to any other term(s)) include Condylomata Acuminata, Gonorrhea, Lymphogranuloma Venereum, Molluscum Contagiosum, Pediculosis Pubis, Pubic Lice, Scabies and Trichomonial Vaginalis. Results revealed that the term “AIDS” co-occurred with Human Papillomavirus Infection in 13 ($S=0.03$) titles, while it co-appeared with the descriptor “Sexually Transmitted Diseases” 9 ($S=0.02$) times. Other co-occurrences involved Hepatitis B (7, $S=0.02$), Syphilis (3, $S=0.01$), Bacterial Vaginosis (1, $S=0.01$) and Genital Warts (1, $S=0.003$). HIV had more co-occurrences than AIDS, results that coincide with all the other analyses. It recorded the highest frequency with Human Papillomavirus Infection (144, $S=0.09$) followed by Sexually Transmitted Diseases (78, $S=0.11$), Genital Warts (26, $S=0.06$), Hepatitis B (21, $S=0.04$), Syphilis (21, $S=0.06$), Bacterial Vaginosis (15, $S=0.06$), Herpes Zoster (10, $S=0.04$), Candidiasis (4, $S=0.03$), Granuloma Inguinale (3, $S=0.03$), Chlamydia (2, $S=0.01$), Pelvic Inflammatory Diseases (2, $S=0.01$), and Trichomoniasis (1, $S=0.01$). There were two terms that co-occurred with HTLV, notably, Human Papillomavirus Infection (3, $S=0.02$) and Sexually Transmitted Diseases (1, $S=0.02$).

Co-occurrence of HIV/AIDS with Other Tropical Diseases

An analysis of the relationship between HIV/AIDS and other diseases (particularly, tropical diseases) showed that a total of 16 titles (or records) contained the words AIDS and Tuberculosis, a relationship that produced a normalized co-occurrence of $S=0.02$, while Hepatitis co-occurred with AIDS in 7 ($S=0.02$) titles. Other terms that co-occurred with AIDS were Malaria (6, 0.02), Meningitis (3, $S=0.02$), Syphilis (3, $S=0.01$), Leishmaniasis (2, $S=0.02$), Sickle Cell (2, $S=0.05$), Cholera (1, $S=0.01$), and Hypertension (1, $S=0.02$). HIV had its co-occurrences with 11 terms which comprised Tuberculosis (198, $S=0.17$), Malaria (39, $S=0.08$), Hepatitis (21, $S=0.04$), Syphilis (21, $S=0.06$), Meningitis (15, $S=0.05$), Malnutrition (5, $S=0.03$), Leshmaniasis (4, $S=0.02$), Schistosomiasis (2, $S=0.01$), Cholera (1, $S=0.01$), Hypertension (1, $S=0.01$), and Polio (1, $S=0.02$). No term was found associated with HTLV. Notably, 12 out of 27 terms did not have any linkages, i.e. Amebiasis, Dengue, Ebola, Giardiasis, Hookworm, Jaundice, Lymphatic Filariasis, Oncocerciasis, Trypanosomiasis, Typhoid and Yellow Fever.

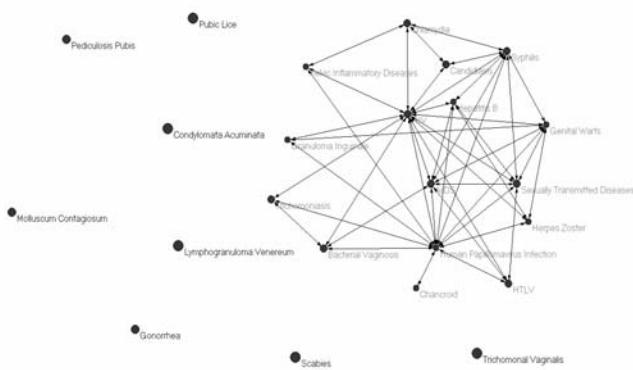


Figure 7. Co-occurrence of HIV/AIDS and Other STDs

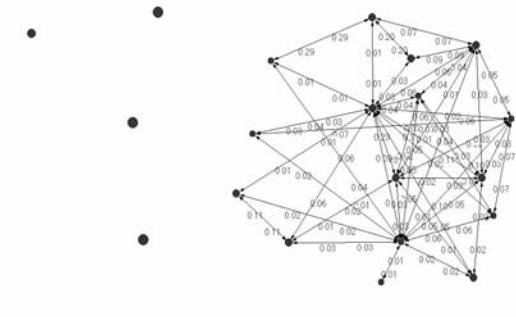


Figure 8. Normalized co-occurrence of HIV/AIDS and other STDs

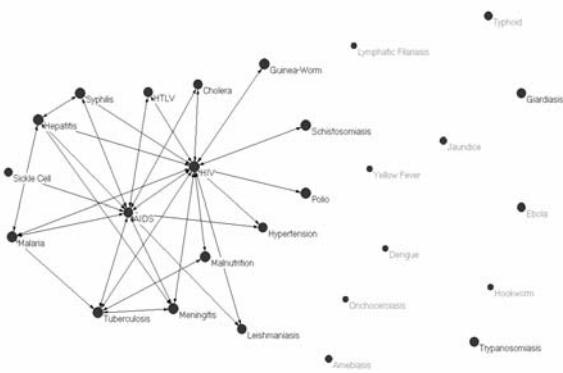


Figure 9. Co-occurrence of HIV/AIDS and other diseases

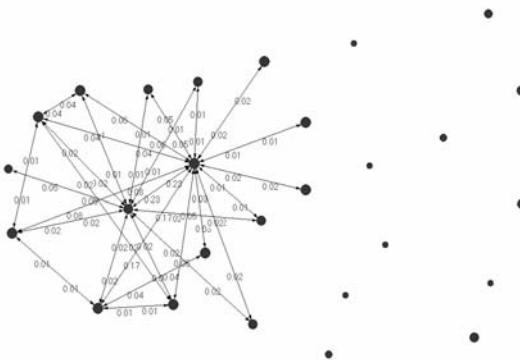


Figure 10. Normalized co-occurrence of HIV/AIDS and other Diseases

Conclusions and recommendations

An analysis of HIV/AIDS and the opportunistic diseases (see Fig 1 and 2) produced patterns that could be said to support arguments that some of the opportunistic infections' associations with HIV/AIDS in Africa is stronger than they could be in industrialized nations or any other geographic region (especially when compared to findings in previously conducted studies (e.g. Cohen, 2000)). Results revealed that HIV/AIDS was associated with 21 opportunistic infections, led by Tuberculosis, followed by Pneumonia, Mycobacterium Avium Complex, Cancer and Kaposi's Sarcoma. This revelation supports documented findings that claim Tuberculosis as the most common ailment in HIV-infected persons in Africa (e.g. Cohen, 2000). Cohen states that Tuberculosis kills more HIV-infected persons in Africa than any other AIDS-related disease. He further notes that the disease is rare in

AIDS patients in the United States and Europe, reporting that one neurologist and pathologist found no TB in all 390 autopsies that were performed on people who had died from AIDS. Other opportunistic infections such as Pneumocystis Carinii Pneumonia (PCP) are more common in HIV-infected persons in developed countries. Cohen (2000b) claims that PCP infected more than 80% of the AIDS patients in developed countries in the 1980s, while only 8% of the HIV-infected people autopsied in Africa were found to have had PCP. A few diseases did not have any connection with HIV/AIDS in Africa, as illustrated in Figures 1 and 2. These were Toxoplasmosis, Isosporiasis, Encephalopathy, Immunoblastic Lymphoma, and Coccidiomycosis. Some of these opportunistic infections (OIs) are missing from the list of the most commonly associated OIs with HIV/AIDS in Onyancha & Ocholla's (2005) study, a pattern that perhaps can be attributable to the international nature of that study. This may probably also support the view that HIV/AIDS pattern differs from one geographic region to another.

Concerning predisposing factors, the findings illustrated some association between several factors and HIV/AIDS in E&S Africa. Factors that could be influencing the spread of HIV/AIDS in the region include culture, substance or drug abuse, malnutrition, rural-related factors and activities, violence, rape, labor migration, ignorance, broken marriages, poverty, inequality, socioeconomic factors, refugees and war. Of these, the most influencing factors are rural and drug or substance abuse related, as illustrated by their high frequency and strength of co-occurrence and association with HIV/AIDS. Most of these factors should be subjects of concern in the intervention programs.

Another factor that this study considered in investigating the uniqueness of HIV/AIDS in Africa is the co-occurrence of AIDS-related risk factors with HIV/AIDS descriptors within the titles of HIV/AIDS papers. Terms that did not have any linkages with HIV/AIDS were adultery, gonorrhea, heterosexuality, promiscuity, and needlestick injuries. Their non-co-occurrence with HIV/AIDS terms should not be misconstrued, however, to mean that the risk factors are not related to HIV/AIDS at all and in this region. Most likely, the authors used related terms or their variants. Notably, most of the risk factors are sex-related. Perhaps this is attributable to the fact that HIV/AIDS is mainly contracted through sexual intercourse, especially between different sexes (i.e. heterosexually) in the case of Africa, as observed by Cohen (2000). Overall, the most common HIV/AIDS-associated risk factors constitute sexual intercourse, vertical transmission (mother to child during birth), blood transfusions and contaminated needles (intravenous drug use, needle stick injuries). According to the findings in Figs. 5 and 6, several AIDS-related risk factors, including the above, were associated with HIV/AIDS in E&S Africa. The highest co-occurrence between HIV/AIDS and the risk factors was recorded by "infected mothers", followed closely by a related descriptor, "mother-to-infant transmission". Sexual intercourse and sexually transmitted diseases also ranked highly. The descriptor "Contaminated needles" was less common.

One of the risk factors (and sometimes a pre-disposing factor) is the sexually transmitted diseases. Amuyunzu-Nyamongo (2001) argues that individuals with ulcerative STIs have an increased risk of transfer of HIV infection by factors of two to four. Of all the sexually transmitted diseases, Papillomavirus Infection was the most common in HIV/AIDS titles. It recorded a co-occurrence and strength of association frequency of 144 and $S=0.09$ with HIV, and 13 and $S=0.03$ with AIDS, respectively. There were other high co-occurrence frequencies from genital warts, hepatitis B, syphilis, bacterial vaginosis, and herpes zoster. Seemingly, HIV/AIDS is mainly linked to un-curable STDs. For instance, the human Papilloma virus is thought to be one of the main causes of cervical cancer and has been linked to other types of cancers of the female reproductive system. While this virus can be treated to reduce signs and symptoms, it does not yet have a cure. Both Herpes and Hepatitis B are other examples of STDs that do not yet have cures. Diseases or viruses that have cures co-occurred less frequently with HIV/AIDS, implying that they are rarely associated with the epidemic in Eastern and Southern Africa.

The effect of other diseases on HIV-infected persons was also considered by analyzing the relationship between HIV/AIDS and the selected diseases through term-co-occurrence analysis. It has long been observed that HIV/AIDS does not actually kill; rather it is the opportunistic infections/diseases (or

other diseases) that kill AIDS patients (Me'decins Sans Frontières, 2003). This study sought to identify the most common HIV/AIDS-associated diseases, especially tropical diseases. Out of the total 24 diseases, slightly over one-half ($\frac{1}{2}$) co-occurred with HIV/AIDS as shown in Figs 9 and 10. The highest frequency of co-occurrence was recorded by tuberculosis, which is said to be killing more HIV-infected persons in Africa than any other disease (Cohen 2000). Other terms that were linked to HIV/AIDS descriptors include cholera, guinea-worm, hepatitis, hypertension, leishmaniasis, malaria, malnutrition, meningitis, polio, schistomiasis, sickle cell, and syphilis. Although most of these diseases have no direct link with HIV/AIDS, it is common knowledge that most have an equally (if not greater) negative impact on the economies of E&S Africa and its peoples. For instance, Malaria is said to be killing millions of people in the region. The World Health Organization (2004) estimates that Malaria accounts for more than a million deaths per year, of which about 90% occur in tropical Africa. Again, it has been observed that HIV infection increases the incidence and severity of clinical Malaria and although the effect of Malaria on HIV is not well documented, UNICEF (2003) states that acute Malaria infection increases viral load. The relatedness of other diseases such as cholera and polio to HIV/AIDS may be attributed to the fact that all are diseases of poverty, which is a common factor in Sub-Saharan Africa. The reasons for the co-occurrence of HIV/AIDS and some of the diseases were, however, not very clear. Perhaps researchers were curious to discover the relationships between these diseases, or simply wanted to find out the impact the diseases have in E&S Africa.

In conclusion, the following diseases and factors produced high/strong co-occurrence patterns with HIV/AIDS:

- Opportunistic infections: *Tuberculosis, Pneumonia, Kaposi's sarcoma, Herpes Simplex, Candidiasis, and Mycobacterium Avium Complex.*
- Pre-disposing factors: *Rural-related issues, Drug abuse, Orphans, Gender, and Violence.*
- Risk factors: *Infected Mothers, Mother-to-infant transmission, Sexual intercourse, Drug abuse, Oral sex, and Breastfeeding*
- Sexually transmitted diseases (infections): *Human Papillomavirus Infection, Sexually Transmitted Diseases, Genital Warts, Hepatitis B, Syphilis, and Bacterial vaginosis.*
- Other diseases: *Tuberculosis, Malaria, Hepatitis, Syphilis, and Meningitis.*

The choice of terms used to conduct the co-word analysis largely influenced the patterns of co-occurrence, shown in section 3 above. It is possible that some terms (e.g. synonyms, related terms, etc) which were left out may have been used by authors in titling their papers. It is also true that authors' choice of terms when formulating article titles (i.e. research topics) differ from author to author. This analysis was also limited to only HIV/AIDS articles written by and/or about E&S Africa. Nevertheless, the data analysis provides partial insight into the uniqueness of HIV/AIDS in Africa. It is worth noting that this observation is not conclusive, as it requires a study into the relationship between these terms in other countries (outside Africa) before deductions can be made as to whether or not AIDS in Africa is a distinct disease. It is also recommended that a co-word analysis be conducted to check for strengths of association between descriptors of opportunistic diseases, pre-disposing factors, risk factors, sexually transmitted diseases and other diseases, as subject headings, and HIV/AIDS. The findings can then be compared to the results of this study so as to draw correct conclusions on the uniqueness of HIV/AIDS in Africa. In this way, intervention programs can be drawn to address the most common infections and/or factors that aggravate the AIDS situation in Africa in general, and Eastern and Southern Africa, in particular.

References

- Aizawa, A., & Kageura, K. (2003). Calculating association between technical terms based on co-occurrences in keyword lists of academic papers. *Systems and Computers in Japan*, 34(3), 85-95
- Amuyunzu-Nyamongo, M. (2001). HIV/AIDS in Kenya: Moving beyond policy and rhetoric. *African Sociological Review*, 5(2)
- Baldwin, C., Hughes, J., Hope, T., Jacoby, R., & Zie bland, S. (2003). Ethics and dementia: Mapping the literature by bibliometric analysis. *International Journal of Geriatric Psychiatry*, 18, 41-54
- Bookstein, A., & Raita, T. (2001). Discovering term occurrence structure in text. *Journal of the American Society for Information Science*, 52(6), 476-486

- Bookstein, A., Kulyukin, V., Raita, T., & Nicholson, J. (2003). Adapting measures of clumping strength assess term-term similarity. *Journal of the American Society for Library Science and Technology*, 54(7), 611-620
- Callon, M., Courtial, J.-P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, 22(1), 153–203.
- Callon, M., Law, J., & Rip, A. (1986). *Mapping of the dynamics of science and technology*. London: Macmillan.
- Cohen, J. (2000b). Is AIDS in Africa a distinct disease? *Science*, 288(5474), 2153-2155
- Conlon, C. P. & Snydman, D. R. (2004). *Mosby's color atlas and text of infectious diseases*. Edinburgh: Mosby.
- Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: a study in co-word analysis. *Journal of the American Society for Information Science*, 49(13), 1206-1223.
- Courtial, J.-P. (1994). A co-word analysis of scientometrics. *Scientometrics*, 31(3), 251–260.
- Courtial, J.-P., & Law, J. (1989). A co-word study of artificial intelligence. *Social Studies in Science*, 19, 301–311.
- Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management*, 37, 817-842
- Hui, S. C., & Fong, A. C. M. (2004). Document retrieval from a citation database using conceptual clustering and co-word analysis. *Online Information Review*, 28(1), 22-32
- Jacobs, N. (2002). Co-term network analysis as a means of describing the informational landscapes of knowledge communities across sectors. *Journal of Documentation*, 58(5), 548-562
- Kopcsa, A., & Schiebel, E. (1998). Science and technology mapping: A new iteration model for representing multidimensional relationships. *Journal of the American Society for Information Science*, 49(1), 7-17
- Kostoff, R. N. (2001). Science and Technology Metrics. Retrieved April 11, 2002, from Defence Technical Information Center. Information for the Defence Community Website http://www.dtic.mil/dtic/kostoff/Metweb5_IV.htm
- Krsul, I. (2002). Co-word analysis tool. Retrieved on December 3^d, 2003, from <http://www.acis.ufl.edu/~ivan/coword/algorithmdescription.pdf>.
- Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics*, 23(3), 417–461.
- Leydesdorff, L. (1988). Words and co-words as indicators of intellectual organization. *Research Policy*, 18:209-223
- Leydesdorff, L. (2004). FullText.exe for Full text analysis. Retrieved June 18, 2006, from <http://www.leydesdorff.net>
- Me'decins Sans Frontie'res (MSF) (2003). HIV/AIDS. Retrieved July 18, 2005, from <http://www.accessmed-msf.org/campaign/hiv01.shtm>
- Nordberg, E. (ed). (2001). *Communicable diseases: a manual for health workers in Sub-Saharan Africa*. Nairobi: African Medical and Research Foundation.
- Onyancha, O.B. & Ocholla, D.N. (2005). An informetric investigation of the relatedness of opportunistic infection to HIV/AIDS. *Information Processing and Management*, 41, 1573-1588
- Schneider, J. W., & Borlund, P. (2004). Introduction to bibliometrics for construction and maintenance of thesauri: Methodical considerations. *Journal of Documentaion*, 60(5), 524-549
- Turner, W., Chartron, G., Laville, F., & Michelet, B. (1988). Packaging information for peer review: New co-word analysis techniques. In A. Van Raan (Ed.), *Handbook of quantitative studies of science and technology*, 291–323. Amsterdam: North Holland
- United Nations International Childrens Fund (UNICEF). (2003). Malaria and HIV/AIDS. Retrieved July 25, 2006, from <http://www.unicef.org/health/files/UNICEFTechnicalNote6MalariaandHIV.doc>
- Whittaker, J. (1989). Creativity and conformity in science: Titles, keywords, and co-word analysis. *Social Science in Science*, 19, 473–496
- World Health Organization. (2004). Malaria and HIV/AIDS interactions and implications: Conclusions of a technical consultation convened by WHO, June 23-25. Retrieved July 25, 2006 from http://www.who.int/hiv/pub/prev_care/en/WHO%20Malaria%20and%20AIDS.pdf
- Yitzhaki, M. (2001). Relation of title length of journal article to length of article. In: M. Davis & C. S. Wilson (eds.). *Proceedings of the 8th International Conference on Scientometrics and Informetrics*, Sydney, July, 16-20, 2, 759-769

Appendix: List of terms used to conduct co-word analysis of hiv/aids literature

Opportunistic Infections

Burkitt's Lymphoma
Cancer
Candidiasis
Carcinoma
Coccidioidomycosis
Cryptococcosis
Cryptosporidiosis
Cytomegalovirus
Encephalopathy
Herpes Simplex
Histoplasmosis
Immunoblastic Lymphoma
Isosporiasis
Kansasii
Kaposi's Sarcoma
Leukoencephalopathy
Lymphoma
Mycobacterium Avium Complex
Pneumocystis carinii
Pneumonia
Progressive Multifocal Leukoencephalopathy
Salmonella
Shigella
Staphylococcus aureus
Streptococcus pneumoniae
Toxoplasmosis
Tuberculosis
Varicella zoster
Wasting Syndrome

Pre-Disposing Factors

Alcoholism
Broken Marriage
Conflict
Culture
Disability
Discrimination
Drug Abuse
Gender
Handicapped
Ignorance
Illiteracy
Inequality
Labor Migration
Marginalization
Malnutrition
Orphans
Poverty

Primitivity
Rape
Refugees
Rural
Sanitation
Socioeconomic Factors
Substance Abuse
Underdevelopment
Uneducated
Unemployment
Urbanization
Violence
War

Risk Factors

Circumcision
Condom Attitudes
Drug Abuse
Extramarital sex
Gays
Genital Herpes
Gonorrhea
Heterosexuality
Homosexuality
Injections
Infected Mothers
Milk
Mother-to-infant transmission
Needlestick injury
Non-usage of Condoms
Oral Sex
Promiscuity
Prostitution
Rape
Saliva
Sex
Sexual Intercourse
Sexually Transmitted Diseases
Substance Abuse
Syphilis
Unprotected Sex

Sexually Transmitted Diseases

Bacterial Vaginosis
Candidiasis
Chancroid
Chlamydia
Condylomata Acuminata
Genital Warts
Gonorrhea
Granuloma Inguinale
Hepatitis B
Herpes Zoster
Human Papillomavirus Infection
Lymphogranuloma Venereum
Molluscum Contagiosum
Pediculosis Pubis
Pelvic Inflammatory Diseases
Pubic Lice
Scabies
Sexually Transmitted Diseases
Syphilis
Trichomonial Vaginalis
Trichomoniasis

Other Diseases (mainly tropical diseases)

Amebiasis
Cholera
Dengue
Ebola
Giardiasis
Guinea-Worm
Hepatitis
Hookworm
Hypertension
Jaundice
Leishmaniasis
Lymphatic Filariasis
Malaria
Malnutrition
Meningitis
Onchocerciasis
Polio
Schistosomiasis
Sickle Cell
Syphilis
Trypanosomiasis
Tuberculosis
Typhoid
Yellow Fever

Mapping the Structure and Evolution of Chemistry Research¹

Kevin W. Boyack*, Katy Börner** and Richard Klavans***

* *kboyack@sandia.gov*

* Sandia National Laboratories, P.O. Box 5800, MS-1316, Albuquerque, NM 87185 (USA)

** *katy@indiana.edu*

SLIS, Indiana University, 10th Street and Jordan Avenue, Bloomington, IN 47405 (USA)

*** *rklavans@mapofscience.com*

SciTech Strategies, Inc., Berwyn, PA 19312 (USA)

Abstract

How does our collective scholarly knowledge grow over time? What major areas of science exist and how are they interlinked? Which areas are major knowledge producers; which ones are consumers? Computational scientometrics – the application of bibliometric/scientometric methods to large-scale scholarly datasets – and the communication of results via maps of science might help us answer these questions. This paper represents the results of a prototype study that aims to map the structure and evolution of chemistry research over a 30 year time frame. Information from the combined Science (SCIE) and Social Science (SSCI) Citations Indexes from 2002 was used to generate a disciplinary map of 7,227 journals and 671 journal clusters. Clusters relevant to study the structure and evolution of chemistry were identified using JCR categories and were further clustered into 14 disciplines. The changing scientific composition of these 14 disciplines and their knowledge exchange via citation linkages was computed. Major changes on the dominance, influence, and role of Chemistry, Biology, Biochemistry, and Bioengineering over these 30 years are discussed. The paper concludes with suggestions for future work.

Keywords

mapping chemistry; journal mapping; dynamics; diffusion

Introduction

Chemistry is a field that is undergoing significant change. Interdisciplinary research has increased over time and the lines between Chemistry and the life sciences have seemingly blurred. Funding for chemistry-related activities now comes from more than just agencies and organizations that have been historically interested only in the physical sciences. For example, a long-time NIH (U.S. National Institutes of Health) grantee, Dr. Roger Kornberg of the Stanford University School of Medicine, was awarded the 2006 Nobel Prize in Chemistry, illustrating the reach of the life and medical sciences into chemistry.

This paper reports on a pilot study that we undertook to map the structure of *Chemistry* over time using journal citation patterns. Of particular interest were the interactions between mainstream *Chemistry* and the fields of *Biochemistry*, *Biology*, and *Bioengineering*, which were presumed to be impinging upon Chemistry. The balance of the paper proceeds as follows. First, we give a brief background on the mapping of science using journals. We then describe the data and processes used to generate our base map of science. Given our need to map the evolution of fields, we describe our method for linking unique journals from additional years into the base map. We then further characterize the maps, and conclude with a discussion of findings and suggestions for future work.

Background

Journals are a unit of analysis that allows one to understand how science is organized at an aggregated level (Leydesdorff, 1987). Thomson Scientific (TS, formerly ISI) has published the Journal Citation

¹. This work was supported by NSF award CHE-052466. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000. Color versions of all figures are available from the authors.

Reports (JCR) for many years now, compiling citation counts between journal pairs that allow for studies of the structure of science.

One of the pioneering journal maps looked at relationships among fields (Narin, Carpenter, & Berlt, 1972). Yet, the majority of such maps have typically focused on single disciplines (Ding, Chowdhury, & Foo, 2000; McCain, 1998; Morris & McCain, 1998; Tsay, Xu, & Wu, 2003). Recently, several larger-scale journal maps have been published. Leydesdorff (2004a) used the 2001 JCR data to map 5,748 journals from the Science Citation Index Expanded (SCIE) and 1,682 journals from the Social Science Citation Index (SSCI) (Leydesdorff, 2004b) in two separate studies. Leydesdorff uses a Pearson correlation on citing counts as the edge weights and the Pajek program for graph layout, progressively lowering thresholds to find articulation points (i.e., single points of connection) between different network components. These network components define journal clusters, which can be considered as disciplines or sub-disciplines. Samoylenko et al. (2006) mapped all journals in the SCIE with an impact factor of 5 or more using minimum spanning trees to show dominant linkages between fields. Leydesdorff (2006) has combined the SCIE and SSCI for a single study. Rather than generating a map of the entire set of journals, he generates centrality measures and shows them in the perspective of local citation environments (small sets of journals where citing is above a certain threshold). Boyack, Klavans & Börner (2005) combined the year 2000 SCIE and SSCI, generating maps of 7,121 journals. They studied the accuracy of maps generated using eight different inter-citation and co-citation similarity metrics, which were compared using an entropy-based measure.

Data and Methods

Prior to this mapping chemistry effort, two of the authors generated a journal-based map of science using the combined SCIE and SSCI from 2002. Although this map has not appeared in a peer-reviewed publication, it has nonetheless been shown in various capacities. In particular, it has been used as a base map on which funding information from several U.S. government agencies has been overlaid. This map, its structure, and the funding overlays are familiar to our project managers, and played a role in generating interest in this project. Thus, we chose to use this particular map as the base map for mapping the structure and evolution of Chemistry.

2002 Base Map

The 2002 journal map was generated using a new multi-step process. This process reduces the effect of over-aggregation due to highly-linked, multidisciplinary journals that tend to distort a journal map because they link to so many other journals in a variety of disciplines. It also helps place journals that might equally well fit in multiple journal clusters. The procedure is as follows:

- Bibliographic coupling counts were calculated at the paper level for the 1.07 million papers using the 24.5 million cited references indexed in the 2002 combined data set. These coupling counts were aggregated at the journal level (7,227 journals), thus giving bibliographic coupling counts between pairs of journals. The counts were then normalized using the cosine index to give a similarity value between 0 and 1 for pairs of journals.
- Using the top 15 similarity values per journal, the position of each journal was calculated using the VxOrd graph layout algorithm. Previous studies have established the accuracy of VxOrd with a variety of similarity measures for journal mapping (Boyack et al., 2005; Klavans & Boyack, 2006a). Details about the algorithms are also available elsewhere (Davidson, Wylie, & Boyack, 2001; Klavans & Boyack, 2006b).
- A breadth value was then calculated for each journal as $\text{SUM}(\text{distance} * \text{counts})$ where distance is the Euclidean distance on the graph, and the counts are the number of bibliographic coupling counts between pairs of journals, summed over all journals with which a particular journal has any counts. The breadth is thus an indicator of how tightly coupled a journal is in its local environment: a small breadth value means that the journal is very tightly coupled to its local environment, while a large breadth value means that the journal has substantial links outside of its local environment, and thus may be distorting the overall graph.
- Journals were ordered by descending breadth, and a scree plot of breadth vs. rank was used to find a natural break in the sequence. A break was identified after 25 journals. Thus, those 25 journals were labelled as distorting journals (e.g. *J Biol Chem*, *PNAS*, *JACS*, etc.).

- The 25 distorting journals were temporarily omitted from the bibliographic coupling matrix (and thus from the resulting map), and cosine values were recalculated for the remaining journals. Once again, using the top 15 similarity values per journal, the position of each journal was calculated using VxOrd. An average link clustering algorithm was then run using the journal positions and edges to generate a cluster solution. 646 clusters of journals were identified.
- The 25 distorting journals include many major journals that should not be omitted from a map of science. These journals were added back into the list, each as its own journal cluster. Thus, the number of journal clusters was now considered to be $646+25 = 671$.
- To produce the final visualization, the bibliographic coupling counts from all 7,227 journals were aggregated at the cluster level, cosine indexes were calculated, and the graph layout algorithm was run again, this time to generate positions for the 671 clusters of journals. A visualization of the clusters is more pleasing than a visualization of all 7,227 journals in that it is far less cluttered, and can show the dominant relationships between fields while preserving the white space that is important to cognition. The resulting visualization of the 671 journal clusters is shown in Figure 1 (top). Lines between the journal clusters indicate the strongest cosine linkages between journal clusters.

The 2002 base map represents journal cluster interrelations but is invariant to rotation and mirroring. The map was oriented to place mathematics at the top and the physical sciences on the right. The ordering of disciplines is similar to what has been shown in other maps of science (Boyack et al., 2005; Moya-Anegón et al., 2004; Small, Sweeney, & Greenlee, 1985): as one progresses clockwise around the map, one progresses from mathematics through the physical sciences (Engineering, Physics, Chemistry), to the earth sciences, life sciences, medical sciences, and social sciences. The social sciences link back to computer science (near the top of the map), which has strong linkages to mathematics and engineering.

Just like a map of the world can be used to communicate the location of minerals, soil types, political boundaries, population densities, etc., a map of science can be used to locate the position of scholarly activity. For example, as mentioned previously, the map shown in Figure 1 has been used to show funding patterns for various government agencies. The profiles for the U.S. NIH and NSF (National Science Foundation) are shown in Figure 1, and were calculated by matching the principal investigators and their institutions from grants funded in 1999 to first authors and institutions of papers indexed in 2002. This type of paper-to-grant matching will produce some false positives. Yet, on the whole it is a conservative approach in that it only considers a single time-lag between funding and publication (3 years in this case), and it does not match on secondary authors. The 14,367 NIH matches, and 10,054 NSF matches are large samples, ensuring that the aggregated profiles are representative of the actual funding profiles of the agencies. An entire paper could be written on these funding profiles and what can be learned from them; we choose not to do so here. Here it serves as a good example of how journal level, or disciplinary, maps can be used to display aggregated information obtained from paper-level analysis.

Maps for Additional Years

The 2002 base map is a static map, yet the goal of this study was to map *Chemistry* and the related fields of *Biology*, *Biochemistry*, and *Bioengineering*, and the changes in their structure and relationships over time. Thus we needed additional data and a way to visualize it in an easy to interpret way. As this is a pilot study, the acquisition of paper-level data for a 30-year study was not feasible due to the associated costs. Hence, it was decided to use journal-level data available in the JCR (which is much less expensive than paper-level data) to do a journal-level analysis and to overlay the results on the 2002 base map that is based on paper-level data, aggregated to journals and clusters.

Our project managers were interested in understanding the dynamics of chemistry over the last 30 years. We therefore obtained SCIE JCR data from TS for the years 1993-2004. To cover the years before 1993 we also obtained raw citing journal/cited journal “citation pairs” for the years 1974, 1979, 1984, and 1989 from the same data source. From these citation pairs we calculated JCR-like counts

between pairs of journals for those years. When combined with the counts data from the JCR for the years 1993-2004, this forms a standardized set of data from which science maps can be generated every five years over a period of 30 years.

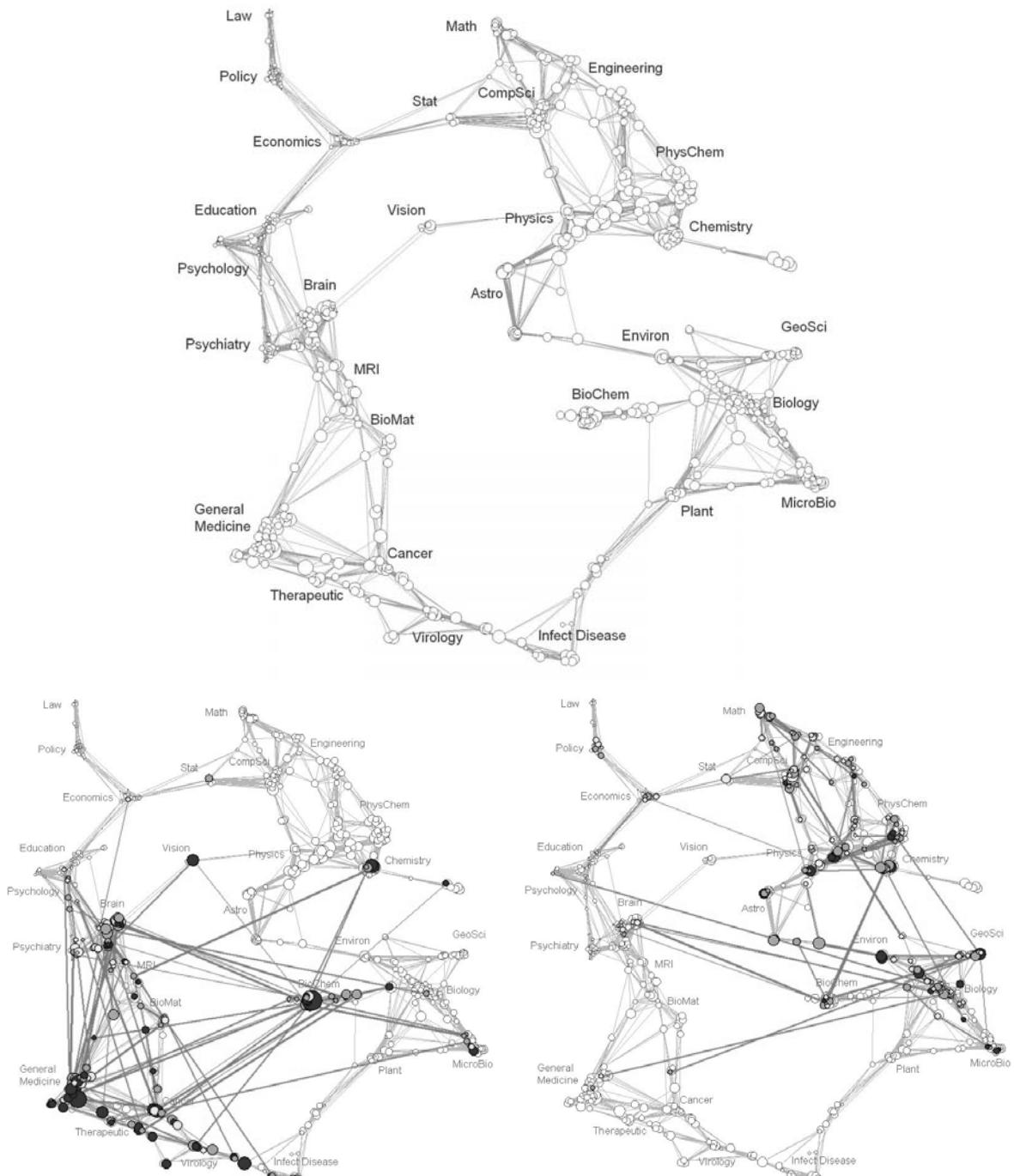


Figure 1. 2002 base map (top). Each node is a cluster of journals, and is sized to show numbers of papers in the journal cluster. NIH (bottom left) and NSF (bottom right) funding profile overlays on the 2002 base map. (Colored nodes show the distribution and numbers of papers tied to grants; red nodes indicate faster moving science than yellow nodes; colored edges show linkages in the funding profiles that are stronger than the corresponding linkages in the base map.)

When choosing to visualize science dynamics there are various options. Maps can be generated for different time periods, and be associated or morphed to communicate structural change (Chen, 2006). We consider this to be an area of research in and of itself. A second option is to use a static map and to

visualize the change in number of papers, citations, and inter-linkage strength using data overlays of changing size, shape, color, etc. This second option is much easier to read as the viewer only needs to understand one reference system, and it will be used here.

Use of a static map presented us with an additional challenge: the journal coverage of the TS databases changes over time. Hence, we needed a way to add 2,350 journals that were not covered in 2002 into the base map. Since we did not have paper-level data, we could not use the bibliographic coupling technique that formed the base map. We chose to use inter-citation data and the cosine index to determine which of the 671 clusters a journal should be added to, using the following process. For each of the years 2004, 1999, 1994, 1989, 1984, 1979, and 1974, in that order:

- Inter-citation counts were obtained for pairs of journals from the JCR-like data source described previously. For each journal pair, we defined inter-citation counts as the sum of the counts from journal A to B and journal B to A. Summing of counts in this way gives a symmetric count matrix with journals as rows and columns. Only those counts to years within the previous 9 years were included. (The JCR only lists counts to individual cited years for the previous 9 years.) For instance, for citing year 2004, all citations to cited years of 1995 and more recent were included, but citations to years 1994 and earlier were not.
- The columns in the count matrix were aggregated by journal cluster number where cluster numbers were available. This gives a matrix with journals as rows and clusters of journals as columns, and thus gives the citation counts of journals to clusters. Cosine index values were then calculated for this matrix, giving each journal-to-cluster a similarity value between 0 and 1. New journals, those not previously assigned to a cluster because they were not in the 2002 data, were then assigned to the clusters with which they had the largest cosine values. This technique makes use of the affinity of journals to an entire cluster rather than to single journals.

The result of this set of calculations was that each journal occurring in any of the data, from 1974-2004, was assigned to one of the 671 clusters of journals in the 2002 base map, thus allowing us to use the 671 clusters for each of the years in the study.

Mapping Chemistry

Once all journals were assigned, we characterized the four fields of interest in this study. This was done using JCR journal categories. Relevant JCR categories were grouped into one of our four fields using the breakdown shown in Table 1. The well-known journals *Science*, *Nature*, and the *Proceedings of the National Academy of Sciences of the USA*, although considered multidisciplinary journals, are in reality highly slanted toward biochemistry. Thus, they were included in the *Biochemistry* field. In addition, the category GC was not available in the data before 1994. Thus, any journal found in category GC in years 1994-2004 was also considered to be a Chemistry journal in the years before 1994.

We also accounted for the fact that many journals are classified in multiple categories by the JCR. For example, the journal *Bioelectrochemistry* has four different JCR category designations:

CQ – Biochemistry
CU – Biology

DA – Bioengineering
HQ – Chemistry

Since we have no detailed information that would allow us to know how much this journal falls into each of the categories, we assume a straight fractional basis. Thus, for the purpose of counting how many papers from *Bioelectrochemistry* should count toward each of our four fields, we count $\frac{1}{4}$ of the papers for each of the four fields. This journal is an extreme example. Most journals are only assigned to one or two categories.

Table 1. JCR categories comprising the fields of Chemistry, Biology, Biochemistry, and Bioengineering

Field	JCR Categories	
Chemistry	DW – Chemistry, Applied DX – Chemistry, Medicinal DY – Chemistry, Multidisciplinary EA – Chemistry, Analytical EC – Chemistry, Inorganic & Nuclear EE – Chemistry, Organic	EI – Chemistry, Physical HQ – Electrochemistry II – Engineering, Chemical GC – Geochemistry & Geophysics UH – Physics, Atomic, Molecular & Chem.
Biology	CU – Biology CX – Biology, Miscellaneous DR – Cell Biology HY – Developmental Biology	HT – Evolutionary Biology PI – Marine & Freshwater Biology QU – Microbiology WF – Reproductive Biology
Biochemistry	CO – Biochemical Research Methods CQ – Biochemistry & Molecular Biology	individual journals: Science, Nature, PNAS
Bioengineering	DA – Biophysics IG – Engineering, Biomedical	DB – Biotechnology & Applied Microbiol. QE – Materials Science, Biomaterials

Given the assignments of journals to clusters, fractional assignments of journals to the four fields of interest, and the number of papers per journal by year, we can calculate the number of papers in each of our four fields for each of the 671 clusters in each year. Figure 2 (left) shows the distribution of *Chemistry* papers on the 2002 map. Although there are some chemistry papers in the medicine area, and some in engineering, the large majority lie within the box that comprises the physics, chemistry, and life sciences portion of the map. Subsequently, we focus on that part of the map, which is shown in an enlarged view in Figure 2 (right), with distributions of papers from all four categories. However, the true fractional distributions cannot be easily discerned as nodes of one color lie on top of nodes of another color, causing partial or complete overlaps.

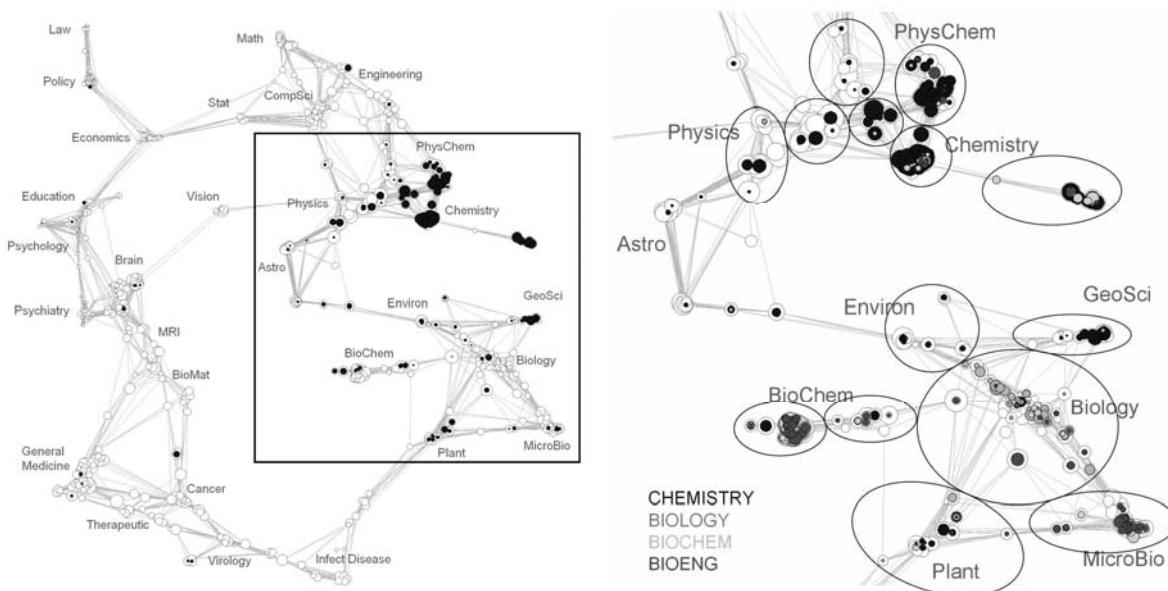


Figure 2. 2002 base map (left) with blue nodes showing the distribution and number of *Chemistry* papers. The inset map (right), also 2002, shows paper distributions for all four fields (*Chemistry*, *Biology*, *Biochemistry*, and *Bioengineering*) along with 14 hand-drawn groupings of 259 journal clusters (disciplines) that are used for further analysis.

In addition, with so many journal clusters, it would be difficult to characterize and visualize diffusion patterns. Thus, we decided to manually group journal clusters into higher-level groupings based on the natural aggregation of journal clusters, spacing between groups of journal clusters, and distributions of the papers of the four fields, as shown in Figure 2 (right). The areas in astrophysics were ignored due to the low chemistry content in that part of the map.

Paper counts for each of the four fields for each journal cluster were summed to give counts by field for each of the 14 groupings (hereafter called disciplines) shown in Figure 2 (right). Figure 3 shows the sizes of the 14 disciplines in 1974. Pie charts are used to show the fraction of papers in each of the four fields for each of the 14 disciplines, which have been labeled using their dominant ISI journal categories. Pie chart diameters are scaled by the square root of the number of papers; thus, the areas of the pie charts are accurate representations of the relative sizes of the disciplines.

Figure 3 also shows the flow of knowledge between pairs of the 14 disciplines. Knowledge flow occurs when one discipline cites another (Narin et al., 1972). Numbers of citations from each discipline to each other discipline were calculated from the original JCR and citation data. The source of the knowledge flow is the cited discipline, while the recipient of the knowledge flow is the citing discipline. Arrows in Figure 3 denote the flow of information from the source to the recipient of the knowledge. Arrows inherit the color of the knowledge source, and are proportional in thickness to the square root of the number of citations. There are knowledge flows between nearly all pairs of disciplines in the diagram; to avoid clutter a threshold of 500 citations was used to show only the dominant knowledge flows.

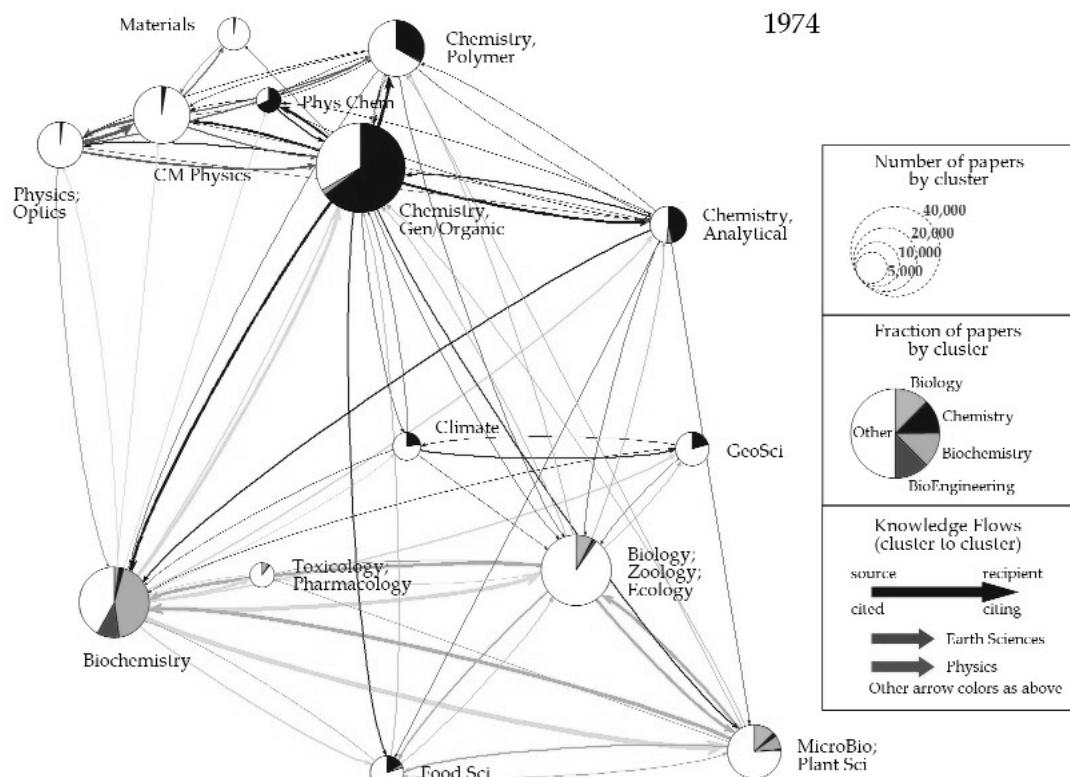


Figure 3. Map of the 14 disciplines, fractions of papers by field for each discipline, and knowledge flows between disciplines for 1974. (These 14 disciplines are further aggregated into six groups, represented by the 6 colors shown in the legend.)

The map in Figure 3 can be interpreted as follows. The majority of *Chemistry* papers are found in the four chemistry-dominated disciplines at the upper right of the diagram. The Gen/Organic Chemistry discipline is the largest, and also has a high fraction of chemistry papers. The Physical Chemistry discipline is the smallest of the chemistry disciplines, but is comprised of about 70% *Chemistry* papers. The remaining 30% of the papers are primarily in physics journals or journals that have both chemistry and physics designations. The three disciplines at the upper left of the diagram have only small fractions (<5%) of chemistry papers, and are primarily composed of Physics and Materials Science papers.

The lower half of the diagram is composed of the earth science (Climate and Geosciences), biology-related, and biochemistry-related disciplines. *Chemistry* is a player in both Climate Science and Geosciences, with around 20% of the papers. *Chemistry* is also a significant part of the Food Science discipline. However, in 1974, *Chemistry* had very little presence in the Biochemistry, Biology, or Microbiology disciplines.

The Biochemistry discipline, although not the largest, has the largest knowledge flows to and from other disciplines, and is a net donor of knowledge; the arrows going out from Biochemistry are thicker than the arrows coming into Biochemistry. In 1974, the dominant chemistry-related knowledge flows were between the four chemistry disciplines. However, the Analytical Chemistry and General Chemistry disciplines were significant sources of knowledge for the Biochemistry discipline. There were also relatively strong flows from General Chemistry to the Food Science and Microbiology disciplines.

Similar diagrams of the 14 disciplines have been created at 5-year intervals to show the changes in size, fractional distribution, and knowledge flow over a 30-year period, and are shown sequentially in 6 different charts comprising Figure 4. The scales for discipline and knowledge flow size have been kept constant in Figures 3 and 4 to enable easy visual inspection of any changes.

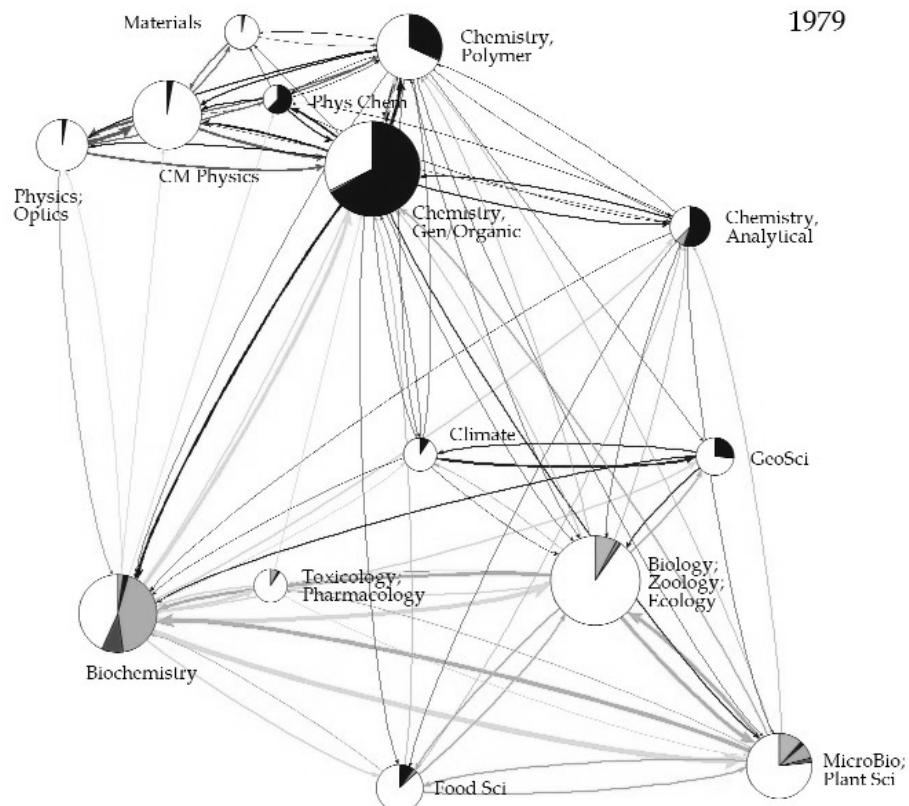
A close inspection of Figures 3 and 4 reveals many changes, more than we can attempt to describe here. Thus, we highlight the dominant features and changes, especially with regard to impacts on the field of *Chemistry*. First, all of the disciplines grow more or less consistently over time. The knowledge flows also grow, but at a higher rate than the growth in publications.

30 years ago, *Bioengineering* had almost no presence outside the Biochemistry discipline. As of 2004, *Bioengineering* had not only increased its presence in the Biochemistry discipline, but had gained a significant role in Microbiology. In addition, *Bioengineering* is starting to be seen in three of the *Chemistry* disciplines: Polymer Chemistry, Analytical Chemistry, and General Chemistry. *Biology* has increased its fractional presence in the Biochemistry, Biology, and Microbiology disciplines, but has not yet gained a foothold in the chemistry disciplines. However, its influence in Chemistry has increased significantly, as shown by the growth in the knowledge flows (green arrows) from the Biology and Microbiology disciplines to the chemistry disciplines. On the whole, *Biology* supplies more base knowledge to *Chemistry* than *Chemistry* does to *Biology*.

The *Biochemistry* field has more or less maintained its place in the Biochemistry, Biology, and Microbiology disciplines over time. However, it has made steady gains in Analytical Chemistry, comprising roughly 20% of its papers in 2004. *Biochemistry*'s presence in General Chemistry is also starting to grow, although it is still small.

As for the field of *Chemistry*, it has maintained or grown its presence in its home disciplines. It has increased its presence in the Geosciences and in Toxicology, but now plays a much smaller role in Climate Science. Interestingly, the knowledge flows from chemistry disciplines to non-chemistry disciplines have not grown as quickly as the knowledge flows from other disciplines into *Chemistry*.

1979



1984

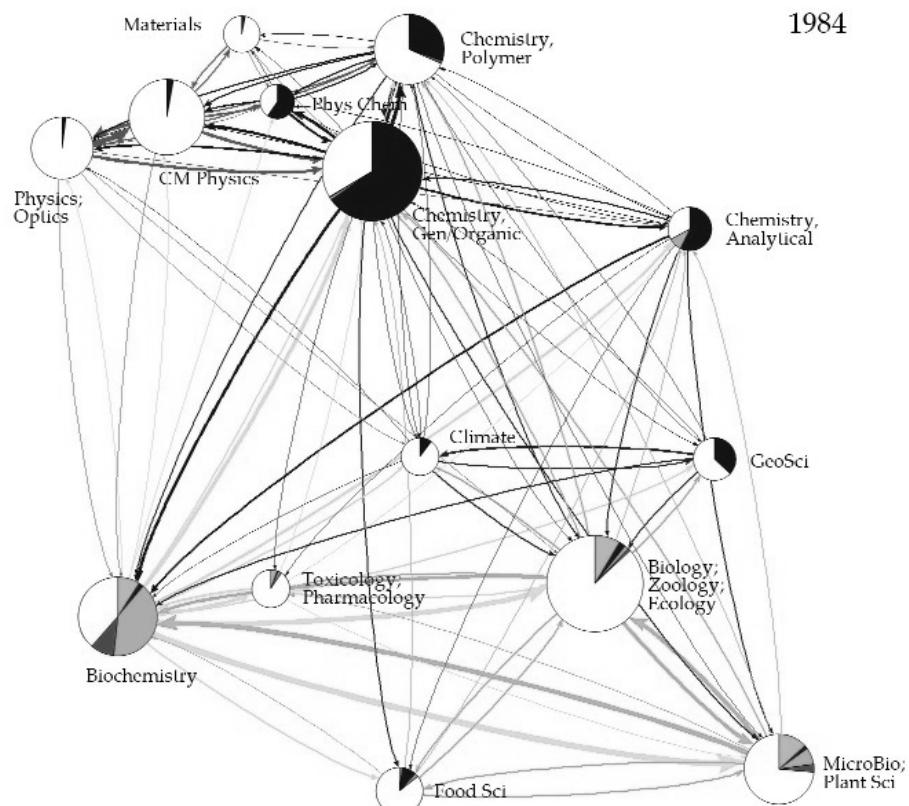


Figure 4a. Maps of the 14 disciplines, fractions of papers by field for each discipline, and knowledge flows between disciplines for 1979 and 1984. (The legend is found in Figure 3.)

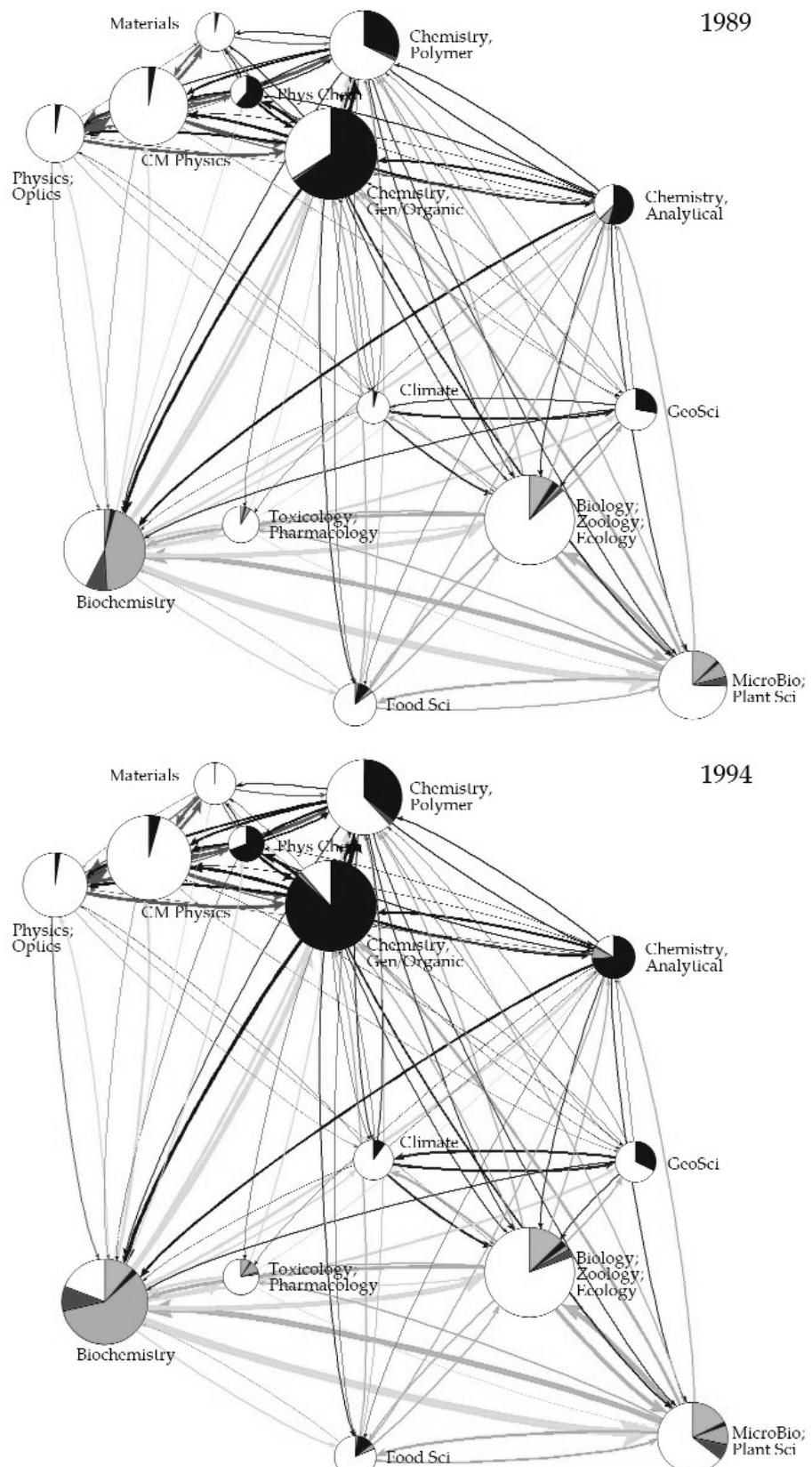


Figure 4b. Maps of the 14 disciplines, fractions of papers by field for each discipline, and knowledge flows between disciplines for 1989 and 1994. (The legend is found in Figure 3.)

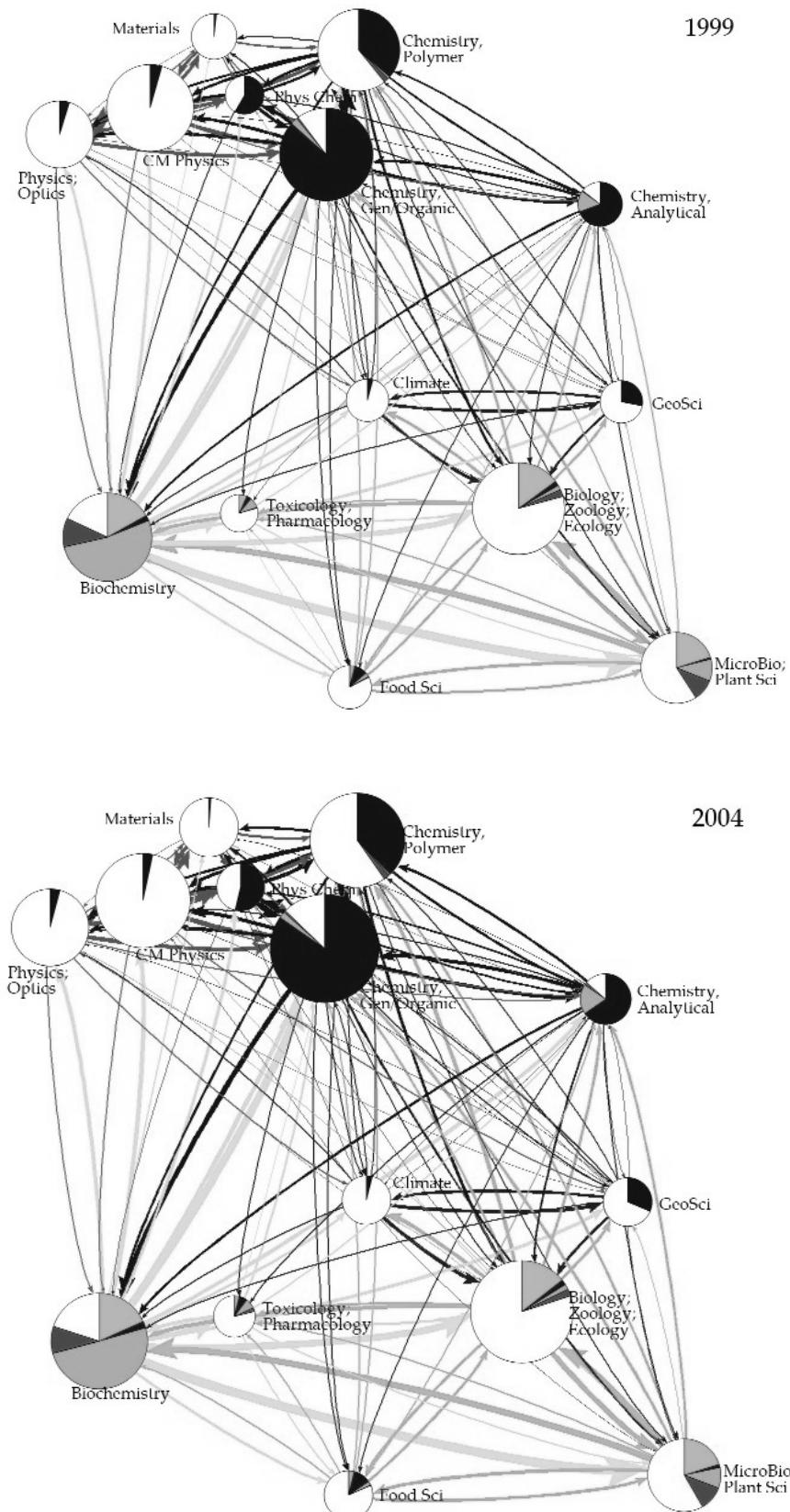


Figure 4c. Maps of the 14 disciplines, fractions of papers by field for each discipline, and knowledge flows between disciplines for 1999 and 2004. (The legend is found in Figure 3.)

Conclusions

Maps showing the growth, distribution, and knowledge flows between *Chemistry*, *Biology*, *Biochemistry*, and *Bioengineering* have been generated from journal-level data, and show many of the changes that have taken place over the past 30 years. Large trends can be seen, suggesting that *Biochemistry* and *Bioengineering* are moving steadily into *Chemistry* territory, and are having a large influence on the general knowledge base. *Chemistry*'s impact on the knowledge base is growing, but at a slower rate. However, journal-level data provide no information about the topics at the interface between fields, thus limiting the strategic decisions that can be made based on the mapping exercise.

Folding in patent and/or commercial data would provide a basis to study the impact of research on innovation and product development. It might very well be the case that some areas of science change their impact from a generator of cited scholarly knowledge to a generator of commercially valuable and hence patented and/or disclosed knowledge. An additional study should be done using paper-level data that can identify topics on the interfaces between fields, knowledge flows at topical levels, and detailed trends at these micro-levels. Paper-level data would also support the analysis of the trajectories and impact of single researchers, teams, institutions, or nations. Correlation of these data with funding data may further support strategic decisions by both funding agencies and researchers.

References

- Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351-374.
- Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377.
- Davidson, G. S., Wylie, B. N., & Boyack, K. W. (2001). Cluster stability and the use of noise in interpretation of clustering. *Proceedings IEEE Information Visualization 2001*, 23-30.
- Ding, Y., Chowdhury, G., & Foo, S. (2000). Journal as markers of intellectual space: Journal cocitation analysis of information retrieval area, 1987-1997. *Scientometrics*, 47(1), 55-73.
- Klavans, R., & Boyack, K. W. (2006a). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2), 251-263.
- Klavans, R., & Boyack, K. W. (2006b). Quantitative evaluation of large maps of science. *Scientometrics*, 68(3), 475-499.
- Leydesdorff, L. (1987). Various methods for the mapping of science. *Scientometrics*, 11(5-6), 295-324.
- Leydesdorff, L. (2004a). Clusters and maps of science journals based on bi-connected graphs in the Journal Citation Reports. *Journal of Documentation*, 60(4), 371-427.
- Leydesdorff, L. (2004b). Top-down decomposition of the Journal Citation Report of the Social Science Citation Index: Graph- and factor-analytical approaches. *Scientometrics*, 60(2), 159-180.
- Leydesdorff, L. (2006). *Betweenness centrality as an indicator of the interdisciplinarity of scientific journals*. Paper presented at the 9th International Conference on Science & Technology Indicators.
- McCain, K. W. (1998). Neural networks research in context: A longitudinal journal cocitation analysis of an emerging interdisciplinary field. *Scientometrics*, 41(3), 389-410.
- Morris, T. A., & McCain, K. W. (1998). The structure of medical informatics journal literature. *Journal of the American Medical Informatics Association*, 5(5), 448-466.
- Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Muñoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61(1), 129-145.
- Narin, F., Carpenter, M., & Berlt, N. C. (1972). Interrelationships of scientific journals. *Journal of the American Society for Information Science*, 23(5), 323-331.
- Samoylenko, I., Chao, T.-C., Liu, W.-C., & Chen, C.-M. (2006). Visualizing the scientific world and its evolution. *Journal of the American Society for Information Science and Technology*, 57(11), 1461-1469.
- Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the Science Citation Index using co-citations. II. Mapping science. *Scientometrics*, 8(5-6), 321-340.
- Tsay, M.-Y., Xu, H., & Wu, C.-W. (2003). Journal co-citation analysis of semiconductor literature. *Scientometrics*, 57(1), 7-25.

Using Detailed Maps of Science to Identify Potential Collaborations¹

Kevin W. Boyack

kboyack@sandia.gov

Sandia National Laboratories, P.O. Box 5800, MS-1316, Albuquerque, NM 87185 (USA)

Abstract

Research on the effects of collaboration in scientific research has been increasing in recent years. A variety of studies have been done at the institution and country level, many with an eye toward policy implications. However, the question of how to identify the most fruitful targets for future collaboration in high-performing areas of science has not been addressed. This paper presents a method for identifying targets for future collaboration between two institutions. The utility of the method is shown in two different applications: identifying specific potential collaborations at the author level between two institutions, and generating an index that can be used for strategic planning purposes. Identification of these potential collaborations is based on finding authors that belong to the same small paper-level community, using a paper-level map of science from the combined 2003 SCIE/SSCI/Proceedings databases containing nearly 1 million papers organized into 117,435 communities.

Keywords

mapping science; paper-level maps; research communities; vitality; collaboration

Introduction

It is a well known fact that scientific collaboration has been increasing over time. Papers co-authored by researchers at more than one institution make up an ever increasing fraction of all scientific papers, accounting for nearly 60% of papers in 2003 (National_Science_Board, 2006). In concert with this increase, measurement of collaboration and investigation of its impact has also increased. Much of this recent work is based on the measurement of co-authorship (Melin & Persson, 1996) and the resulting networks (Newman, 2001) at institutional (Havemann, Heinz, & Kretschmer, 2006), regional, or national levels (Glänzel & Schubert, 2001). Despite this body of work, few have asked the question “Who should I collaborate with?” from a strategic viewpoint. Identification of the best collaboration opportunities is an important part of institutional strategy and planning.

Techniques have recently been developed for clustering very large segments of the technical literature using sources such as Thomson Scientific’s Science Citation Index (Boyack, Klavans, & Börner, 2005; Klavans & Boyack, 2006a, 2006b). The primary objective of this work has been to develop indicators of potential impact at the paper level. Indicators aggregated at different levels enable profiling of departments, institutions, agencies, etc., and are useful for institutional planning and evaluation of research. This work is often presented as maps of science and technology with various overlays corresponding to the indicators associated with a particular search or question. Such maps of science, if created at a highly detailed level, are suitable for identifying potential opportunities for collaboration.

Here we report on two advances. First, given that the author’s institution (Sandia National Laboratories) is interested in collaborations in technology as well as science, we constructed a map of science and technology using the Science Citation Indexes and the ISI Proceedings database. The Proceedings database provides a technology component not present in the standard citation indexes. This new science and technology map shows the impact of including Proceedings in a science mapping effort. Second, once this map was constructed, it was used to identify potential collaborations in two different ways: a) specific collaboration opportunities between Sandia and the University of Texas system were targeted, and b) a general collaboration potential index ranking all U.S. universities in areas of interest to our institution was generated. This paper will describe methods and results in each of these areas.

¹ Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000..

Science and Technology Map

The science and technology map to be described here has both paper-level and journal-level components. Paper-level and journal-level science mapping efforts have long histories that are covered in detail elsewhere (Börner, Chen, & Boyack, 2003). The majority of this history has dealt with small maps covering specific topics or journal sets. It has only been recently that much effort has been expended on mapping “all of science”. Background on journal-level (Boyack, Börner, & Klavans, 2007) and paper-level (Klavans & Boyack, 2006b) efforts to map yearly segments of the literature is available elsewhere.

Our composite map of science and technology was generated using a multi-step process, and consists of two maps calculated at different levels and joined by coupling between the levels. The first map is a disciplinary map, a journal-based map, which shows the various disciplines in science and technology and their relationships to one another. This disciplinary map provides an easily understandable visual picture of science and technology as a whole, and is the template upon which query results can be presented. The second map is a paper-level map, generated by clustering the individual papers.

Disciplinary Map

Data from the combined 2003 Thomson Scientific’s Science Citation Index Expanded (SCIE), Social Science Citation Index (SSCI), and Proceedings databases were used to generate our composite map of science and technology. These data consisted of 1.35 million records (papers) from 7,506 journals and 1,206 conference proceedings, and contained a total of 29.23 million references. The process for calculating the disciplinary map is very similar to that shown in (Boyack et al., 2007):

- Bibliographic coupling (BC) counts were calculated at the paper level for the 1.35 million papers using the 29.2 million references in the combined data set. These were aggregated at the journal level (with proceedings considered as journals), thus giving bibliographic coupling counts between pairs of journals, and then normalized using the cosine index.
- Using the top 15 similarities per journal, journal positions were calculated using the VxOrd graph layout algorithm (Klavans & Boyack, 2006b). Minimum distances were calculated for each journal to its nearest neighbour in the graph, and this distribution was used to calculate a threshold value.
- Journals were ranked by summed bibliographic coupling counts, and the resulting scree plot was consulted to break the journals into four different groups, which will be detailed below. There was a distinct break in the scree plot at 40 journals, and there was a knee in the lower part of the curve at approximately 1000 counts.
 - MD – the 40 journals with the highest total BC counts were labeled as distorting journals because of their high level of linkage, and were temporarily omitted from the next portion of the map calculation. The first 10 journals on this list were *J Biol Chem*, *PNAS*, *J Chem Phys*, *Biochemistry-US*, *JACS*, *J Phys Chem A*, *Biochem Biophys Res Co*, *J Bacteriol*, *Phys Rev B*, and *J Mol Biol*. Notably, two proceedings were in the top 20: *Lect Notes Comp Sci* and *P Soc Photo-Opt Ins. Science and Nature*, two journals that one would expect to be on this list, were not, both appearing in the top 50-100 range.
 - FLOAT – 435 journals with fewer than 1000 total BC counts and with a minimum distance below the threshold were included in the calculation, but were assigned to the single journal with which they had the highest cosine relationship. These journals were thus tag-alongs, and were not allowed to be single entities in the next stage of the calculation.
 - REMOVE – 45 journals with fewer than 1000 total BC counts and a minimum distance above the threshold were not included in the calculation. These were excluded since they had few counts and did not form a close affiliation with a single cluster of journals as evidenced by a large distance value.
 - OTHER – the remaining 8,192 journals were fully included in the balance of the calculation.
- Cosine values were recalculated using the matrix of bibliographic coupling counts between the set of 8,192 journals identified in the previous step. MD journals were left out of this phase of the calculation so that they would not over-aggregate the journal graph. Counts from the

FLOAT journals were included with the journals they were assigned to. Once again, using the top 15 similarities per journal, positions were calculated with the Vxord graph layout algorithm. The resulting positions and edges were used to generate a cluster solution using a modified average-link clustering algorithm (Klavans & Boyack, 2006b). 812 clusters of journals were thus identified.

- The 40 MD journals were added back into the calculation at this point. Each was considered to be its own cluster, thus giving $812+40 = 852$ clusters. Bibliographic coupling counts were then aggregated at the cluster level, cosine indexes were calculated, and the graph layout algorithm was run once more to generate positions for the 852 clusters. The resulting visualization is shown in Figure 1 (upper).

This journal cluster map (hereafter called a disciplinary map) was oriented to place mathematics at the top and the physical sciences on the right, corresponding to the convention established in our previous mapping efforts (Boyack et al., 2005). A disciplinary map generated using a nearly identical process from the 2002 SCIE/SSCI, but without the Proceedings, is shown in Figure 1 (lower) for comparison.

Juxtaposition of the 2003 map including Proceedings with a 2002 map omitting Proceedings allows examination of the effect of the ISI Proceedings database on a disciplinary science map. The high-level similarity between maps is striking – the ordering of fields is very similar with the physical sciences and engineering at the upper right of each map, the earth and biological sciences at the right, medical fields at the bottom and lower left, and the social sciences at the upper left.

However, the impact and weight of the Proceedings database can also be seen. The 2003 map shows where the Proceedings papers appear in the map through the use of colored nodes: the darker the node, the higher the fraction of Proceedings papers in the node. A majority of the Proceedings papers are in the Computer Science (CS) region of the map, with a significant number in the space between CS and Physics. This includes *P Soc Photo-Opt Ins*, which is the large black node just above the Physics label. *Lect Notes Comp Sci* comprises another black node in the middle of the CS region, but it is hidden by smaller nodes appearing on top of it. Other areas of science in which Proceedings (at least those indexed by ISI) play a non-trivial role include Physics, Engineering, some areas in Earth Science, and Statistics (the darker area at the interior of the Social Sciences region). There are several fields in which the Proceedings add very little additional information, and thus have almost no impact. These include Chemistry, Biology, most of the Medical Sciences, and the Social Sciences. These observations correlate well with the analysis by Glänzel et al. (2006).

The weight of the added papers in CS and Physics also have an impact on the structure of the map. With the addition of the Proceedings, the number of papers in the CS area has roughly tripled over that shown in the 2002 map, and the number of journal clusters has roughly quadrupled. Accordingly, when using our force-directed graph layout algorithm to place the journal clusters, the additional clusters in CS have enough influence to push other fields away. Thus, we see more space between CS and Math, CS and Engineering, and CS and Statistics in the 2003 map than in the 2002 map. In addition, the overall weight of the additional papers in the combined CS, Physics, and Engineering areas have pushed the Earth Sciences further to the right.

Inclusion of the Proceedings database in a map of science has implications on the research planning and evaluation process. The distribution of Proceedings papers suggests that they can be ignored in many fields of science, but must be included in any exercises including the fields of CS, Physics, or Engineering. This is particularly true in CS, where the culture is to publish in Proceedings rather than in journals.

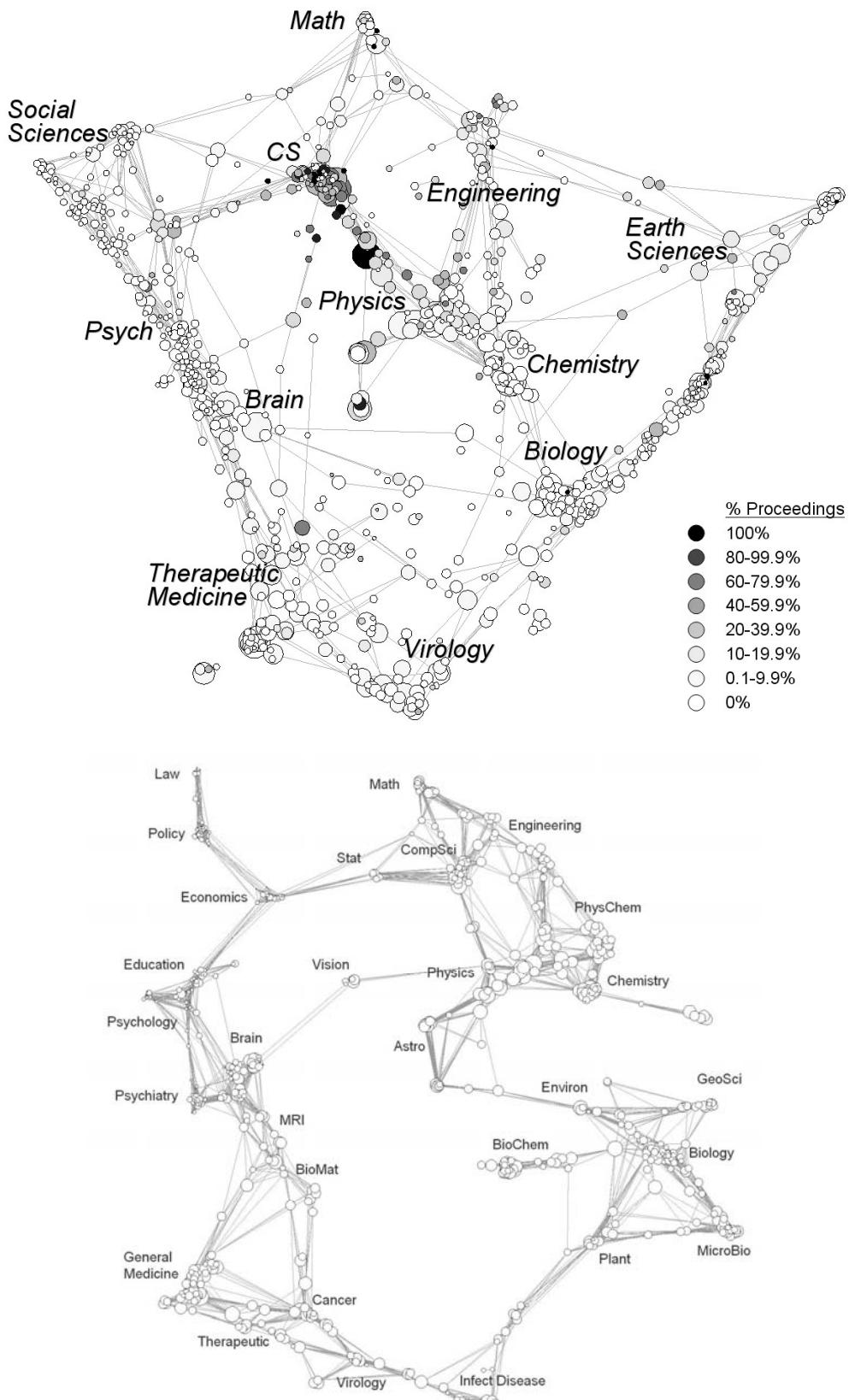


Figure 1. 2003 disciplinary map (upper) including SCIE/SSCI/Proceedings databases. (Each node is a cluster of journals, and is sized to show numbers of papers by cluster. Node colors correspond to the fraction of papers coming from Proceedings as opposed to journals. 2002 disciplinary map (journal clusters) without Proceedings (lower). Node sizes in the two maps are of different scales).

Paper-level Map

The second portion of our combined map of science and technology was generated from the individual papers themselves. A very low threshold was used to limit the map to those papers that could reasonably be expected to contain some sort of scientific advance; papers with at least two bibliographic coupling counts (two co-occurrences in the reference lists) to another paper in the set were included. Of the original 1.35 million papers, 997,775 papers were included in this map. Bibliographic coupling counts were then normalized using the cosine index. The top 10 similarities (cosines) per paper were used as input to the VxOrd algorithm, which calculated positions for each paper. A modified average link clustering algorithm was then used to assign papers to clusters based on distances and the existence of edges (or a link in the top 10 similarity file). A total of 117,435 clusters of papers were identified using this method.

A cluster of papers, hereafter referred to as a research community, is the unit of analysis that will be used in the balance of the paper, and generally represents a single research topic. Previous work has shown that the topical coherence of these clusters is very high (Klavans & Boyack, 2006b). Statistical distributions related to the communities are shown in Figure 2. The graph layout of our paper-level map is not shown here; indeed, if the resulting layout of 1 million nodes were viewed on a 1 megapixel display, each node would be represented by one pixel. We find it much more instructional to display the research communities on the disciplinary map. Positions for each community are calculated according to the journal distribution in the community. For instance, if a community has 7 papers in journal cluster A, and 3 papers in journal cluster B, the position of the community would be calculated as $x = 0.7^2 x_A + 0.3^2 x_B / (0.7^2 + 0.3^2)$. A similar calculation would ensue for the y-dimension. Squares are used on the fractional components to keep the communities near their dominant journal locations.

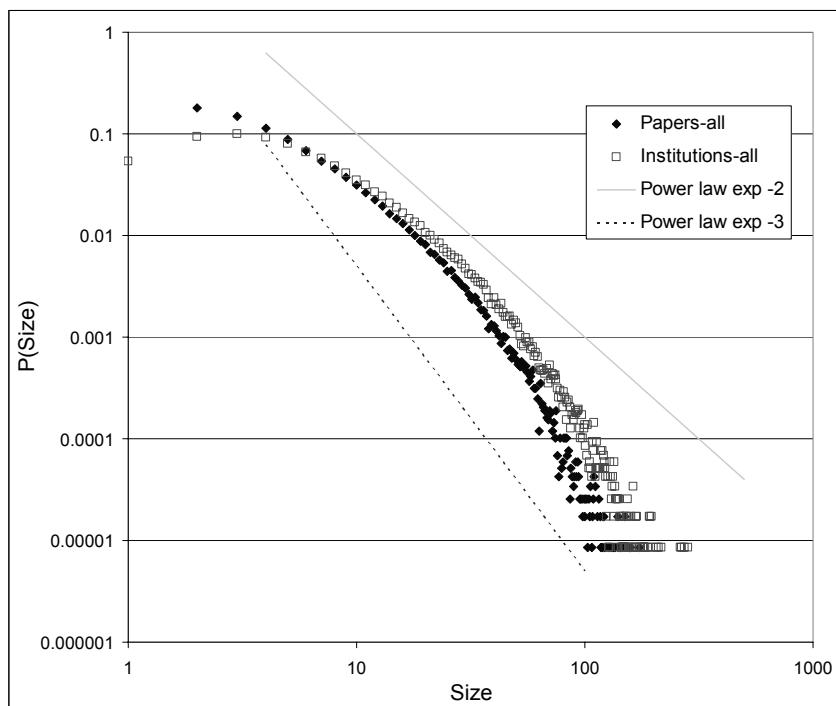


Figure 2. Distributions of numbers of papers and research institutions for the 117,435 communities. Power law slopes for exponents of -2 and -3 are shown for comparison.

The distribution of communities is shown in Figure 3, where the disciplinary map is superimposed on top of the individual communities. This map is interesting in that it shows in a visual way the interdisciplinary nature of science, and the relative interdisciplinarity of different fields. For example, the Medical Sciences are much more interdisciplinary than is Physics, as shown by the relative densities of the communities in between nodes in the disciplinary map. There is also a high degree of

interdisciplinarity between Chemistry and Biology as indicated by the highly dense region of communities between the large groups of journal clusters in those two areas.

Using the Map to Identify Collaboration Potential

Institutional Profiles

Although Figure 3 shows communities in representative positions, we also assign each community to a single journal cluster for purposes of aggregation and presentation of the results of an analysis. Communities are assigned to their dominant journal clusters. Thus, for our example in which a community has 7 papers in journal cluster A, and 3 papers in journal cluster B, the community would be assigned to journal cluster A. In the case of ties, the community is assigned to the smaller journal cluster since it has the higher fractional relationship. We can then use the disciplinary map to display information related to the communities, aggregated to the journal cluster level.

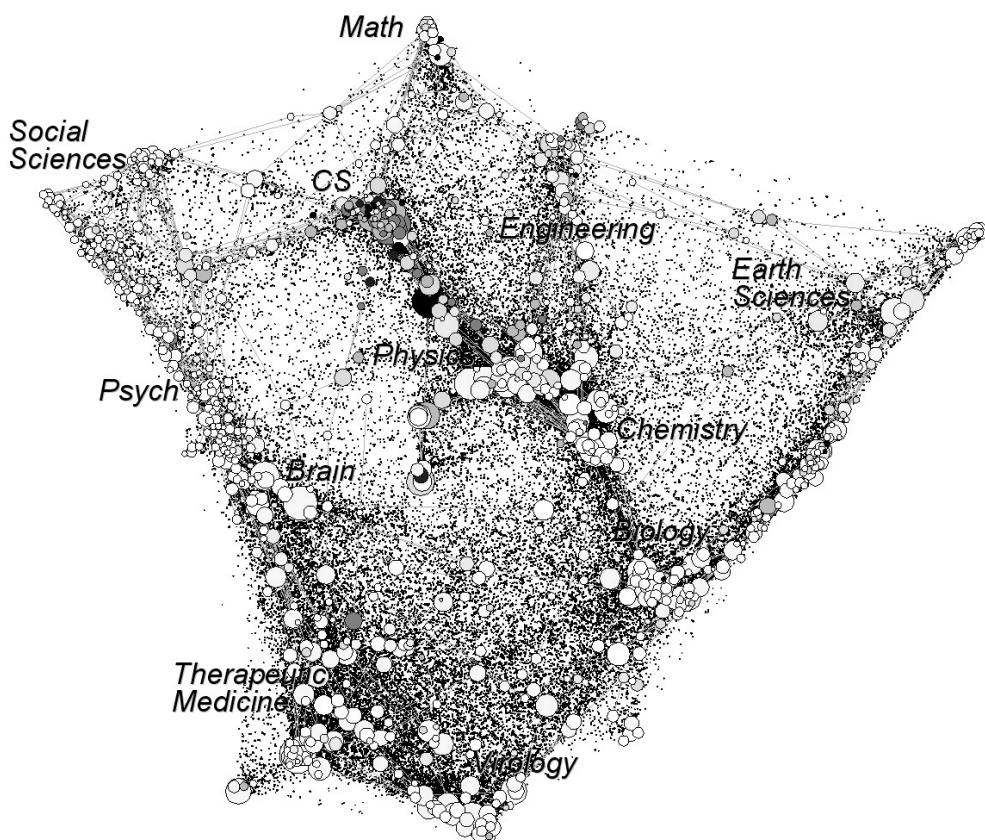


Figure 3. Combined science and technology map from 2003, showing journal nodes, as in Figure 1, and communities or clusters of papers. Each small black dot is one community.

One primary purpose for such a map is to show institutional profiles. Figure 4 shows the publishing activity in 2003 for two different institutions as overlaid on the 2003 map of science and technology. To generate an institutional overlay, all papers authored by the institution are identified. A list of the communities to which those papers are assigned is then generated, and the number of communities in each discipline is counted. Figure 4 displays the sizes and the vitalities of each of the journal clusters for the two institutions. Node size indicates the number of communities by cluster, while node color indicates the relative vitality of the communities in which the institution is active within the cluster.

Vitality is a measure that is related to the average age of all cited references from the papers in a community (Klavans & Boyack, submitted). The vitality for community c is calculated as:

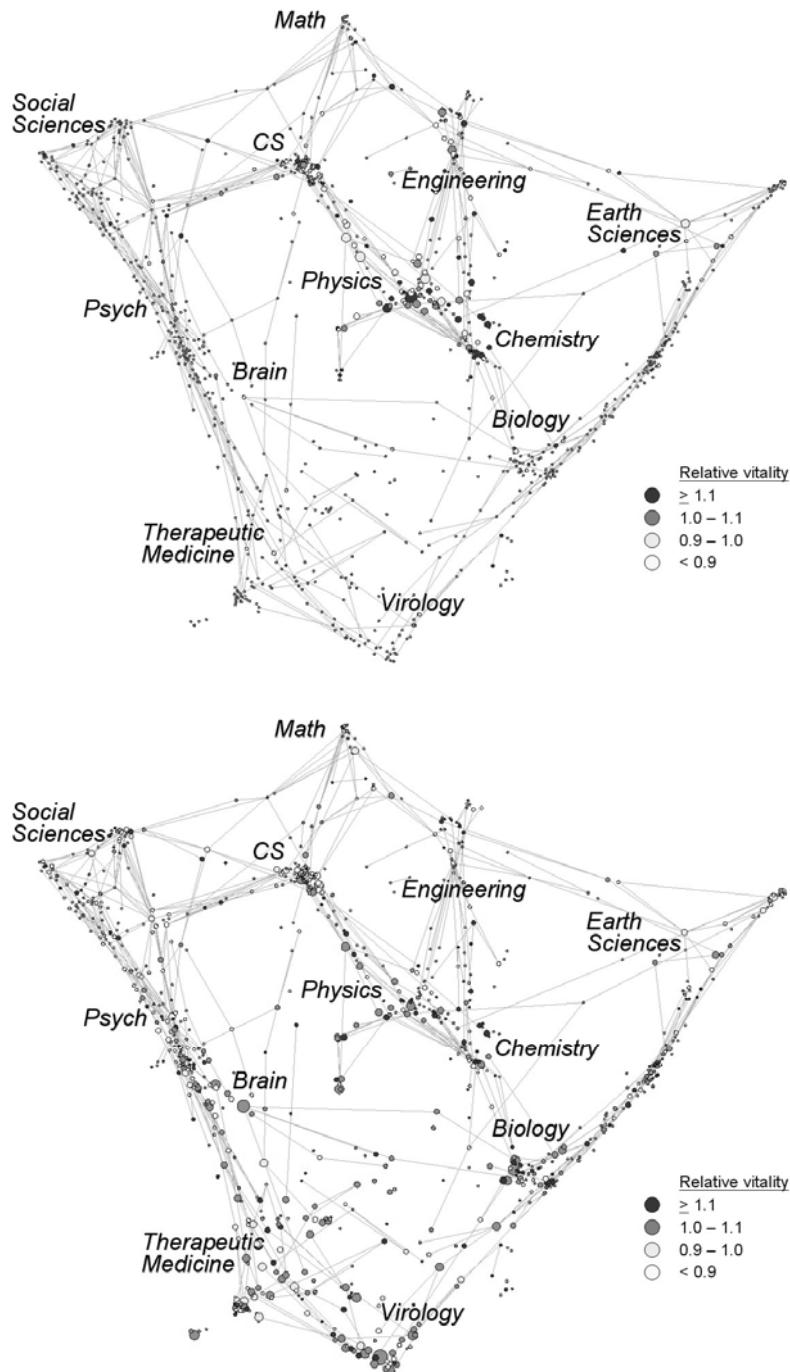


Figure 4. Publishing profiles for two institutions, Sandia National Laboratories (top) and the University of Texas system (bottom), overlaid on the disciplinary map.

$$V_c = \frac{1}{n} \sum_{j=1}^n \left(\frac{1}{Age_j + 1} \right),$$

where n is the number of references from all current papers assigned to a community, and Age_j is the age of reference j in years. Vitality is thus bounded between zero and one. Communities that refer to more recent research (in the form of younger papers) have a higher vitality. The research in those topics is updating itself more quickly. Highly vital topics are, thus, fast moving areas of research. Vitality is one metric by which we can compare different communities within a discipline or journal

cluster. A high vitality does not necessarily mean that the research in one community is better than that in another, but merely that it is in a faster moving, or more vital, topic. We do not compare vitalities between disciplines because different disciplines in science and technology have different citing cultures, and thus different natural vitalities. Rather, we calculate the mean vitality for each discipline and use it as a reference vitality. The relative vitality for an institution in a particular discipline, or journal cluster, is thus the ratio of the average vitality for the communities in which the institution has published to the reference vitality for that discipline.

For instance, suppose that an institution published in three communities in a particular discipline, with vitalities of 0.20, 0.23, and 0.26, and suppose that the reference vitality for that discipline was 0.20. The relative vitality would be $\text{Avg}(0.20, 0.23, 0.26) / 0.20 = 1.15$. In this case, the vitality of the institution would be 15% greater than that of the world at large for the particular discipline. Using our color scale from Figure 4, this vitality would merit a red node.

The institutional profiles of Figure 4 show two very different types of institutions. Sandia is centered in physics, chemistry, engineering and computer science. In general, it has a higher than average vitality (red and orange circles) in many of its research areas. However, it has a lower than average vitality (yellow and white dots) in some of the areas around physics, particularly between physics and computer science.

By contrast, the profile for the University of Texas shows that this university system publishes in nearly all areas of science and technology. This is no surprise; one would expect a large university system to have departments in nearly all potential fields. However, the work of this university system encompasses a range of vitalities. Their communities in physics, biology, and virology tend to be of higher vitality, while communities in the rest of the sciences and technology are a mixed bag, with some higher vitality areas and some lower vitality areas. For example, in computer science, while there are some high vitality areas, the majority of the journal clusters are of lower than average vitality (yellow nodes).

Identifying Targets for Future Collaborations

Communities can also be used as the basis for identifying both existing and potential future collaborations. The case of existing collaborations is trivial in that we are simply identifying papers co-authored by two institutions. We do not need a map of science or a list of research communities for this, although the map does provide a good visual template on which to display the results.

By contrast, identification of targets for future collaboration does require some sort of very detailed clustering at the paper level. We need a way to identify researchers who could easily collaborate because they are working on the same topic. Our paper level mapping and the clustering of papers into communities provides a means to identify those researchers who are working on the same focused topics.

We define a ‘potential collaboration’ as a research community in which two institutions have each authored papers. This includes co-authored papers because although there is already an existing collaboration, it is nonetheless a collaboration that may continue into the future, and is thus a potential future collaboration as well. Given that research communities are tightly focused around topics, and are based on common referencing patterns, it is highly likely that the researchers in a community know each other either directly through conference attendance or common associates, or indirectly by reputation or reading each others’ work. Researchers from a single community are those who could easily collaborate with each other given common interest, expertise, and past research activity. From a personal or an institutional standpoint, one can thus identify potential collaborators as the researchers in those communities in which the person or institution publishes.

Figure 5 shows the existing and potential collaborations between Sandia National Laboratories and the University of Texas system, along with their relative vitalities. Existing collaborations are based on co-authored publications, and potential collaborations are based on finding communities in which each

institution published. Using maps from both 2002 and 2003, we found a total of 11 co-authored papers, distributed on the map as shown in Figure 5 (top). In addition, 74 different research communities were identified as potential collaborations between the two institutions. These communities aggregate into 48 separate disciplines as shown in Figure 5 (bottom). A list of the target communities, key phrases (research topics) associated with each community, primary authors at the two institutions, and the size and vitality of the communities, have been provided to management at both institutions to use in discussions about strategic directions (see example in Table 1). Use of the vitality metric will allow management to focus research and future collaboration in the fastest-moving topics within a discipline.

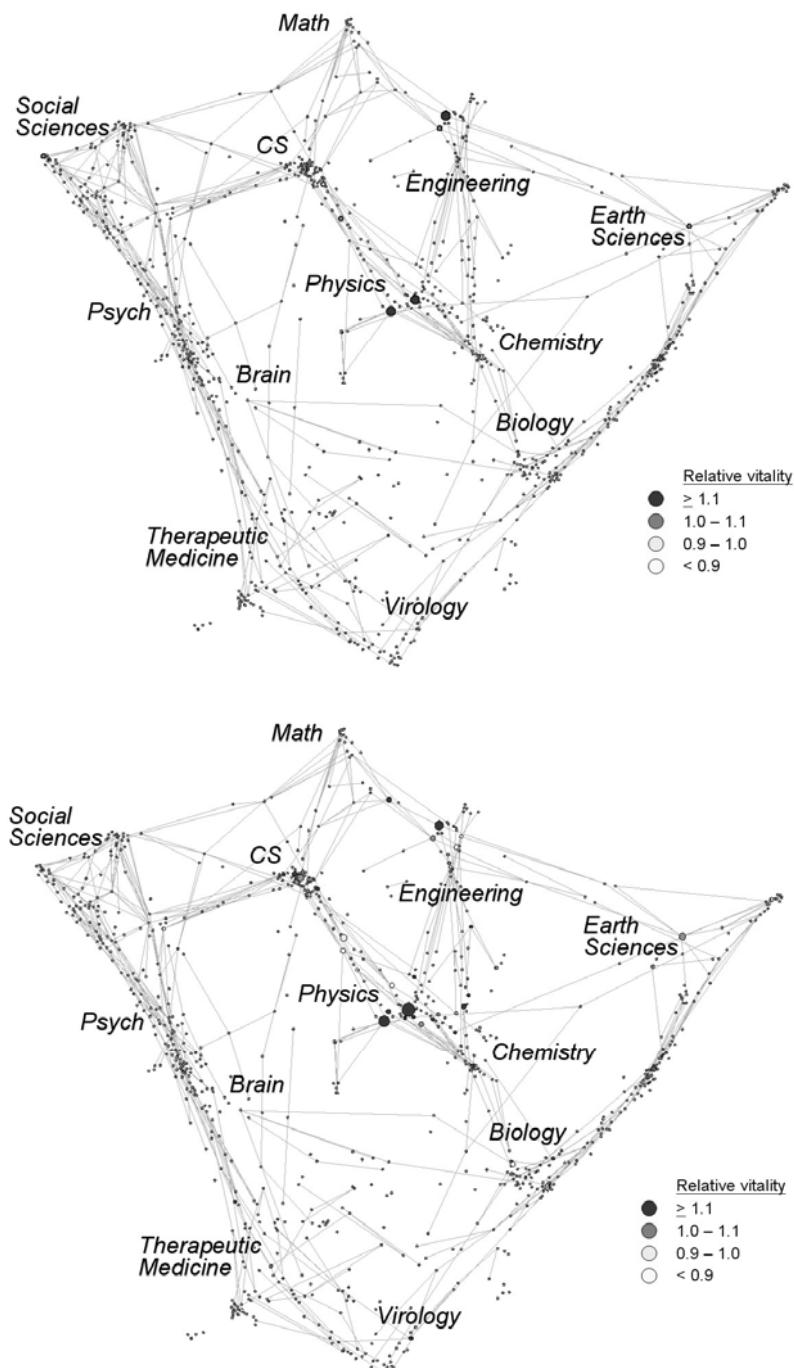


Figure 5. Existing (top) and potential (bottom) collaborations for two institutions, Sandia National Laboratories and the University of Texas system, overlaid on the disciplinary map.

Table 1. Example target communities for future collaboration (author names not listed).

Dsc	Discipline	Community #	Vitality	Ref. Vitality	Key Phrases
649	<i>Physics, fluids and plasmas</i>	2003_C_108168	0.348	0.237	magnetic shear; transport barrier; heating power; density profiles; ELM My H-mode
749	<i>Materials science</i>	2003_C_113040	0.340	0.297	quantum dots; CdSe nanocrystals; core/shell nanocrystals; quantum rods; box nanocrystals
13	<i>Engineering, electrical</i>	2003_C_53745	0.338	0.289	PDMS interconnects; negative pressure; ceramic packaging; native oxide; water plugs
473	<i>Biochemistry & mol. biology</i>	2003_C_57437	0.302	0.256	pre-mRNA splicing; splice site; cell cycle; U6 snRNA; splicing factor

Creating a Potential Collaboration Index

The idea of identifying potential collaboration can be generalized to all institutions. Sandia National Laboratories is working to identify which of the major universities in the U.S. have the greatest potential overlap with our work in a variety of fields. Among the many reasons for this activity are the need to identify strategic academic partners by scientific field, and the desire to hire from institutions doing work of high quality that aligns with our strategic missions. We have thus generated a collaboration index that ranks U.S. universities by the number and vitality of potential collaborations, using the research communities in which we have published as a basis. This index was generated for eight of the primary hire fields for Sandia: computer science and information technology (CS/IT), mechanical engineering (ME), electrical engineering (EE), physics, chemistry, chemical engineering (ChE), materials (MS), mathematics; and one growth field, biology.

The nine fields were identified by grouping journal clusters in the 2003 disciplinary map into nine groups, as shown in Figure 6. The boundaries and groupings on the map were chosen after detailed exploration of the journal clusters, their major journal constituents, and the ISI category assignments for journals and clusters. Large portions of the map are left uncategorized, but are not needed given our institutional work profile.

Potential collaborations were identified based on the communities from three consecutive years: 2002-2004. Although the method section of this paper has only described the calculation of communities for 2003, an identical method was used for the 2002 and 2004 paper-level data, and the resulting communities were assigned to journal clusters using the dominant journal counts, as was described for the 2003 communities.

We designed an index to rank the potential for collaboration that would take into account not only the number of communities (and thus the number of topics) in common with a university, but also the vitality of those communities. However, we did not want the number of communities to completely overwhelm the effect of the fastest moving, high vitality science. Thus we used the product of the cubed root of the number of communities and the average vitality of those communities as our index. Table 2 shows the results of this analysis for the physics field as defined in Figure 6.

The effects of both the number of communities and the vitality can be seen in the listing in Table 2. For example, the top four universities listed all have potential collaborations in over 40 different communities. However, as we proceed down the list the numbers of communities do not decrease monotonically. Some institutions with fewer communities are ranked higher than other institutions with more communities. For example, the University of Wisconsin is ranked higher than either UCLA or the University of Michigan due to its work in higher vitality communities. The same is true for Lehigh University, which has the highest average vitality of those in the table.

Similar calculations were done for each of the other eight fields shown in Figure 6, and the results have been used by management in making strategic decisions as regards our university programs.

Table 2. Potential collaboration index between Sandia National Laboratories and 20 U.S. universities in physics. Sandia published in a total of 417 different physics communities over the time period of the study, 2002-2004.

University	No. communities	Avg. vitality	Index
<i>MIT</i>	44	0.271	0.957
<i>Univ Illinois - Urbana-Champaign</i>	41	0.264	0.910
<i>Princeton Univ</i>	45	0.254	0.903
<i>Univ Calif Berkeley</i>	44	0.255	0.900
<i>Univ Calif San Diego</i>	32	0.282	0.895
<i>Univ Calif Santa Barbara</i>	33	0.279	0.895
<i>Univ Texas - Austin</i>	28	0.285	0.865
<i>Univ Florida</i>	21	0.295	0.814
<i>Princeton Plasma Phys Lab</i>	18	0.305	0.799
<i>Univ Wisconsin - Madison</i>	20	0.291	0.790
<i>Univ Calif Los Angeles</i>	24	0.268	0.773
<i>Univ Michigan</i>	28	0.250	0.759
<i>Columbia Univ</i>	16	0.300	0.756
<i>Cornell Univ - Ithaca</i>	20	0.278	0.755
<i>Univ Maryland - College Park</i>	24	0.255	0.736
<i>Lehigh Univ</i>	10	0.333	0.717
<i>Univ New Mexico</i>	21	0.250	0.690
<i>Arizona State Univ</i>	18	0.258	0.676
<i>Johns Hopkins Univ</i>	15	0.270	0.666
<i>CalTech</i>	28	0.218	0.662

Summary

This paper has presented a method for identifying targets for future collaboration between two institutions, and has shown its utility in two different applications: identifying specific potential collaborations at the author level between two institutions, and generating an index that can be used for strategic planning purposes. Potential collaborations are those where authors belong to the same small paper-level community. Although the examples shown here deal only with institutions within the U.S., the method is equally applicable to the identification of potential international collaborations. Identification of potential collaborations at this detail is only possible because of our ability to map and cluster papers at an extremely refined level. The paper-level map presented here from the combined 2003 SCIE/SSCI/Proceedings databases contained nearly 1 million papers organized into 117,435 communities. The average size of a community is just over 8 papers, thus the potential collaborations identified using this method are extremely focused, at the research topic level.

This method could be expanded to include potential collaborations from neighboring communities. Such an exploration, while it would undoubtedly introduce a higher fraction of so-called false positives into the result set, might also identify more potential areas for interdisciplinary collaboration. Although this is not on our short-term research agenda, we invite research and discussion on better ways to identify high-quality opportunities for collaborative research.

We also note that just because a potential collaboration is identified based on a common topic focus, this does not necessarily mean that the collaboration should occur. Many other factors are typically considered when choosing collaborations, including funding, detailed skill sets, and personal relationships. However, we also believe that far more fruitful collaborations could be occurring than currently are, and endorse the method presented here as a means to that end.

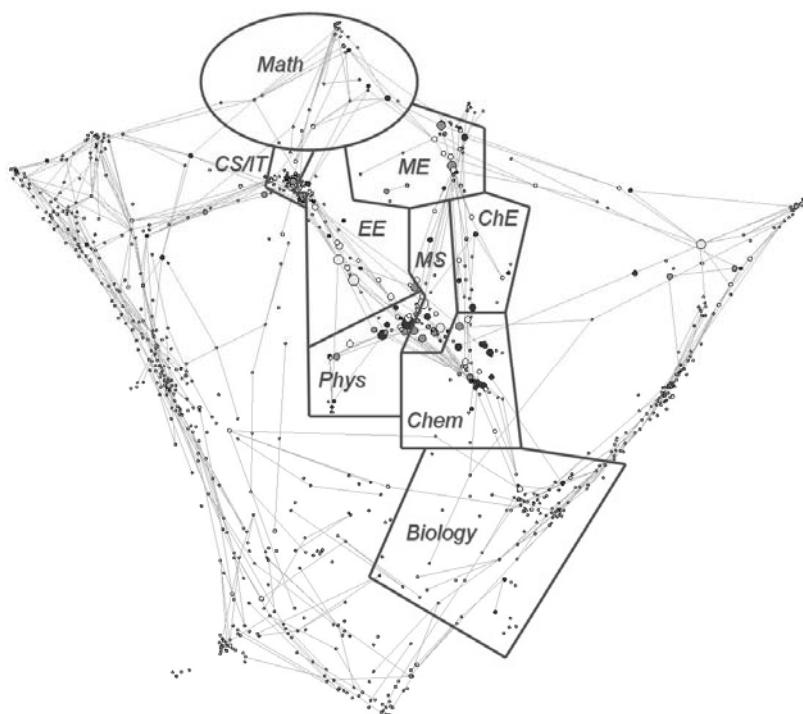


Figure 6. Journal clusters comprising field groupings on the disciplinary map. (The numbers of communities in which Sandia National Laboratories published during 2002-2004 are: Physics (417), EE (283), MS (238), Chemistry (237), ME (185), CS/IT (83), ChE (67), Math (34), and Biology (31)).

References

- Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37, 179-255.
- Boyack, K. W., Börner, K., & Klavans, R. (2007). *Mapping the structure and evolution of chemistry research*. Paper presented at the 11th International Conference of the International Society for Scientometrics and Informetrics.
- Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351-374.
- Glänzel, W., Schlemmer, B., Schubert, A., & Thijs, B. (2006). Proceedings literature as additional data source for bibliometric analysis. *Scientometrics*, 68(3), 457-473.
- Glänzel, W., & Schubert, A. (2001). Double effort = double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, 50(2), 199-214.
- Havemann, F., Heinz, M., & Kretschmer, H. (2006). Collaboration and distances between German immunological institutes - a trend analysis. *Journal of Biomedical Discovery and Collaboration*, 1, 6.
- Klavans, R., & Boyack, K. W. (2006a). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2), 251-263.
- Klavans, R., & Boyack, K. W. (2006b). Quantitative evaluation of large maps of science. *Scientometrics*, 68(3), 475-499.
- Klavans, R., & Boyack, K. W. (submitted). Thought leadership: A new indicator for national and institutional comparison. *Scientometrics*.
- Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36(3), 363-377.
- National_Science_Board. (2006). *Science and Engineering Indicators 2006*. Arlington, VA. National Science Foundation (volume 1, NSB 06-01; volume 2, NSB 06-01A).
- Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the USA*, 98(2), 404-409.

Everything about Genes: some Results on the Dynamics of Genomics Research¹

Robert Braam

r.braam@rathenau.nl

*Rathenau Institute, National Centre for Science System Assessment (SciSA),
Anna van Saksenlaan 51, 2593 HW Den Haag (The Netherlands).*

Abstract

In this study some novel indicators and publication data resources are explored to study the dynamics of genomics research at different levels. The growth of genomics research worldwide seems to be stabilizing, whereas genomics research in the Netherlands is aiming at getting ‘ready for the next step’. Our results suggest that arranging governmental support for this ‘next step’ could mean different things for different Genomics Centers, to better fit steering measures to the underlying research dynamics of these Research Centers. For this purpose, a general model of research dynamics and timing of research management is introduced, building on ideas of Price and Bonacorsi, that may function as a tool for policy makers and research management to discuss timing and matching of interventions in relation to the dynamics of research. As a first exercise, indicator results for genomics research are presented that will be used to inform coming steering discussions at the Netherlands Genomics Initiative (NGI).

Keywords

research dynamics; science policy; search regimes; genomics; indicators.

Introduction: Genomics Research Dynamics explored by using bibliometric indicators

In 2001 the Dutch Government established the National Genomics Initiative (NGI) to promote collaboration in the scattered genomics research activities in the Netherlands. The NGI’s strategic plan for the period 2008 – 2012, aims at adding another € 300 million to the national genomics research infrastructure, to emphasize social and economic returns, and further strengthen the knowledge base in research, technology and education (NGI, 2006). This ‘next step’ includes strengthening subfields as proteomics and bioinformatics (or: e-bioscience) and creating platforms for new subfields as metabolomics, or ecogenomics.

In our research we try to provide (bibliometric) research information that helps enhancing the interplay of research dynamics and research policy. Besides calculating research indicators, we propose a model of research dynamics and timing of research management. This model may serve as a tool for policy makers and research management to follow the dynamics of research over time, and to better discuss options for timing and matching of steering interventions in relation to the internal dynamics of research. Our first main question is how Dutch research fits into the dynamics of Genomics worldwide, and what differences, if any, can be found in the dynamics of the several Centers and Platforms related to the national genomics initiative. Is the genomics research field worldwide expanding, or stabilizing, and in what ways, and what can be said about our national research efforts in this respect? Our second main question is how NGI’s steering measures will best fit the ‘next step’ in the dynamic development of genomics research in the Netherlands.

We look at the presence of genomics research on the internet, in the Web of Science and in a digital library repository, for the period 1990 – 2006, to get a rough picture of the dynamic development of the field of genomics and its subfields worldwide. Next, we look at Dutch research contributions in the research output found in the Web of Science. Next, we looked at output growth, research focus and collaboration in the Research Centers and Platforms, to get an impression of the dynamics of Dutch genomics research as promoted by NGI. Finally we visited the Genomics Momentum by NGI (2006) to get an impression of the NGI atmosphere.

¹ We kindly thank QANU for permission to use the publication data from the self-evaluation reports of the various NGI Centers for the analysis presented in this paper.

The results are set in a framework of science dynamics and management (inspired by Bonacorsi (2005) that looks at science as a creative process of knowledge construction, with alternating periods of diverging and converging research strategies (Fig. 1). In this process, we distinguish two main dimensions. One dimension includes the contents of research fields, with differing richness to be explored that are approached and interpreted by researchers. Bonacorsi (2005a) denotes this dimension as a ‘search space’, or even ‘search regime’: the pattern of existing theories, hypotheses, objects and techniques, that scientists will follow in their search for solutions to particular scientific problems. Research in this dimension is characterized as ‘convergent’ if each conclusion adds to a more general one’, or as ‘divergent’ if each conclusion gives origin to many sub-hypotheses and new (re)search programs’. We add to this, that in the course of time within a research field or program the search process may alter from a divergent to a convergent regime, or vice versa, depending on novel content produced. The way scientists proceed herein; we here call their ‘research strategy’ (Fig. 1).

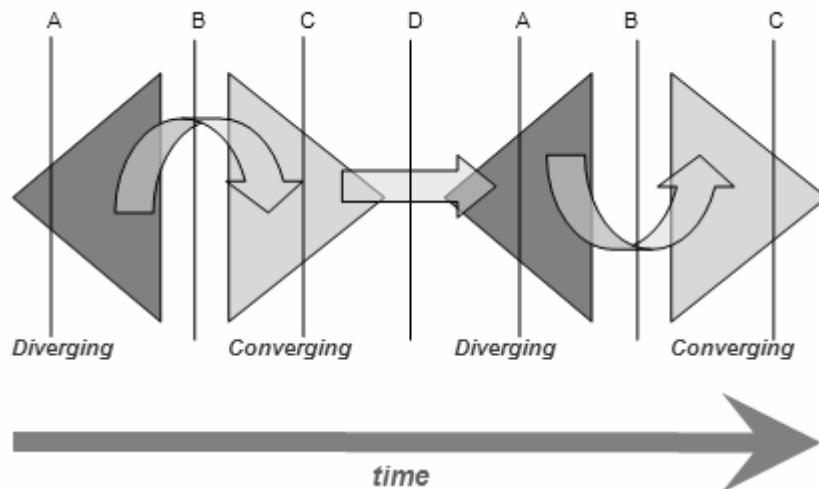


Figure 1. Timing of research management in line with research dynamics

According to Whitley (1984, 2000), the autonomy of researchers over their research strategies is limited by the need to convince specialist colleagues (or peers) of the significance of their work. Though the level of this restriction differs between scientific fields, alternation of diverging and converging search periods may still occur well within the bounds of each field. The other dimension relates to the institutional settings of research rather than to its contents. It includes the prevailing institutional arrangements, organizational settings and strategies that are superimposed on research and researchers by research organizations, by government or business firms that take an interest in steering and harvesting science results. Steering initiatives, and policy measures within this dimension may effect the research strategies of groups, in addition to the prevailing search regime. As to the first dimension, an indication of a search regime may be found in both research input, such as growth and diversity of research personnel and collaborations, and output, such as publication growth and diversity of topics chosen, as reflected in use of keywords (Bonacorsi, 2005a/b), or in the scope of journals groups publish their work in, as we do here.

Looking at the second dimension, a main aspect of concern is how policy steering measures fit the development of the research (Fig. 1 A-D). If the research agenda is expanding, as in divergent periods, the institutional setting and strategic funding measures are probably more effective if they provide critical mass and stimulate openness of project proposals based on excellent quality. On the other hand, if the research changes to a convergent pattern, policy steering instruments are to be used that accommodate to the requirements of gaining focus and stability in funding this research, e.g. stable budgets for clear research targets. Steering measures may, however, also take a more pro-active form, anticipating or forging changes.

Genomics as a growing research field worldwide

The word ‘genomics’ is said to have been invented in 1920 by a German botany professor, as a contraction of the words *gene* and *chromosome*, thus starting Genomics as the study of an organism’s genome, its hereditary information as encoded in the chromosomes on the genes of the DNA, to provide answers in biology, medicine, and industry (In: Wikipedia, Nov. 2006).²

The first genome entirely sequenced in 1980 was that of a single cellular organism. The international Human Genome Project (HUGO) completed a rough first draft of the human genome early 2001. Today, genomics is seen as a potential source of novel applications in medicine, health, food and sustainable production (NGI, 2006). In the wake of the technical successes of genomics, the use of the Greek suffix ‘ome’, meaning ‘all’, ‘every’ or ‘complete’ in ‘gen-ome’ has been translated to other areas, such as ‘nutri-’ or ‘eco-genomics’.

Looking at genomics research worldwide it is clear from the data we gathered that the field has grown very rapidly over the last decade, after its ‘birth’ under the common denominator of ‘genomics’ around the year 1990 (Fig. 2)³.

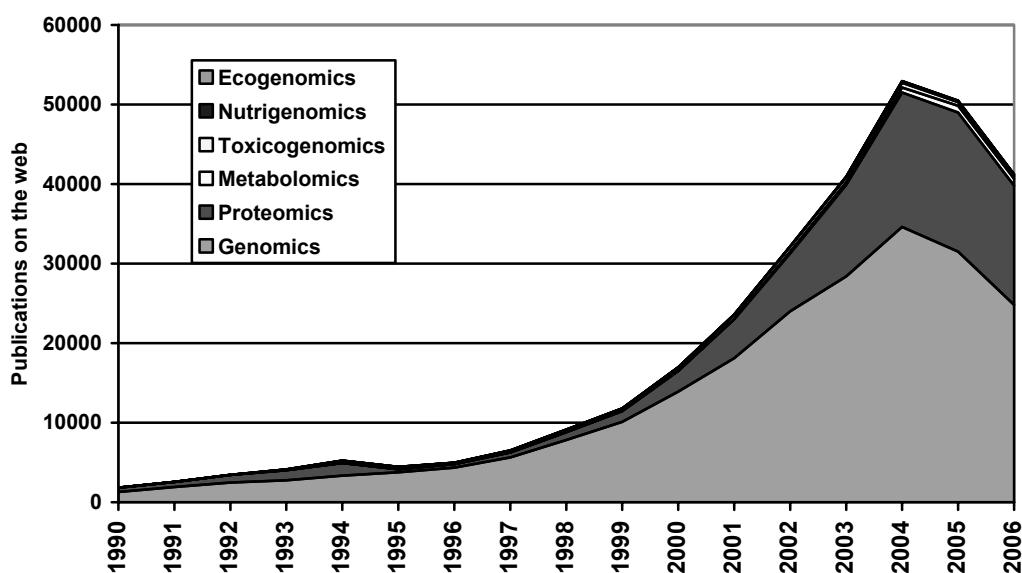


Figure 2. Genomics Worldwide Source: data searched at Google Scholar, 6th February 2007.

The newer sub-areas of ‘ecogenomics’, ‘nutrigenomics’, ‘toxicogenomics’, ‘metabolomics’, form a relatively small portion on top of ‘genomics’ and ‘proteomics’ as seen from publications presented on the worldwide web.

If we zoom in on the top layers of the graph, it is clear that ‘toxicogenomics’ is a rather stable area, at least as perceived on the internet, that is present from early in the nineties, whereas ‘metabolomics’, ‘nutrigenomics’ and ‘ecogenomics’ are the more recent genomics subfields.

It seems clear from these results, that genomics research worldwide has branched into some novel subfields, or: lent its success to fields that also want to have a try in ‘omics’ research. The growth of these subfields, as of yet, however, doesn’t provide a second boost of efforts.⁴

² Or: “Genomics is the area of scientific study that aims to decipher and understand the genetic information contained within living organisms”. http://ghi-igs.nrc-cnrc.gc.ca/whatisgenomics_e.html

³ We assume here that the development of the genomics research field and its sub-fields, is reflected in the use of keywords as ‘genomics’ in a linear fashion: that is, proportionally with total research output.

⁴ According to Derek de Solla Price’s (1961,’63) theory of the exponential growth of science, an area of research stabilizes after a period of exponential growth, due to saturation, and then may start a new exponential growth.

As a rough statement we conclude that genomics research showed dynamic growth in the last decade, starting to grow rapidly in the late nineties of the last century, and is now branching into several subfields, whereas the earlier expanding growth is now turning to a slower pace.

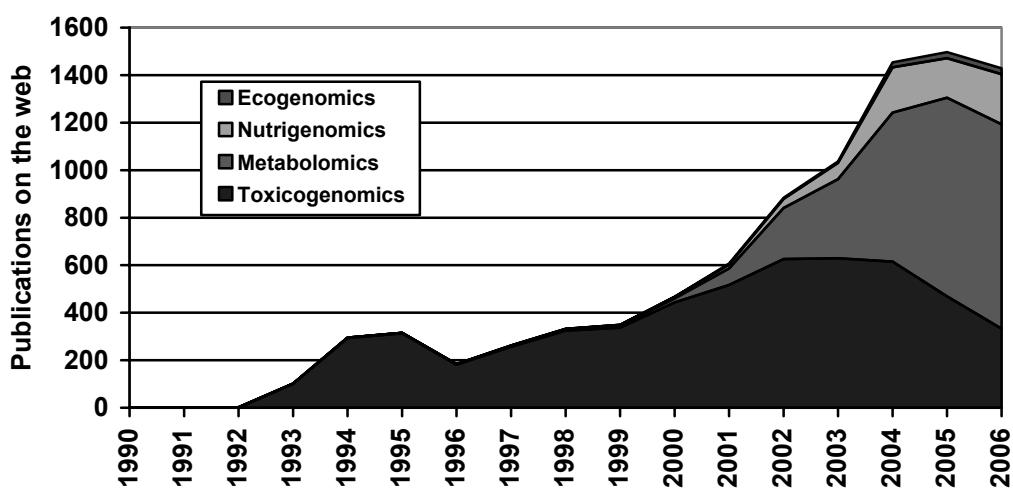


Figure 3. Genomics' minor subfields
(Source: data from Google Scholar, 6th Feb. 2007)

The data were gathered from the internet using the ‘Google Scholar Advanced Search’ (GS) search engine (Google, 2006) for publications including ‘Genomics’, respectively ‘Proteomics’, ‘Metabolomics’, ‘Toxicogenomics’, ‘Nutrigenomics’ or ‘Ecogenomics’; anywhere in the article, for the specified years (<Return articles published between ‘1990 – 1990’, up to ‘2006 – 2006’>⁵). Both in the genomics field as in the larger and minor subfields it seems that the growth curve is stabilizing or even going down somewhat. Anyhow, it seems, from presence of publications on the internet, that the field is now far from exponential growth, indicating lower dynamics.

It must be said that this conclusion is rather tentative, as it is known that the coverage of (formal) research publications by Google Scholar is limited, according to Jacso (2005). On the other hand, citation counts across a wide range of disciplines, including molecular biology and ecology, are shown to lead to essentially the same results, as may possibly be explained by posting of journal back-issues and author-draft versions of articles on the web, as shown by Pauly and Stergiou (2005). We therefore conclude, with some caution, that our data and analysis results on publication growth, seen on the web, indicate real developments in genomics. For an additional perspective we also performed a ‘genomics’ search on the HighWire Press, Stanford University Libraries, which hosts a repository of (links to) full-text articles from nearly 1000 peer reviewed journals. The results are quite similar to what we found at GS, although the downward movement at the end of the time period is not so clear.

In the HighWire Press online hosted journals, the figure of ‘genomics’ is steadily going up starting from the late nineties (Fig. 4), indicating a steady-state growth of the genomics field.

If we look at ‘genomics’ and its subfields⁶ in the Web of Science by Thomson Scientific⁷, the picture indicates the genomics field entering gradually a more steady growth pace too.

⁵ It should be marked that the GS data for 2006 include publications up to the month November only; the figures for 2006 have been corrected for the remaining time-period using a linear estimate.

⁶ Subfields included ‘genomics’, ‘proteomics’, ‘metabolomics’, ‘toxicogenomics’, ‘nutrigenomics’ and ‘ecogenomics’, i.e. the same subfield indicators were taken as in searching at Google Scholar (Fig 2).

⁷ The Thomson *Web of Science* provides access to current and retrospective multidisciplinary information from some 8700 high impact research journals worldwide. Approximately 850,000 fully indexed journal articles have been added to *Web of Science*, from 262 scientific journals.

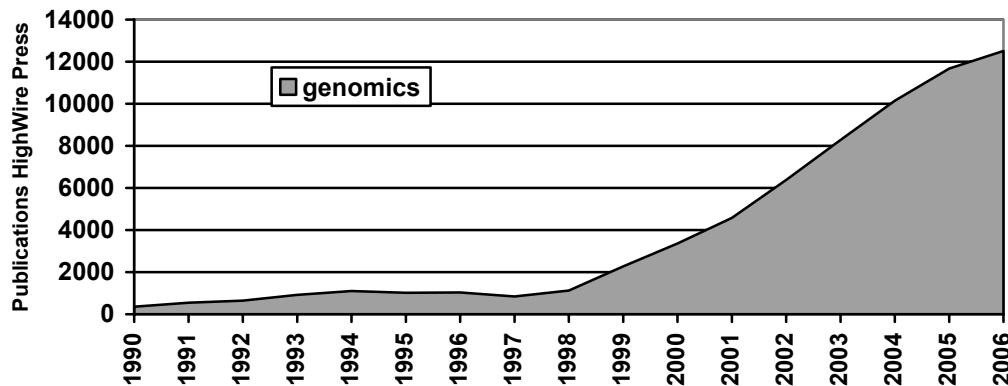


Figure 4. Genomics worldwide, at HighWire Press, Stanford University Libraries.⁸

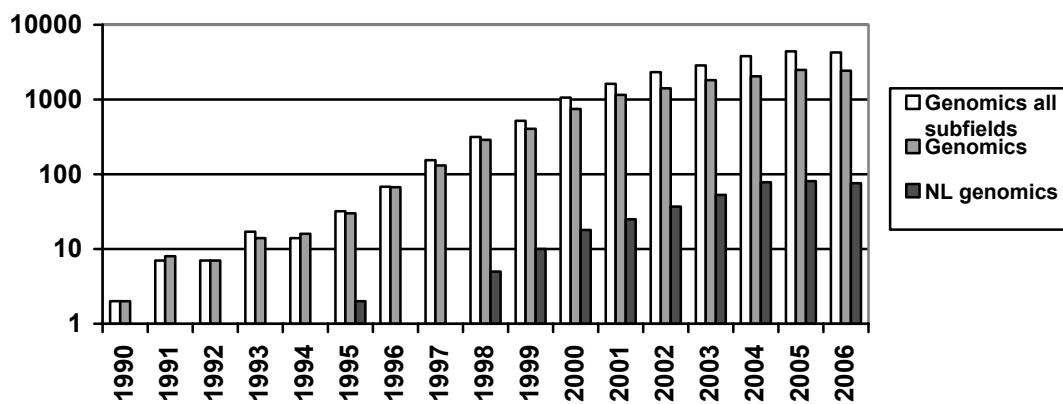


Figure 5. Genomics worldwide at Web of Science, Thomson (see note 4)

(Source: Web of Science, Thomson, 28 November 2006, updated 6th Feb. 2007.)

If we look at the contribution of genomics research in the Netherlands, it seems that the figure more or less follows the same path, though at a lower level, which is quite understandable if we take the size of the country into account. The Dutch contribution to genomics and its several subfields will be discussed in more detail in the next paragraph. For the moment we conclude, based on these rough indications by our data, that genomics and its subfields worldwide is in steady growth, or maybe even declining somewhat in popularity amongst researchers and their financers. However, this might be a period of rest before a new boost of growth, or stagnation due to lack of high throughput data analysis facilities to be developed

The National Genomics Initiative in the Netherlands

In the Netherlands the government has been supporting genomics research from the early years of this century, by establishing a National Genomics Initiative that stimulates novel research collaboration in genomics research networks and innovative application clusters. The contribution to genomics by research from the Netherlands, as seen in the Web of Science, is not evenly distributed amongst the several subfields, as seen in Fig. 6 below. The numbers in this figure give an underestimate of

⁸ Source: High Wire Press Online, 08-11-2006, figure for 2006, updated 14-02-07. High Wire Press of Stanford University Libraries provides on-line repository hosting of some 1000 journals and about 4 million full text articles from over 130 scholarly publishers, including PubMed.

contributions, as only publications explicitly including ‘genomics’, or ‘proteomics’ etc., in the article text were selected for this instance.

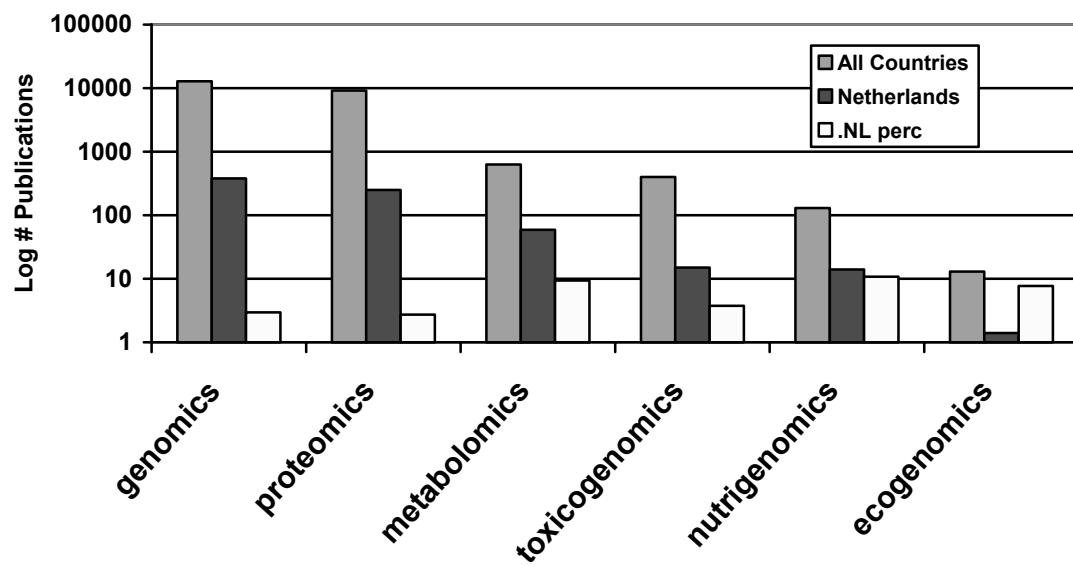


Figure 6. Dutch presence in genomics as reflected in Web of Science 1990-2006
(Source: search results at Web of Science, Thomson, 28 November 2006.)

In order to study the dynamics of Dutch genomics research in more detail we used data from self-evaluation reports of these centers that were provided on behalf of the NGI and QANU.⁹

For the several Centers of Excellence, Platforms and Innovative Clusters, we have performed a somewhat more detailed analysis of academic research output, focusing on levels of growth, topical diversity and collaboration networks. Contrary to the worldwide data, for these centers only a limited number of years that can be traced, as the subsidized research output is not older than 2002, and in some cases even of more recent date. Following Bonaccorsi (2005) we looked at the growth of the number of publications from the different centers, as a measure of expansion of output in their research areas, and at the choice of journal titles chosen to publish in, as a measure of diversity of the scope of research at the centers. Collaboration patterns will also be looked at as a further clue to the dynamics of genomics research. These three data together will be used to establish indicators of research dynamics at the different groups, or centers: do their research efforts follow a more diverging or converging (re)search regime?

So far, we have tracked down publication data and plotted tables for the programs of the genomics centers for growth of academic peer reviewed journal articles and for journal scope. As most of the programs just run for some years as of yet, and as journal scope is only a rough indicator of research scope, the results offer not more than a tentative indication of the search development in the programs and some differences between program dynamics. Not for all the twelve centers publication data were already available, or were not available in a form that could be used for our data gathering and analysis purposes. Therefore, some of the centers are not included here. If possible, figures for these centers will be presented later on.

The research output of the NGI Genomics Centers is given below (Fig. 7) as a first indicator of the dynamics of genomics in the Netherlands.

⁹ The Netherlands Genomics Initiative (NGI) recently commissioned an evaluation of its centres by QANU: Quality Assurance Netherlands Universities. SciSA was kindly permitted to use these data to provide additional information.

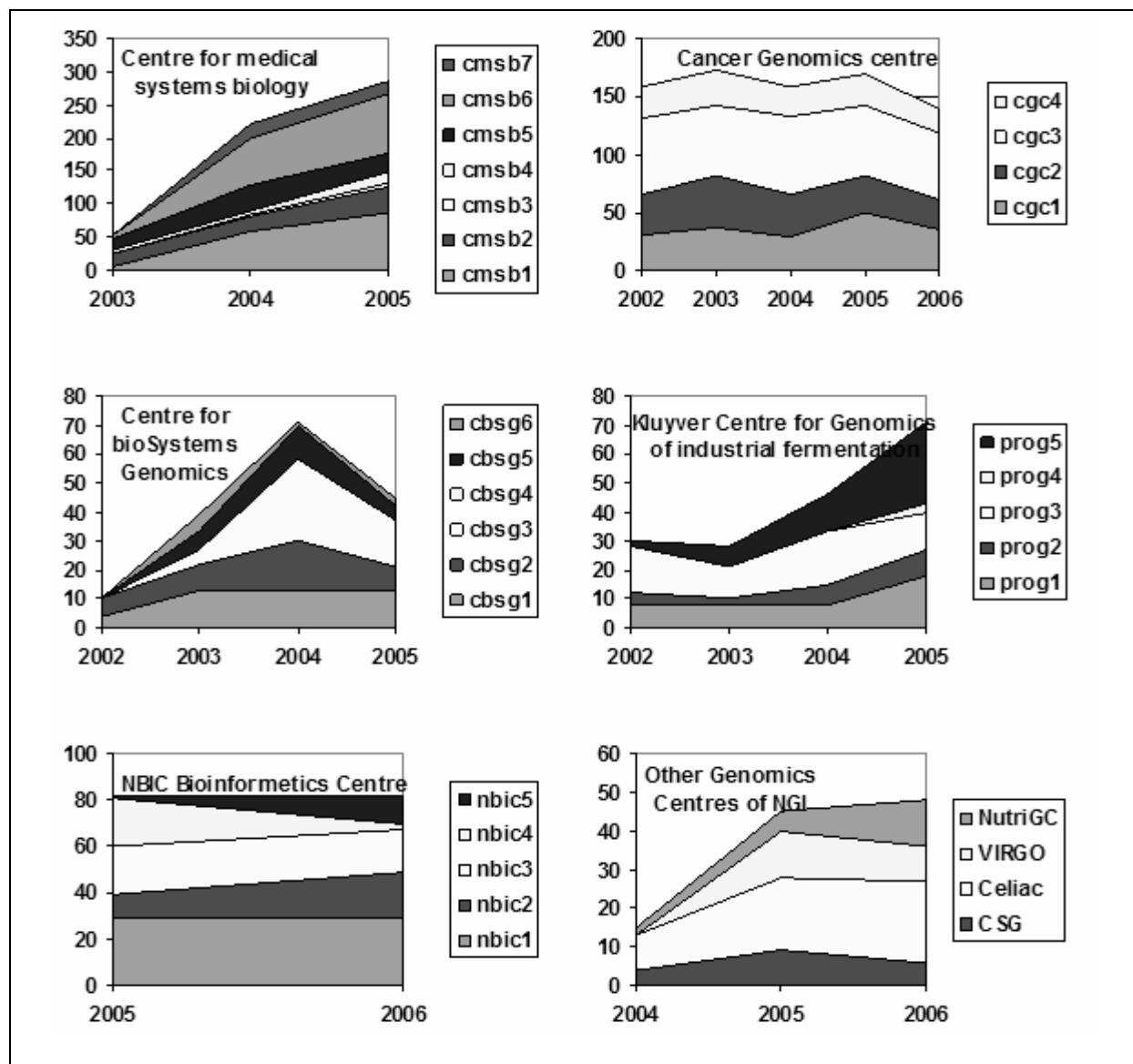


Figure 7. Growth of research output NGI Centers, as numbers of publications¹⁰

It seems clear from these figures that the dynamics at the several centers included are not the same: some programs clearly expand, while others seem to stabilize or decline in output somewhat, perhaps only for the moment. Also, the starting date of the Centers' connection to NGI differs, which is of course reflected in both the output data and results. For some of the centers, in particular 'Ecogenomics', and 'Toxicogenomics', data have so far not been processed or are not available yet, and results are not included here. The establishment of a Centre for Metabolomics is part of the next strategic plan of NGI. It is apparent that the NGI covers with all these centers a very broad range of the worldwide genomics research universe. If we look at the NBIC Centre, output seems to be stabilized. As bioinformatics is a major tool for genomics research, and probably crucial for the future flourishing of genomics, it is good to remark here that bioinformatics is programmed in other NGI Centers as well. The same holds for research efforts in Society and Genomics (CSG), for which a dedicated Centre is established, but also attention is (to be) paid at each of the other NGI Genomics Centers.

¹⁰ Only articles in peer reviewed international journals included (and bioinformatics/computing conferences); figures for 2006 are here corrected by a factor 3/2 (adding a third of a years output) as a rough estimate of 2006 research output. For three of the NGI Genomics Centers, no sufficient data were available as of yet: Ecogenomics Center; Toxicogenomics Center; Proteomics Center.

We now turn to another indicator of research dynamics: change in journal scope (Table 1). The idea behind this indicator is that in an expanding field of research where new topics are explored and/or research efforts incline sharply, the choice of journals will reflect the broadening of scope and/or search of available extra journal space to publish the new amounts of output in. Therefore we looked at the number of articles that were published by the Centers in journals that were (first) used as output channels in the period under consideration (Fig. 8).

Table 1. Example of journal scope change in a research program

Celiac Disease Consortium, journal articles 2004-2006	2004	2005	2006	Total
PLOS Medicine (PLoS=public Library of Science)	x			1
<i>Journal of Autoimmunity</i>	x			1
<i>GUT</i>	xx		x	3
<i>European Journal of Human Genetics</i>	xx		xxx	5
<i>Genes</i>	x			1
<i>Gastroenterology</i>	x	xxx	x	5
<i>American Journal of Gastroenterology</i>	x			1
<i>Journal of Life Sciences</i>		x		1
<i>Nature Genetics</i>		x		1
<i>Human Genetics</i>		x		1
<i>New England Journal of Medicine</i>		x		1
<i>Expert Review of Molecular Diagnostics</i>		x		1
<i>Biotechnology Advances</i>		x		1
<i>Am J Physiol Gastrointestinal Liver Physiology</i>		x x		2
<i>Clin Chem Lab Med (Clinical Chemistry and Laboratory Medicine)</i>		x		1
<i>Best practice & research: clinical gastroenterology</i>		xxx		3
<i>J Pediatr (The Journal of Pediatrics)</i>		x		1
<i>Genes and Immunity</i>		x		1
<i>Immunogenetics</i>		x		1
<i>European Journal of Gastroenterology and Hepatology</i>		xx	xx	4
<i>Tissue Antigens</i>			x	1
<i>Trends in Biotechnology</i>			x	1
<i>Human Immunology</i>			xx	2
<i>American Journal of Human Genetics</i>			x	1
<i>BMC Genomics</i>			x	1
Total articles	9	19	14	42
Journals	7	14	10	25
New journals (not used before)	7	13	5	-
Articles in new journals	7	16	6	-
Scope: % articles in new journals	100	84	43	-

Source: www.celiac-disease-consortium.nl/publications.asp

This example shows the basic journal publication data of the Celiac Disease Consortium used and clarifies how the scope indicator is calculated: for each journal all articles are plotted for subsequent publication years; the number of articles in new journals is given as a percentage. The table shows that 84% of the publications of the Celiac Disease Consortium in the year 2005 (16 out of 19 articles) was in newly used journals. For each Centre that we could gather and analyze

publication data, the results for this indicator of scope change are presented here (Fig. 8) for all individual programs at these NGI Centers¹¹.

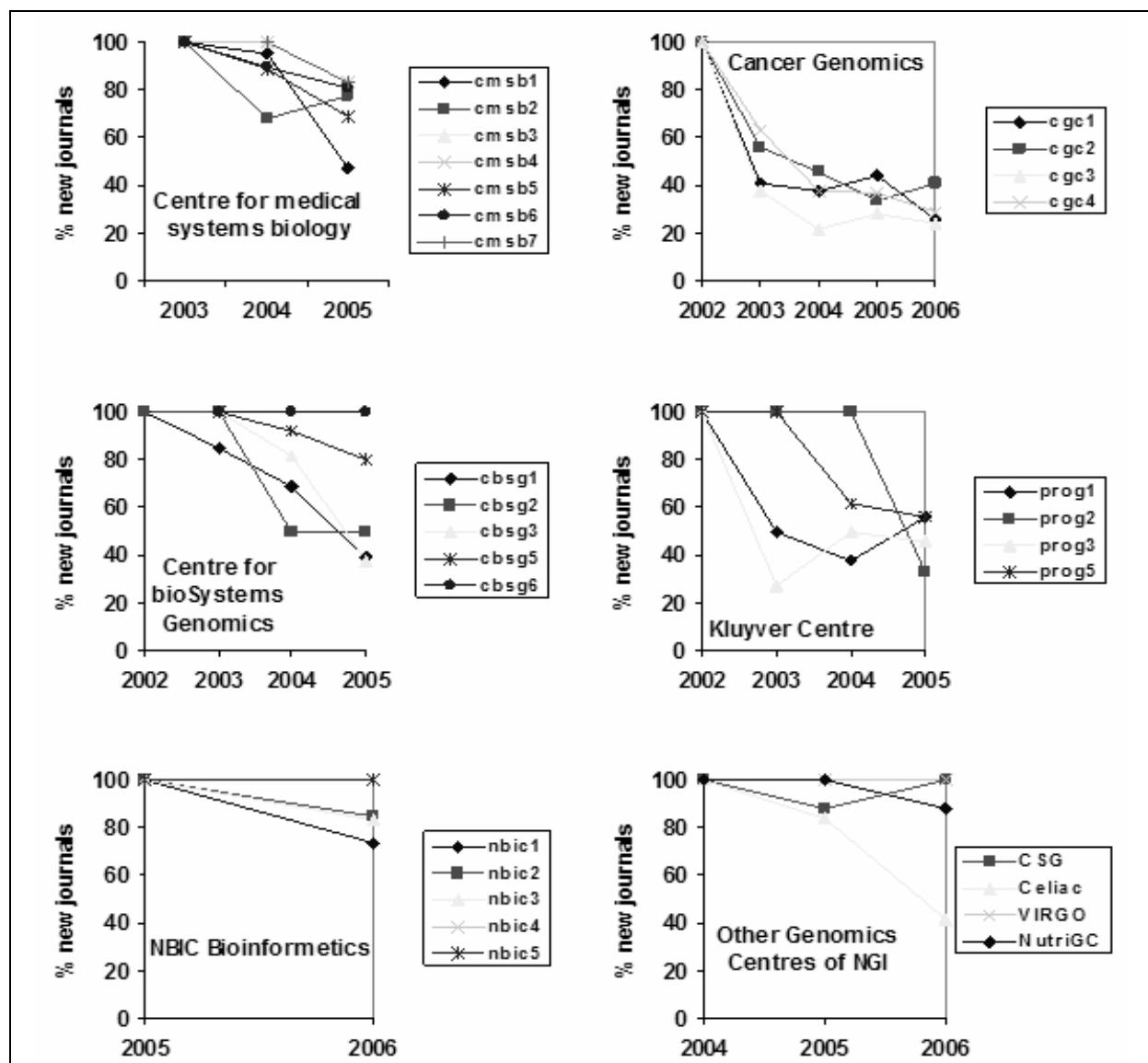


Figure 8. Scope changes of research output of NGI Centers, indicated by journal choice

From these figures it is clear that the dynamics at the Centers' programs show differences. In some cases, as in the Cancer Genomics Centre, the research programs seem quite stable in focus, whereas in other Centers, such as the Kluyver Centre and CBSG, the programs give a more mixed picture, some changing in scope more than others. For most of the programs at the different Centers it seems that the change in journals chosen to publish research output is rather high: many programs have over 40-50% published output in journals not used before for almost every year (in the period under consideration).

The general picture that arises from these results, preliminary as they are, may probably be that alongside some stable research areas in cancer research, biomedical research, industrial fermentation research and in systems biology, a probably expanding exploration of new concepts, tools and

¹¹ Data were gathered at the level of programmes, from the output lists on the web pages of the Centres, and/or - in collaboration with QANU - from the self-evaluation reports the Centres provided for an NGI-research review. Three Centres are left out here, as available publication data were insufficient.

empirical data is going on in the programs, carried out by researchers in the networks associated with and around the NGI Centers.

According to Bonaccorsi (2005), the way researchers work in networks, and the collaboration of many and different types of parties, from public to private, from basic to applied research and industrial innovative efforts provides a third indication of dynamics of research. For this we looked at the number and type of partnerships established in the working networks for the NGI Genomics Centers, as a rough indicator, to get a first impression of research dynamics.

If we look at the NGI Genomics Centers, it shows (from analysis of their fact sheets¹²) that nearly all NGI Centers have many partnerships in their network, including different types of institutes, such as university departments, public financed laboratories, private company laboratories, university medical centers, and some societies and boards related to industry.

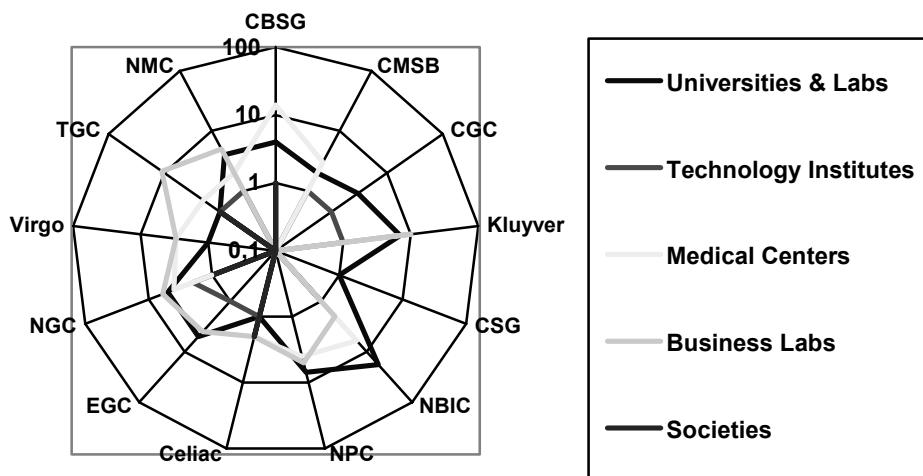


Figure 9a. Partnerships in Netherlands Genomics Initiative Centers

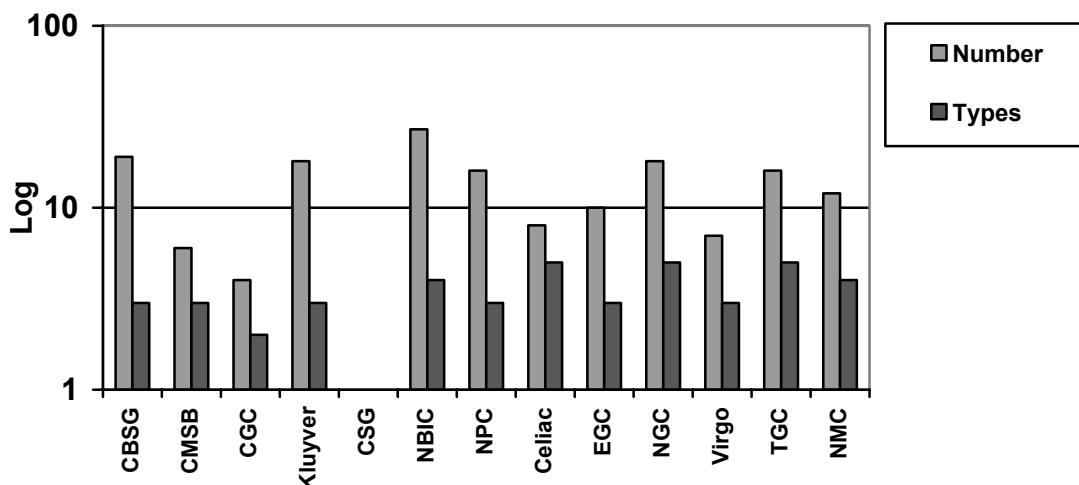


Figure 9b. Partnerships in Netherlands Genomics Initiative Centers

¹² All NGI Genomics Centres provide a fact sheet including data on partnerships, that is available via the NGI website www.genomics.nl. These data are used for analysis in this article.

According to Bonaccorsi (2005a) the variety and abundance of (research)collaboration with non-academic laboratories and business research laboratories and industry, is indicative of research areas with high dynamics and innovative potential. The number and variety of partnerships in the NGI Centers (Fig. 9a) can be taken to reflect something of the dynamics of the research. The NGI Centers include from 4 to 27 institutional partnerships, with up to five different types (Fig. 9b). The only exception is the Centre for Society and Genomics (CSG) that, doing research of a different nature, doesn't work in such an institutional network. Thus, we tentatively conclude from this indicator, as a rough estimate, that the Centers are set in an environment that enhances dynamic expansion of their findings and innovative opportunities.

The future of Genomics research: “Ready for the next step”

A sense of the dynamics of Genomics also comes up from visiting the Genomics Momentum 2006 (NGI, 2006), where the various NGI Genomics Centers presented themselves together with government and business parties, in some 50 exhibit stands. The Genomics Momentum 2006, sponsored by four large industrial firms a government agency for innovation and the City of Rotterdam, explicitly addresses NGI's future expectations under the title given: “Genomics: Ready for the next step”. The several workshops provide an idea of what is meant by this follow up of “The ‘Big Bang of ‘-omics’ that basically just happened”:

Table 2. NGI Genomics Momentum 2006 workshops

Workshop	Titles
<i>Workshop1</i>	Metabolomics and quality of life: towards new initiatives and strategies
<i>Workshop2</i>	Toxicogenomics: assuring safety without animal testing
<i>Workshop3</i>	Towards a bio-based economy
<i>Workshop4</i>	Biobanks: from individual to collective concern
<i>Workshop5</i>	Living Healthier for longer: unexplored opportunities
<i>Workshop6</i>	E-bioscience: a new way of life (science)

The Genomics Momentum 2006 shows that the Netherlands Genomics Initiative is aiming at a broad strategy for research on the one hand, and holds high expectations on the societal and economic revenues to be explored and harvested on the other.

Looking at the three calculated indicators of research dynamics for the NGI Centers, growth of publications, research focus and collaborations, we placed the centers in the timeframe of the above introduced research dynamics and management model (Fig. 1 A-D). The estimates for the Centers are presented in Table 3, together with citation impact results from CWTS¹³.

The three indicators for a number of centers seem to point differences in research dynamics between the NGI Centers, some stabilizing (B), or converging (C), others to more divergent research dynamics (A), particularly centers in the non-classical subfields. For those centers that we could find citation impact results for, it seems these are very well cited internationally. Apart from CMSB, these are more stabilized, or converging Centers. Most of the other centers seem to be in an expanding dynamics phase of the model timeframe. As of yet there are no comparative citation impact data available for these centers. As in the more classical Genomics fields the NGI Centers are very influential up to 2004, and as the other centers seem to be in an expanding research dynamics mode, the NGI statement that the Netherlands are making ready for the ‘next step’ in Genomics may not be far from the truth.

In the Netherlands funding is often discussed along lines of providing ‘focus’ and/or ‘mass’ to research institutes and initiatives. If we take these two dimensions of policy measures and compare them to research dynamics, we may draw a matrix of research policy effects. For each combination of

¹³ Van Leeuwen and Nederhof (2006), Centre for Science and Technology Studies, Leiden University.

research dynamics and policy measures in the matrix, policy steering options are specified that relate to the timeframe of the research management model (Fig. 10).

Table 3. Indicators of research dynamics for NGI Centres

NGI Centre	Growth	Focus	Collaboration	Phase in Model	Citation impact*
<i>CMSB</i>	<	<	<	A	1,33 / 2,74
<i>CGC</i>	=>	>	>	C	1,28 / 1,72
<i>CBSG</i>	>	<=	<	C	1,23 / 1,87
<i>Kluyver</i>	<	>	<	B	1,30 / 1,56
<i>NBIC</i>	=	<	<	A	-
<i>NutriGC</i>	<	<	<	A	-
<i>Virgo</i>	>	<	<	A	-
<i>Celiac</i>	<	>	<	A	-
<i>CSG</i>	=	<	>	A	-

Symbols: < expansion; divergent; complementary; > shrinkage; convergent; redundant; = stable; stable; neutral // *Citation impact compared to average of the centers' journal set / field average (1.0), data from CWTS, 2006.

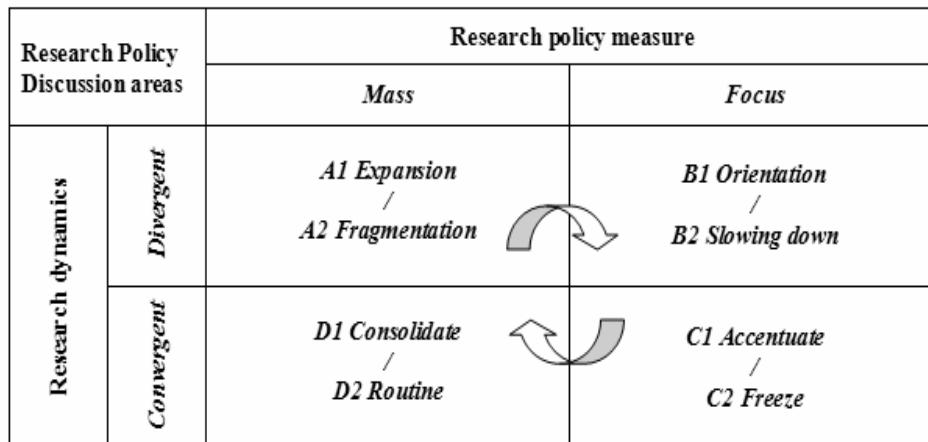


Figure 10. Research policy discussion areas

Providing 'mass' to research in a divergent regime, helps in expanding the research agenda, but it also may lead to fragmentation of research if the area is 'exploding'. Overcoming fragmentation of research efforts is precisely what the NGI tried to reach in the previous funding round. Providing more focus to research in a divergent regime (B) may then be a better policy solution, although one should be careful not to slow down research too much. Some of the NGI Centers (CCG and CBSG) seem to be more in a converging mode, and focused support (C) may be of helping research efforts zooming in here, but one should be on guard of too much rigor here (both Centers here are highly influential internationally).

Conclusions

The growth of genomics research worldwide has been found to be stabilizing, and coming in a period where new branches or subfield are coming up. Genomics research in the Netherlands, associated with the National Genomics Initiative, seems to be going along with this, but also prepares to enter a new

phase of expanding both research and innovative explorations and applications in various sectors. As put by NGI at the Genomics Momentum 2006, the Genomics Movement is getting ‘Ready for the next step’. If Dutch Genomics research is to be at the forefront of this movement, as seems to be the aim, the question arises how to arrange both governmental and industrial support for this in a stimulating way. Moreover, stimulating support may well mean different things for different centers, as underlying dynamics differ. If a new boost of genomics indeed will take place in the next decade, the question will also be if support on the whole range of all these promising subfields will ask for intensified research and technology collaborations and shared research efforts. Before discussing the amounts of money involved, these more qualitative questions will have to be answered in order for the NGI strategic plans to be successful in putting the next step in Genomics to work as it should.

We think, the indicators and publication resources explored provide a rough but informative picture of the research dynamics studied: the international field of genomics and its sub-disciplines, and the Dutch part therein played by the NGI-related Genomics Centers and their individual programs. The framework developed for policy makers and research management to follow the dynamics of research, and discuss timing and matching of interventions, offers a way to differentiate between group and discuss steering options in relation to research dynamics. We hope the results and conclusions in this study will be of help in doing just that.

References

- Bonacorsi, A. (2005a). Search regimes and the industrial dynamics of science. *Paper presented to the PRIME General Assembly, Manchester, January 7-9, 2005*. Draft version (pp. 38), December 2004.
- Bonacorsi, A. (2005b). Better Institutions vs better policies in European Science. *Paper presented at the PRIME General Assembly, Manchester, January 7-9, 2005*. Draft version (pp. 40), January 2005.
- Google (2006). Advanced Scholar Search. Retrieve November, 2006 from: <http://www.google.com>
- Jacso, P. (2005). As we may search – Comparison of major features of the web of science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89 (9) 1537-1547, 10 Nov. 2005.
- NGI (2006). Strategic Plan 2008 – 2012. National Genomics Initiative, official website: www.genomics.nl .Also: Genomics Momentum 2006. Retrieved November 2006 from: www.gm2006.org
- Pauly, D., and Stergiou, K. (2005). Equivalence of results from two citation analyses: Thomson ISI’s Citation Index and Google’s Scholar service. *Ethics in Science and Environmental Politics, ESEP*, Dec.22, 2005:33-35.
- Solla Price, D. de (1961). *Science since Babylon*, Paperbound edition 1962. New Haven and London: Yale University Press.
- Solla Price, D. de (1963). *Little Science, Big Science*, Paperback edition 1965. New York and London: Columbia University Press.
- Whitley, R. (2000). *The Intellectual and Social Organization of the Sciences, second edition*. Oxford: Oxford University Press. Enlarged second edition; reprinted 2006; first edition 1984.
- Wikipedia (2006). Genomics. Entry retrieved November 2006 from: <http://en.wikipedia.org/wiki/Genomics>.

Functional Use of Frequently and Infrequently Cited Articles in Citing Publications. A Content Analysis of Citations to Articles with Low and High Citation Counts

Lutz Bornmann* and Hans-Dieter Daniel**

*bornmann@gess.ethz.ch

** ETH Zurich, Professorship for Social Psychology and Research on Higher Education (Switzerland)

**University of Zurich, Evaluation Office (Switzerland)

Abstract

Using publication and citation data from a study on the selection procedure of the Boehringer Ingelheim Fonds (B.I.F.), this study investigated the extent to which frequently and infrequently cited articles were used differently by the scientists that cited them. The data set consisted of 31 articles by B.I.F. grant applicants that had received 451 citations in 270 citing publications. In a comprehensive content analysis each reference to the B.I.F. article in the citing publication was classified according to two categories: 1) the location of the citation within the citing publication (section of the paper in which the citation appears) and 2) meaningful or cursory mentioning of the article in the citing publication. The results showed statistically significant differences between the B.I.F. applicants' articles with low or high citation counts. All in all, the results indicate that an article with high citation counts had greater relevance for the citing author than an article with low citation counts.

Keywords

citation; citation behavior; citation content analysis; citation context analysis; evaluative bibliometrics

Introduction

The central problem in the use of citation counts to evaluate scientific work is that it is not certain what is being measured by the citations (Bornmann & Daniel, accepted for publication). Are frequently cited articles used by a citing author differently than articles that are infrequently cited? For citing authors, does a frequently cited article have greater relevance – in terms of ‘intellectual influence’ and ‘contribution to scholarly progress’ (Moed, 2005, p. 221) – than an infrequently cited article? According to the social constructivist sociology of science (Latour & Woolgar, 1979) the significance of an article depends largely on the manner in which it is used by other scientists. If scientists intensively use the content of an article, knowledge claims that are made in this article become scientific facts and are gradually integrated into the stock of scientific knowledge (Amsterdamska & Leydesdorff, 1989).

The present study investigated to what extent frequently and infrequently cited articles were used differently by the scientists that cited them. In a comprehensive content analysis we classified citations to cited articles in citing publications using two different categorizations of citations. Firstly, we noted the location of citations with respect to one of the sections of the citing publication: introduction, methods, results, and discussion. According to Voos and Dagaev (1976) there are obvious indications that it is possible to calculate the value of a cited article for the author of the citing publication using its location in the citing publication. In a citation content analysis, Maricic, Spaventi, Pavicic, and Pifat-Mrzljak (1998) attach the highest importance to citations in the methods or results section of a cited publication. Citations in the discussion section are rated somewhat lower, and citations in the introduction section are ascribed the lowest importance. For Cano (1989), citations located in introductory sections represent a “setting of the stage” (p. 288) and have little informational utility to the authors of the citing publications.

Secondly, we classified citations according to intensity of mentioning of the cited article by the citing authors. We followed Bonzi (1982) and chose a simple three-level distinction that captured both cursory mentioning and more meaningful mentioning of the cited article (Hooten, 1991; Maricic, Spaventi, Pavicic, & Pifat-Mrzljak, 1998). Other schemes that was used for citation content analyses

also include cursory or meaningful mentioning of cited articles in the citation categories (on this, see Maricic, Spaventi, Pavicic, & Pifat-Mrzljak, 1998): cursory citation is called perfunctory (Murugesan & Moravcsik, 1978), peripheral (McCain & Turner, 1989), or non-essential (Cano, 1989), and meaningful citation is called organic (Murugesan & Moravcsik, 1978), central (McCain & Turner, 1989), or essential (Cano, 1989).

Methods

The sample of articles cited in the publications analyzed

We previously investigated committee peer review for awarding long-term fellowships to young researchers as practiced by the Boehringer Ingelheim Fonds (B.I.F.) – a foundation for the promotion of basic research in biomedicine (Bornmann & Daniel, 2006). Assessing the validity of the B.I.F. selection decisions, bibliometric analyses for articles published previous to the post-doctoral applicants' approval or rejection for a B.I.F. fellowship were conducted. All in all, 1,586 articles had been published by 397 applicants previous to their applications to the B.I.F. (on average four articles). Using the same data set of articles used to evaluate the B.I.F. selection procedure (B.I.F. applicants' articles and their citing publications), the present study examined to what extent frequently and infrequently cited papers were differently used by scientists who cited them. As content analysis of citations with different classifications is very time-consuming (it entails finding the citation in the article, reading the whole sentence, incorporating the reference, and classifying the citation five times) we did not include all B.I.F. applicant articles (and their citations) in the analysis but instead draw a stratified random sample from the total data set of articles, selecting a separate random sample from each of two strata. The stratification variable was the decision by the B.I.F. Board of Trustees to approve or reject an applicant for a post-doctoral fellowship, as it can be assumed that the articles published by approved applicants were of higher quality than the articles published by rejected applicants. In total, 34 articles written by B.I.F. applicants between 1987 and 1994 were selected randomly: 17 articles each by approved and rejected applicants.

The different strata of a sample selected according to a stratification variable should be relatively homogeneous, and the strata should also be mutually exclusive. For this reason, when selecting articles by the B.I.F. applicants within the two strata, we made sure that the distribution of (1) the publication years and (2) the articles' citation counts were nearly the same and that (3) the articles were published in journals of similar scope and (4) having a similar Journal Citation Report (JCR) impact factor (provided by Thomson Scientific, Philadelphia, PA, USA).

The citations to the articles published by the B.I.F. applicants

The 34 articles of the B.I.F. applicants in our sample were cited by 308 citing publications, with an average of 11 citing publications per cited article (median). The sample of the citing publications was adjusted by excluding those that listed the B.I.F. applicants' articles only in a bibliography without mention in the text ($n=2$) and those that were published in non-English language journals ($n=5$).

In order to test the extent to which the number of citations to articles by the B.I.F. fellowship applicants correspond with the categories of both citation classifications, we divided the total of 34 articles into two groups by using the citations' median value as threshold (see Preacher, Rucker, MacCallum, & Nicewander, 2005): 1) articles with low citation counts ($n=24$), that is, articles with fewer than 11 citations (3 to 10 citations), and 2) articles with high citation counts ($n=7$), that is, articles cited 12 to 23 times. Three articles with citation counts equal to the median value were not included into the statistical analyses.

The sample for the statistical analyses consisted of 270 citing publications. As some articles of the B.I.F. applicants were cited multiple times in one citing publication, the total number of citations was 451. On average, one article by the B.I.F. applicants was cited 1.7 times in one of the 270 citing publications.

Statistical methods

The associations between the categorical variables low or high citation counts for the articles by the B.I.F. applicants and the categories of the categorizations were calculated using the Cochran Mantel-Haenszel test (Agresti, 2002, section 7.5.3-7.5.6; Cytel Software Corporation, 2005). Since the result of the statistical significance test is dependent on sample size and “statistical significance does not mean real life importance” (Conroy, 2002, p. 290), it is the strength of the association that is more interesting and important for interpreting the empirical finding. For calculating strength we have to employ an additional measure of association, here *Cramer's V* coefficient (Cramér, 1980).

Results

The location of the citations to the articles by the B.I.F. applicants with respect to one of the sections of the citing publicationTable 1Table shows the sections in the citing publications where the B.I.F. applicant articles are cited: a total of 32% of the articles are cited in the introduction, 24% in the methods, 13% in the results, and 31% in the discussion section. This result agrees approximately with citation distributions reported by Voos and Dagaev (1976) and Cano (1989). Their findings indicate that the largest concentration of citations is located in the beginning sections of the citing publications. A look at the differences in the percentages of citations in the different sections of the citing publications between articles by B.I.F. applicants with low or high citation counts in Table shows expected differences between the methods, results, and discussion sections. As expected, articles with high citation counts are more frequently cited in the methods (27% of the citing publications) and results (15% of the citing publications) sections than articles with low citation counts (methods: 20% of the citing publications; results: 11% of the citing publications). Articles with low citation counts (39% of the citing publications) are more frequently cited in the discussion section than articles with high citation counts (25% of the citing publications). But contrary to our expectations, articles with high citation counts are more frequently cited in the introduction section (34% of the citing publications) than articles with low citation counts (30% of the citing publications).

The differences in the distribution of the citations in sections of the citing publications between articles by the B.I.F. applicants with low or high citation counts are statistically significant; $T (n=350) = 8.82, p=.03$; with small effect size, *Cramer's V*=.17. (see Table 1).

Table 1. Sections in the citing publications containing the citations to the B.I.F. applicants' articles with low or high citation counts

*Section of citing	**Art. low citation	***Art. high citation	Total
<i>Introduction</i>	30	34	32
<i>Methods</i>	20	27	24
<i>Results</i>	11	15	13
<i>Discussion</i>	39	25	31
<i>Total</i>	100	100	100
<i>No of classified citations</i>	162	188	350

- Notes1. * Section of citing publication where article cited ; ** Articles with low citation counts (3 to 10 citations) ; *** Articles with high citation counts (12 to 23 citations)

- Notes2. $T (n=350) = 8.82, p=.03$ (Cochran Mantel-Haenszel test adjusted for potential effects of qualitative differences between articles of the approved and rejected B.I.F. applicants); *Cramer's V*=.17. A total of only 350 citations could be assigned to a section, because 101 citations to the B.I.F. applicants' articles were located in citing publications that had no (clear) section headings.

Cursory or meaningful mentioning of the B.I.F. applicants' articles in the citing publications. The citation content categories provided by Bonzi (1982) are based on the premise that one measure of true relevance to a citing publication is the extent of treatment of the cited article in the citing publication. An article simply mentioned in a citation can be expected to be less relevant for the author than a citation where the cited article is discussed in depth within the citing publication. For our citation content analysis we used three categories provided by Bonzi (1982) to measure citation relevance: (1) not specifically mentioned in text (e.g., "Several studies have dealt with ..."), (2) barely mentioned in text (e.g., "Smith has studied the impact of ..."), and (3) one quotation or discussion of one point in text (e.g., "Smith found that ...").

For this type of content analyses it is customary for two persons to conduct the coding of text material for purposes of determining the interjudgmental reliability of the codings, using measures of agreement. In the present study two independent coders classified the citations as to cursory or meaningful mentioning of the cited article in the citing publications. The reliability of the two coders' ratings was very high, kappa coefficient = .93 (on interpreting the coefficient, see von Eye & Mun, 2005, pp. 5-6).

Table 2. Cursory or meaningful mentioning of the B.I.F. applicants' articles with low or high citation counts

Citation content category	**Art. low citation	***Art. high citation	Total
(1) Not specifically mentioned in text (e.g., "Several studies have dealt with ...")	31	13	22
(2) Barely mentioned in text (e.g., "Smith has studied the impact of ...")	32	48	40
(3) One quotation or discussion of one point in text (e.g., "Smith found that ...")	37	39	38
<i>Total</i>	100	100	100
<i>No Number of classified citations</i>	226	223	449

- Notes 1. * Section of citing publication where article cited ; ** Articles with low citation counts (3 to 10 citations) ; *** Articles with high citation counts (12 to 23 citations)

- Notes 2. $T(n=449) = 22.84$, $p=.00$ (Cochran Mantel-Haenszel test adjusted for potential effects of qualitative differences between articles of the approved and rejected B.I.F. applicants); *Cramer's V*=.22

The distribution of the citations to the article of the B.I.F. applicants across the three citation content categories in Table 2 shows that the greatest percentage of articles (40%) are barely mentioned in the citing publications (second citation content category: e.g., "Smith has studied the impact of ..."). In another 38% of the citations, either a passage from a B.I.F. applicant's article is cited directly or the content of the article is discussed (third citation content category: e.g., "Smith found that ..."). Twenty-two percent of the citations to the articles are simple mentions, with no discussion of the content of the cited article (first citation content category: e.g., "several studies have dealt with ...").

The results in Table show, as expected, that B.I.F. applicants' articles with low citation counts are clearly used more frequently (31% of the citing publications) by the citing authors for cursory mentioning (first citation content category) than articles with high citation counts (13% of the citing publications). B.I.F. applicants' articles with high citation counts are more frequently barely mentioned in the citing publication (48% of the citing publications) and quoted directly or discussed in the citing publication (39% of the citing publications) than articles with low citation counts (second citation category: 32% of the citing publications; third citation category: 37% of the citing publications).

The differences between the frequencies are statistically significant; $T(n=449) = 22.84$, $p=.00$; the association between both variables has a medium effect size; *Cramer's V*=.22. These findings suggest

that when infrequently cited B.I.F. applicants' articles were used, they tended to have lower relevance than frequently cited articles for the authors of the citing publication (and vice versa).

Discussion

Using publication and citation data from the fellow selection procedure of the B.I.F., the present study investigated to what extent frequently and infrequently cited articles were differently used by the scientists that cited them. We utilized two different categorizations to capture the functional use of the articles by the authors in the citing publications: 1) the location of the citation to the article within the citing publication (section), and 2) meaningful or cursory mentioning of the article in the citing publication.

Our results show that for both classifications of citations in the citing publications there are statistically significant differences between B.I.F. applicants' articles with low or high citation counts. B.I.F. applicants' articles with high citation counts were more frequently cited within the citing publications in the methods and results sections than articles with low citation counts. Articles with high citation counts were more frequently cited in meaningful mentions in the citing publications than articles with low citation counts. We proved whether these associations are still hold when the threshold for the categorizations of the citations is changed. Using three groups (low, medium, and high citation counts) instead of two in the statistical analyses, we got nearly the same results. All in all, our findings suggest that the more an article is cited the more intensively its content is used by the citing scientists. Therefore, citation counts are not only an indication of the (superficial) relevance of research but are also an indicator for the relevance of this research for scientific work in a research field.

References

- Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Amsterdamska, O., & Leydesdorff, L. (1989). Citations: indicators of significance? *Scientometrics*, 15(5-6), 449-471.
- Bonzi, S. (1982). Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science*, 33(4), 208-216.
- Bornmann, L., & Daniel, H.-D. (2006). Selecting scientific excellence through committee peer review – a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, 68(3), 427-440.
- Bornmann, L., & Daniel, H.-D. (accepted for publication). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*.
- Cano, V. (1989). Citation behavior: classification, utility, and location. *Journal of the American Society for Information Science*, 40(4), 284-290.
- Conroy, R. M. (2002). Choosing an appropriate real-life measure of effect size: the case of a continuous predictor and a binary outcome. *The Stata Journal*, 2(3), 290-295.
- Cramér, H. (1980). *Mathematical methods of statistics*. Princeton, NJ, USA: Princeton University Press.
- Cytel Software Corporation. (2005). *StatXact: version 7*. Cambridge, MA, USA: Cytel Software Corporation.
- Hooten, P. A. (1991). Frequency and functional use of cited documents in information science. *Journal of the American Society for Information Science*, 42(6), 397-404.
- Latour, B., & Woolgar, S. (1979). *Laboratory life: the social construction of scientific facts*. London, UK: Sage.
- Maricic, S., Spaventi, J., Pavicic, L., & Pifat-Mrzeljak, G. (1998). Citation context versus the frequency counts of citation histories. *Journal of the American Society for Information Science*, 49(6), 530-540.
- McCain, K. W., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics*, 17(1-2), 127-163.
- Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht, The Netherlands: Springer.
- Murugesan, P., & Moravcsik, M. J. (1978). Variation of nature of citation measures with journals and scientific specialties. *Journal of the American Society for Information Science*, 29(3), 141-147.
- Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: a critical reexamination and new recommendations. *Psychological Methods*, 10(2), 178-192.
- von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement. Manifest variable methods*. Mahwah, NJ, USA: Lawrence Erlbaum Associates.
- Voos, H., & Dagaev, K. S. (1976). Are all citations equal? Or, did we *op. cit.* your *idem*? *Journal of Academic Librarianship*, 1, 19-21.

Spiritualised Medicine? A Bibliometric Study of Complementary and Alternative Medicine

Jenny-Ann Brodin Danell and Rickard Danell

jenny-ann.brodin@soc.umu.se, rickard.danell@soc.umu.se
Department of Sociology, Umeå University SE-901 87 Umeå (Sweden)

Abstract

Recent research has shown that complementary and alternative therapies (CAM) are increasingly popular in the western world. As a consequence, CAM is becoming more integrated into conventional medical practices. However, this could be seen as a paradox since most CAM relies on spiritual and religious assumptions, sometimes at odds with western scientific traditions. The purpose of this paper is to study the development of research concerned with CAM. To define and retrieve publications on complementary and alternative therapies we use MEDLINE Medical Subject Heading (MeSH). In the retrieval process we use all entries under category complementary therapies, except traditional medicine since this does not denote a specific therapeutic tradition. In the article we analyse general patterns concerning the development of research activity in different CAM traditions. In the article we conclude that the publication activity in CAM increases rapidly, and that the changing growth rate of CAM articles is not due to a general expansion of Medline. We also find that the character of CAM articles has changed, especially at the beginning of the 1990th, towards more clinical oriented research.

Keyword

complementary and alternative medicine; bibliometrics; publication behaviour; scientific journals

Introduction

Recent research, American, Australian and European, indicate that complementary and alternative medicines (CAM) are increasingly popular – and frequently used by the general population. For example, Eisenberg et. al. (1993; 1998) has shown that the use of 16 CAM-therapies, such as homeopathy, massage and energy healing, has increased from 34% in 1993 to 42% seven years later in the USA (see also Harris & Rees 2000). In an Australian study, the researchers found that 48,5% of the respondents had used at least one non-medical prescribed alternative medicine, and that 20,3% had visited at least one alternative practitioner (MacLennan, Wilson & Taylor 1996). However, there are also indications on that the use of alternative therapies are not equal among populations. Women, relatively well educated, with poorer health status, and with a holistic orientation to health, seem to be more likely as users. Many of them are also skeptical towards traditional biomedicine (e.g. Austin 1998, MacLennan, Wilson & Taylor 1996, Millar 1997).

From the historical and sociological research of CAM we can conclude that the role and status of CAM, in western societies, has changed dramatically over time (e.g. Salmon [ed.] 1984). During the late nineteenth century, many CAM-practitioners, such as homeopaths and herbalists, practiced alongside with medical professionals. Later on, with the growth of biomedical research and professionalization, development of new technologies, and diagnostic tools, many CAM-occupations became marginalized and legally restricted (Kelner et. al. 2004). During the twentieth century, the pendulum has turned again. Many CAM-therapies has gained increased societal recognition and acceptance – and gone from a marginal or fringe positions to become integrated into biomedical practices and public health sectors (for overviews, see Eklöf [ed.] 2004; Jütte, Eklöf & Nelson [ed.] 2001). For example, this is the case of such therapies as psychotherapy, acupuncture, naprapathy, osteopathy, and chiropractic – which in many western countries are possible treatments within the public health sector – either performed by medical professionals (such as midwives and doctors) or by CAM-professionals in collaboration with public health. In some cases, as with psychotherapy, the integration has advanced to a point where the therapies no longer are regarded as complementary or alternative – they are fully integrated.

However, acceptance and integration are far from simple processes – they can be identified at many levels and from different perspectives. In most countries there are both symbolic and institutional divisions between conventional/regular/scientific/biomedicine at one hand – and unconventional/fringe/alternative/holistic medicine on the other (see Worsley 1982; Jütte 2001). One crucial aspect is that most CAM therapies, at least to some extent, relies on spiritual, religious, or magic assumptions – in conflict with dominating scientific norms and traditions. Neither is it given that CAM practitioners strive for integration, since it often requires compromises, or even giving up central ideas and/or parts of practice. As David Hess points out, when CAM do make their way into biomedical research projects, they usually shorn off their non-western theoretical frameworks and reinterpreted in terms of current biomedical and psychological knowledge (1995: 201). As a consequence, many practitioners protect ideas of alternative rationality, concerning such aspects as tools of diagnose and mode of treatment (cf. Hess 1993). From that perspective, being alternative is an advantage – especially in those cases when conventional medicine is regarded as insufficient to cure illness or reduce various sorts of pain. The alternative modes of treatment, or diagnosis, are not restricted to what is accepted knowledge, according to established medical or scientific communities.¹ Among CAM practitioners we also find varying attitudes towards the division between science and non-science. While some, such as many parapsychologists and chiropractors, view their own tradition as scientific, albeit not generally accepted by the broader scientific community (see e.g. Hess 1993; Martin 1994) – others emphasize spiritual or religious roots and engage in debates with established medical and scientific communities only to a limited extent. The line of division, between science and non-science, is, of course, not only of interest to those who strive for recognition and acceptance – but also to the medical and scientific establishment. As Thomas Gieryn (1999) points out, to keep the privileges of being trusted authorities it is crucial to set up clear boundaries to other producers of knowledge. Therefore, it is hardly surprising to find hard resistance among medical professionals and various scientists towards clinical studies on effects of CAM practices or collaboration with CAM practitioners.

There are many arenas where it is possible to observe so-called boundary work (e.g. Gieryn 1999; Star & Griesemer 1989) both between CAM and medical establishment and within each camp. There are several actors engaged (such as CAM and medical professionals, organizations, public authorities, and even states) in varying negotiations, on different boundaries (for example, what is considered as reliable or scientific knowledge, rights to exercise practice, economical subsidies), and/or boundary objects. In this article we will analyze one type of boundary work, namely how CAM research makes its entry into established scientific arenas. With help from bibliometric methods, we will study the general publication activity of CAM-articles, during the period 1966-2003. We will also analyze the content of CAM research and if/how it has changed over time. Finally, we will analyze in what journals the publications are found. Is there, for example, CAM research found within specialized CAM-journals?

Data

This study is restricted to published items, classified as *Journal Articles*, during the period 1966-2003, in the Medline database. We selected the CAM-articles as defined by the MESH (Medical Subject Headings) category *Complementary Therapies*. However, in this category we excluded items indexed as *Traditional Medicine*, such as African medicine and Chinese medicine, since these are not considered as alternative or complementary in their own cultural context. Traditional medicines are neither therapies in a restricted sense, but general traditions. It should be noted that traditional medicines only are excluded in the search profile. As we will see later on in this study, they will appear in other categories.

¹ It is crucial to recognize that CAM practitioners exercise their occupation under varying circumstances. The degree of regulation varies among countries. As Fisher & Ward (1994) point out, in most western countries only registered health professionals may practice, while in others, such as in the United Kingdom, practice is almost unregulated.

General development of publication activity

One common utterance about CAM research, not at least among medical professionals, is that there is no such thing as CAM research, at least if we talk about reliable research, conducted in accordance with established scientific procedures (such as clinical trials) and published in well-reputed academic journals. Therefore, let us begin with a simple overview of the publication activity. Figure 1 shows the number of CAM journal articles in Medline during the period 1966 to 2003. The articles grow at a steady rate, with exception from a small increase in the middle of the 1970th, from 1966 to 1996, with an average growth rate of 44 articles per years. However, in 1996-97 something dramatic happens. During the rest of the period, the growth rate increases to an average of 424 articles per year. It should be noted that these numbers tells nothing about the content of the CAM articles; they can be almost everything related to CAM – clinical trials on specific therapies, general research about CAM, overviews, debates – as long they are classified as a journal articles in Medline.

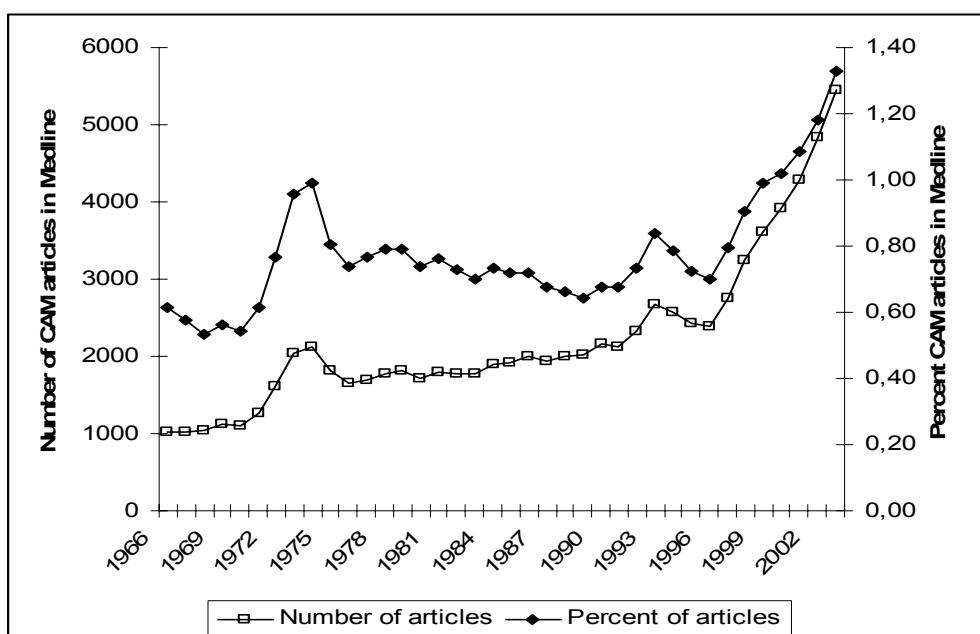


Figure 1. Articles indexed as Complementary Therapies in Medline during the period 1966-2003

How is it possible to explain this change in publication activity? One answer could be that the numbers not are reflecting a real change, but an increase in the number of journals indexed in Medline. Another answer could be that the increase in publication activity reflects a general increase in research and/or publication activity. To test these questions we have compared the share of CAM articles in Medline over time.

Figure 1 also shows the percentage of CAM articles in Medline as a whole. Here we find that the share of CAM articles is relatively constant from 1966 to 1996 (again; with exception for a peak in the beginning of the 1970th). After 1996 the share of CAM articles increases rapidly, i.e. faster than Medline as a whole. We can therefore conclude that the changing growth rate of CAM is not due to a general expansion of Medline.

Content and development of CAM research

In order to grasp the development of content in CAM research we have analyzed the prevalence of clinical oriented research. Clinical trials are, not at least by researchers and clinicians, considered as the most credible way to test different kinds of medical treatments. For example, the U.S. National Institute of Health declare that.Carefully conducted clinical trials are the fastest and safest way to find treatments that work in people and ways to improve health². From a CAM perspective,

²<http://clinicaltrials.gov/ct/info/whatis#whatis>

accomplishment of clinical trials are reasonably crucial (even though not the only way) to get legitimacy from a broader scientific community. However, in the boundary work, over scientific claims, clinical trials are also crucial to the opponent of CAM – as a legitimate way of falsifying CAM theories and practices.

In MESH, clinical trials are defined as Pre-planned studies of safety, efficacy, or optimum dosage schedule (of appropriate of one or more diagnostic, therapeutic or prophylactic drug, devices, or techniques selected according to predetermined criteria of eligibility and observed for predefined evidence of favourable and unfavourable effects³). Furthermore, clinical trials can be divided into treatment-, prevention-, diagnostic-, screening-, and quality of life trials. They are also conducted in different phases (I-IV). Clinical trials can take place in various locations, such as universities, community clinics, and hospitals, and can be funded by both private and federal agencies.

In figure 2 we can see that the character of CAM articles has changed, especially at the beginning of the 1990th, towards more clinical oriented research. A similar development can be identified in Medline as a whole. The clinical trials expands from about 3% of the published articles in the early 1990th up to 6% at the end of the decade. However, the change is much more dramatic among the CAM articles; from 7% in 1993 up to the double in a few years. From Figure 2 we can also conclude that the expansion of clinical trials takes place both within and outside the journal category of Complementary Therapies.

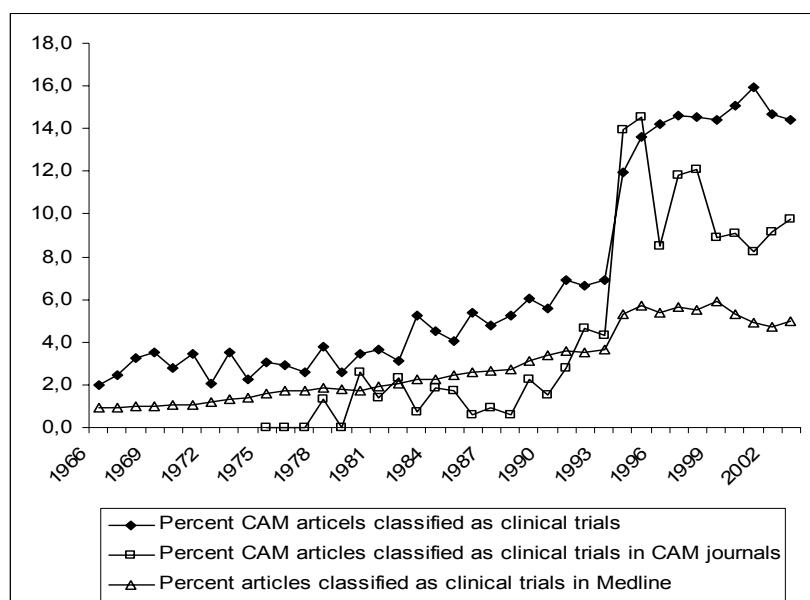


Figure 2. Percent articles classified as clinical trials.

The fact that the share of clinical trials in CAM journals increases rapidly in the early 90th might indicate that new types of journals are established. It can also reflect a general strive for scientific proofs (or from an antagonistic perspective; falsification) – and acceptance from the scientific and medical establishment (cf. Hess 1993) in order receive integration of CAM therapies in public health sectors. Another question, that arises from the figures above, is if there is a general increase in clinical trials among CAM-therapies – or just among a few (for example, among those therapies that already are relatively well integrated – such as acupuncture, naprapathy, and osteopathy)?

In Figure 3, with help from the hierarchical categories in MESH, we separate four dominating sub fields (Mind Body therapies, Spiritual therapies, Musculoskeletal manipulations, and Acupuncture) in the overall CAM category. Interestingly, we can see that there is a substantial increase in all four sub

³http://mesh.kib.ki.se/swemesh/show.swemeshtree.cfm?Mesh_No=E05.318.760.535&tool=karolinska

fields. However, the increase is, as suspected, most dramatic among Musculoskeletal manipulations and Acupuncture, therapies that in many western societies are fairly well integrated in public health sectors.

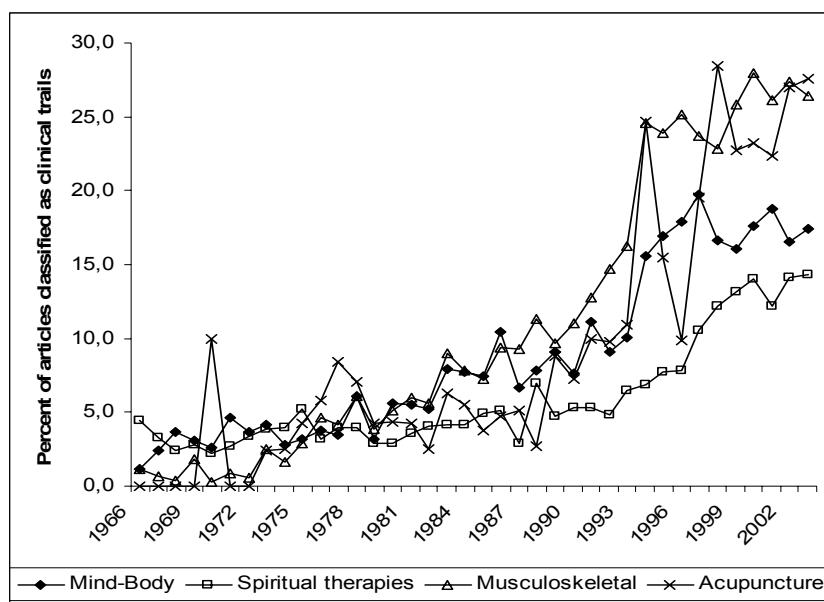


Figure 3. Percent of Clinical Trials in for CAM subfields; Mind Body therapies, Spiritual therapies, Musculoskeletal manipulations, and Acupuncture.

Establishment of a CAM field of journals

In Figure 1 we displayed all articles in Medline, indexed as Complementary Therapies. However, in 1975 Complementary Therapies also appears as a journal category in Medline. This can imply that this is the point of time when CAM in general, or CAM research and journals in particular, is enough recognized and/or accepted to be included in Medline. This can also imply that some journals are re-indexed, and/or that the journals in this category are fairly new (and therefore not included in Medline before). To find answer, we have compared when the journals appear in the CAM journal category, if they were indexed in another Medline category before, and when the journal originally was founded.⁴

The first journal indexed as Complementary Therapies in Medline was *Chinese Medical Journal*, founded as early as 1887 but not indexed in Medline until 1975. However, most journals in this category follow another pattern; they were founded during the 1970th to the 1990th and got included in Medline in a few years, or in some cases about a decade, delay. Two American journals are, exceptionally, indexed in Medline from their first issue. Two other journals, one homeopathic and one fitotherapeutic, are, in similarity with *Chinese Medical Journal*, old journals, not included in Medline until they got included in the Complementary Therapy-category. In other words, there are no re-indexations, and the majority of the journals in this category are relatively new.

One relevant question, in order to get a clearer picture of the assumed increase in publication activity (see Figure 1), is in what journals the CAM articles appear. Is the observed change in growth rate, especially after 1996, due to the increase in the number of journals indexed as Complementary therapies?

⁴ Since year of foundation not is included in Medline, this information has been taken from the official webpages of the journals. In some cases, as for the Chinese journals, the year of foundation has been estimated from the number of volumes.

Table 1. Journals indexed as Complementary Therapies in Medline, 1975-2003.

Name of Journal	First year in Medline	Year of foundation
<i>Acupuncture and Electro Therapeutics Research</i>	1981	1975
<i>Acupuncture in Medicine – J. of the British Med. Acupuncture Society</i>	2001	1981
<i>Advances in Mind Body Medicine</i>	1999	1985
<i>Alternative Therapies in Health and Medicine</i>	1995	1995
<i>The American Journal of Chinese medicine</i>	1979	1972
<i>China Journal of Chinese Material Medica</i>	1989	1976
<i>Chinese Medical Journal</i>	1975	1887
<i>Chinese Medical Sciences Journal</i>	1991	1986
<i>Complementary Therapies in Medicine</i>	1999	1993
<i>Fitoterapia</i>	2000	1929
<i>Forschende Komplementarmedizin und Klassische Naturheilkunde</i>	2000	1994
<i>Homeopathy – the Journal of the Faculty of Homeopathy</i>	2002	1912/1995
<i>Journal of Alternative and Complementary Medicine</i>	1995	1995
<i>Journal of Chinese Medicinal Materials</i>	1997	1978
<i>Journal of Traditional Chinese Medicine</i>	1981	1981
<i>Phytomedicine International Journal of Phytotherapy and Phytopharmacology</i>	1999	1994
<i>Phytotherapy Research PTR</i>	1999	1987
<i>Chinese Journal of Integrated Traditional and Western Medicine</i>	1992	1981

In figure 4 we subtract articles published in CAM journals and journals indexed as Medicine. It can be observed that the steady increase in CAM articles between 1975 and 1996 in great part is due to the increase in CAM journals. However, after 1996 the field displays a rapid growth both *within* and *outside* this journal category. The expansion within the CAM category evolves from 2,7% in 1975 (when just one journal is indexed in this category) to 36,4% in 2003. In this figure we can also see that the number of articles published in journals indexed as Medicine is relatively stable, although the share is decreasing. These results are very much in line with the results of Table 1, where we can see that as many as 12 of the 18 journals, in the category of Complementary therapies, are included in Medline after 1996 – i.e. during the period of dramatic expansion of CAM as a scientific field.

The few number of journals in the category of Complementary Therapies can probably be explained by the fact that there are few multiple classifications in the database (in this specific category, there are none). If we look at other journal categories, such as Nursing, with a great share of the CAM articles, we find many journals that *also* could be classified as CAM (such as Complementary Therapies in Nursing and Midwifery, Holistic Nursing Practice, Journal of Holistic Nursing, Holistic nursing, British Homeopathic Journal). Measures of expansion within and outside journal categories, such as CAM, therefore seems relatively rough. We can for that reason expect an underestimation of the number of CAM journals and the volume of articles published in these.

Establishment of specific CAM-journals can be interpreted in many ways. First of all, it can be a sign of greater public interest in CAM in general – and in different CAM subfields. The specialized journals can also be an indication of greater research activity and a scientific need of specialization. However, these journals could also be interpreted as strategies to create separate scientific fields, governed by, at least in part, other scientific standards than biomedicine. This, in turn, can also be the result of exclusion from conventional medical forums. These strategies have been observed in the case of parapsychological research, when J. B. Rhine founded a separate journal off campus. As David Hess points out, would the parapsychological interpretation be that Rhine, to a large extent, was forced into a separatist position by the scepticisms by his colleagues (Hess 1993: 26; Mauskopf & Mc Vaugh 1980).

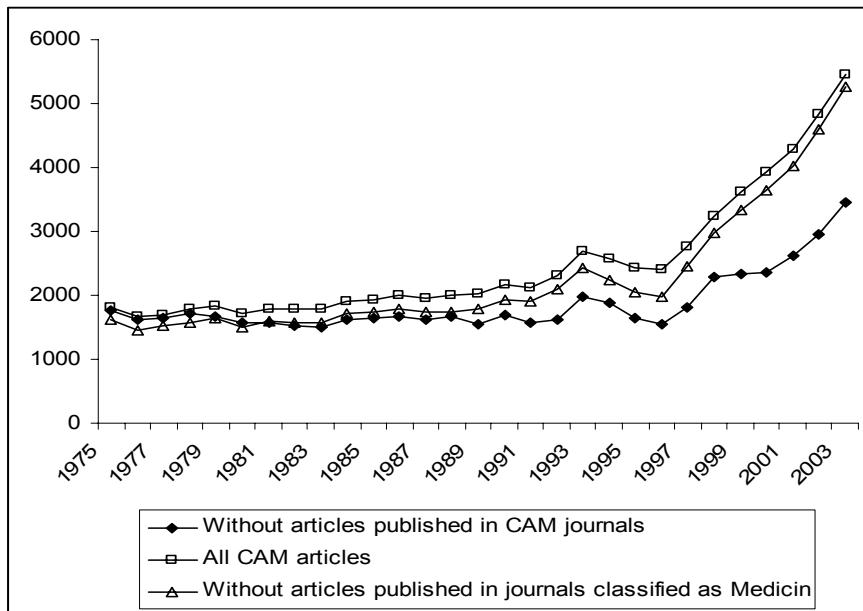


Figure 4. Decomposed articles sets by journal categories

Conclusions

In this article we have analyzed the general publication activity of CAM-articles and orientation of CAM research has changed during the period 1966-2003. Based on our observation we conclude that the publication activity in CAM increases rapidly, and that the changing growth rate of CAM articles is not due to a general expansion of Medline. It is observed that the number of CAM articles in MEDLINE grows at a steady rate from 1966 to 1996, with an average growth rate of 44 articles per years. However, in 1996-97 the growth rate increases to an average of 424 articles per year. The steady increase in CAM articles between 1975 and 1996 in great part is due to the increase in CAM journals. However, after 1996 the field displays a rapid growth both *within* and *outside* this journal category. We also conclude that the character of CAM articles has changed, especially at the beginning of the 1990th, towards more clinical oriented research. We observe an expansion of CAM articles classified as clinical trials both within and outside the CAM journals, and that the increase in clinical research is substantial all major CAM sub fields. However, the shift toward more clinical trials is most dramatic among Musculoskeletal manipulations and Acupuncture, therapies that in many western societies are fairly well integrated in public health sectors.

References

- Austin, J. A. (1998). Why Patients Use Alternative Medicine – Results of a National Study, *JAMA – Journal of the American Medical Association*, 279 (19), 1548-1553.
- Eisenberg, D. M. et. al. (1993). Unconventional Medicine in the United States – Prevalence, Costs, and Patterns of Use, *The New England Journal of Medicine*, 328 (4), 246-252.
- Eisenberg, D. M. et. al. (1998). Trends in Alternative Medicine Use in the United States, 1990-1997, *JAMA*, 280, 1569-1575.
- Eklöf, M. [ed.] (2004) *Perspektiv på komplementär medicin*. Lund: Studentlitteratur.
- Fisher, P. & Ward, A. (1994). Complementary Medicine in Europe, *British Medical Journal*, 309 (6947), 107-11.
- Gieryn, T. F. (1999) *Cultural Boundaries of Science – Credibility on the Line*. Chicago: University of Chicago Press.
- Harris, P. & Rees, R. (2000). The Prevalence of Complementary and Alternative Medicine among the General Population: a Systematic Review of the Literature, *Complementary Therapies in Medicine*, (8), 88-96.
- Hess, D. (1993) *Science in the New Age – the Paranormal, Its defenders and Debunkers, and American Culture*. Madison: The University of Wisconsin Press.
- Hess, D. J. (1995) *Science and Technology in a Multicultural World*. New York: Columbia University Press.
- Kelner, M. et. al. (2004). The Role of the State in the Social Inclusion of Complementary and Alternative Occupations, *Complementary Therapies in Medicine*, 12, 79-89.

- Jütte, R. (2001).Alternative Medicine and Medico-Historical Semantics in Jütte, R. , Eklöf, M. & Nelson, M. C. [ed.] (2001) *Historical Aspects of Unconventional Medicine – Approaches, Concepts, Case Studies*. Sheffield: European Association for the History of Medicine and Health Publications.
- Jütte, R. , Eklöf, M. & Nelson, M. C. [ed.] (2001) *Historical Aspects of Unconventional Medicine – Approaches, Concepts, Case Studies*. Sheffield: European Association for the History of Medicine and Health Publications.
- MacLennan, A. H., Wilson, D. H. & Taylor A. W. (1996) Prevalence and cost of alternative medicine in Australia, *Lancet*, 347 (9001), 569-573.
- Martin, Steven C. (1994).'The Only Truly scientific Method of Healing' Chiropractic and American Science, 1895-1990, *Isis*, 85, 207-27.
- Mauskopf, S. & Mc Vaugh, M. (1980) *The Elusive Science: Origins of Experimental Psychology*. Baltimore: The Johns Hopkins University Press.
- Millar, W. J. (1997).Use of Alternative health care by Canadians, *Canadian Journal of Public Health*, 88 (3), 154-158.
- Salmon, J. W. [ed.] (1984) *Alternative Medicines –Popular and Policy Perspectives*. New York: Tavistock Publications.
- Star, S. L. & Griesemer, J. R. (1989).Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkley's Museum of Vertebrate Zoology, 1907-39, *Social Studies of Science*, 19, 387-420.
- Worsley, P. (1982).Non-western Medical Systems, *Annual Review of Anthropology*, 11, 315-348.

Hirsch's h-index and Egghe's g-index

Quentin L. Burrell

q.burrell@ibs.ac.im

Isle of Man International Business School, Isle of Man IM2 1QB (United Kingdom)

Abstract

In a recent issue of the ISSI Newsletter, Egghe (2006a) proposed the g-index, claimed to be an improvement on the original h-index proposed by Hirsch (2005). The aim of this paper is to investigate the inter-relationships between these two measures and also their time dependence using the stochastic publication/citation model proposed by Burrell (1992, 2007a). We find that an author's g-index is directly proportional to career length and hence to the h-index. We also make some tentative suggestions regarding the relative merits of these proposed measures.

Keywords

H-index; G-index; stochastic model; publication/citation process.

Introduction

The proposal by Hirsch (2005) to introduce a single index to quantify a scientist's published research impact created an unprecedented response from the scientometric community. Within a year we saw responses from, among others, Banks (2006), Bornmann & Daniel (2005), Braun et al. (2005), Cronin & Meho (2006), Diniz Batista et al. (2005), Egghe (2006a, b, 2007), Glänzel (2006a, b), Liang (2006), Popov (2005), Rousseau (2006a, b) and van Raan (2006). Some of these have sought to demonstrate empirical applications of the index, some to extend its applicability and others to provide mathematical models, most notably Egghe & Rousseau (2006), Glänzel (2006a) and Burrell (2006, 2007a).

Others have proposed alternative measures, similar to or based upon the h-index. The "size" of the h-core (Rousseau, 2006b) and the A-index (Jin, 2006 and Rousseau, 2006b) have been analysed by Burrell (2007b) using a stochastic model for the publication/citation process proposed by Burrell (1992, 2007a). The idea that it might be more appropriate to use the h-rate, rather than the h-index, has been argued by Burrell (2007c) based upon the work of Liang (2006). In this paper we use this same stochastic model to investigate aspects of the g-index proposed by Egghe (2006a, b).

The two indexes

According to the preprint version of Hirsch (2005), the h-index for an author is that integer h such that h of his/her papers have at least h citations each, while the rest have fewer than h citations. Actually this is not quite well-defined, see the print version of Hirsch (2005), Glänzel (2006a) and Rousseau (2006b), since there is ambiguity if there are several papers with the same number of citations at h . To get round this, let us introduce

Notation. Write $f(n; T)$ for the number of an author's papers receiving exactly n citations by time T , and $N(n; T)$ for the number of an author's papers that have received at least n citations by time T so that $N(n; T) = \sum_{j=n}^{\infty} f(j; T)$ and note that $N(n; T)$ decreases as n increases.

In this notation, the total number of citations received by those publications receiving n citations each is $nf(n; T)$ and hence the total number of citations received by all those receiving at least n citations each is given by $C(n; T) = \sum_{j \geq n} j f(j; T)$. We refer to this as *the size of the n-core* and note that $C(n; T)$ decreases with increasing n .

Remarks. (i) Here and later we include the time parameter T explicitly since one of our aims is to consider how the indexes develop in time as dynamic processes. But note that when we talk of time T , this refers to the time that has elapsed since the start of a particular author's publication career so that

it is important to realise that when we are considering several authors, “now” or “the current time” may correspond to different values of T.

(ii) Note that $f(0;T)$ gives the size of the zero class, the number of an author’s papers that have received no citations by time T; $N(0;T)$ gives the total number of his/her publications by time T; $C(0;T)$ gives the author’s total number of received citations by time T.

Definition 1. Hirsch’s h-index at time T is, for any particular author, the integer $h(T)$ satisfying

$$h(T) = \max \{n : n \leq N(n; T)\}$$

For instance, if this maximal $n = 25$, say, then at least 25 of the author’s papers have received 25 or more citations while fewer than 26 have received 26 or more citations. Note that this is an empirical measure, requiring observation of the actual values of $N(n; T)$.

Egghe’s g-index is rather different in that it switches attention from the number of most productive sources (publications) to the actual number of items (citations) attracted by the most productive sources. In this sense, it is related to the size of the Hirsch core, see Rousseau (2006b) and Burrell (2007b), and Jin’s A-index, see Jin (2006), Rousseau (2006b) and Burrell (2007b).

Another novelty in Egghe’s approach is that a paper’s rank, as determined by its number of received citations, is included explicitly. If we write $r = r(n)$ for the rank of a paper receiving n citations then in this notation Egghe’s g-index can be defined as:

Definition 2. Egghe’s g-index at time T is, for any particular author, the integer $g(T)$ satisfying

$$g(T) = \max \{r : r(n)^2 \leq C(n; T)\}$$

In Egghe (2006a,b) an author’s papers are ranked according to the number of citations received with the convention that when there are several papers with the same number of citations they are ranked serially, with no indication as to how they are to be ordered. (This last point is of no concern in determining the author’s index, only when we wish to decide which papers are included and which excluded from any sort of core. It is analogous to the original ambiguity in the definition of the h-index, and the definition of the h-core (see Burrell, 2007c).) It will be convenient for our analysis to modify slightly this notion of rank. Our approach is that all papers receiving the same number of citations are given the same rank. (This conforms with standard practice in probability and economics, although it is not crucial in what follows, which is intended purely as an illustrative analysis.)

Thus, for any $n = 0, 1, 2, \dots$ we define the rank of n (or of a paper receiving exactly n citations) as the number of papers receiving at least n citations, and note that this is defined whether or not there are any papers receiving (exactly) n citations. In other words, our modified notion of rank is given by what we have already denoted by $N(n; T)$. We thus slightly modify the earlier definition of the g-index to:

Definition 2.* Egghe’s g-index at time T is, for any particular author, the integer $g(T)$ satisfying

$$g(T) = \max \{N(n; T) : N(n; T)^2 \leq C(n; T)\} \quad (1)$$

The stochastic model

Here we just recap the essentials of the model and refer the reader to Burrell (2007a) for full details. The basic idea is that an author publishes papers at certain times and that these papers subsequently attract citations following their publication, where both the publication and citation accumulation processes are random. We further assume that some papers are more citable than others so that the citation rate varies between different publications. The basic ideas behind this were originally put forward in Burrell (1992). The precise technical assumptions, without the mathematical details, are:

Assumptions

1. From the start of his/her publishing career (at time zero), an author publishes papers according to a Poisson process of rate θ which gives the mean number of publications per unit time, called the *publication rate*.

2. Any particular publication acquires citations according to a Poisson process of rate Λ , where Λ varies from paper to paper. Here Λ denotes the mean number of citations to the paper per unit time following publication, called the *citation rate*.

3. The citation rate Λ for this author varies over the set of his/her publications according to a gamma distribution of index $v \geq 1$ and scale parameter $\alpha > 0$. Note that $E[\Lambda] = v/\alpha = \mu$, say, gives the *mean citation rate*.

See Burrell (2007a) for the precise details.

In all of the following analysis, the basic result concerns the distribution of the number of citations garnered (up to the current time) for a typical paper by this author, whenever it was published during his/her career. This is given in the following:

Theorem (Burrell, 2007a)

Under the assumptions of the model, the distribution of X_T , the number of citations to a randomly chosen paper by time T , is given by

$$P(X_T = r) = \frac{\alpha}{(v-1)T} B\left(\frac{T}{\alpha+T}; r+1, v-1\right) \text{ for } r = 0, 1, 2, \dots \quad (2)$$

where $B(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x y^{a-1} (1-y)^{b-1} dy$ is the cumulative distribution function of a beta distribution (of the first kind) with parameters a and b .

To apply the theoretical model, we need to re-interpret the empirical quantities $f(n; T)$, $N(n; T)$ and $C(n; T)$ as expected values of the corresponding random quantities. We will, however, retain the same notation as for the empirical quantities so that, in terms of the stochastic model we have:

Proposition 1.

- (i) $f(n; T) = \frac{\alpha\theta}{(v-1)} B\left(\frac{T}{\alpha+T}; n+1, v-1\right)$ for $n = 0, 1, 2, \dots$. (3)
- (ii) $N(n; T) = \theta T \left(1 - \frac{\alpha}{(v-1)T} \sum_{j=0}^{n-1} B\left(\frac{T}{\alpha+T}; j+1, v-1\right)\right)$ for $n = 0, 1, 2, \dots$. (4)

$$(iii) C(n; T) = \theta T \left(\frac{vT}{2\alpha} - \frac{\alpha}{(v-1)T} \sum_{j=0}^{n-1} j B\left(\frac{T}{\alpha+T}; j+1, v-1\right)\right) \text{ for } n = 0, 1, 2, \dots \quad (5)$$

Proof. See the Appendix.

Remark. Given that our model provides explicit formulae, it is worth noting that, according to the model,

$$N(0; T) = \text{Expected total number of papers published by time } T = \theta T \quad (6)$$

and

$$C(0; T) = \text{Expected total number of citations received by time } T$$

$$= \frac{1}{2}\theta T^2(v/\alpha) = \frac{1}{2}\theta\mu T^2 \quad (7)$$

(For this last result, see the proof of the Proposition in Burrell (2007b).)

From the above it follows that:

Proposition 2.

Given the model assumptions, $g(T)$ is defined if and only if $\theta > v/2\alpha = \mu/2$

This is an interesting result. It says that an author needs a publication rate at least half as big as his/her mean citation rate in order to have a defined g -index. For instance, if you are publishing 2 papers a year, but your papers are attracting on average 5 citations a year, then your g -index is not defined, at least according to the theoretical model. That the *empirical* g -index is not always well-defined was acknowledged by Egghe (2006b) in a “Note added in proof”. His proposed solution was, in such cases

“fictitious articles with 0 citations have to be added (until the g-index can be determined)”. We do not pursue this in what follows, merely note that the g-index is not necessarily defined.

Numerical investigations

The theoretical model involves four parameters:

- (i) The author’s publication rate, θ .
- (ii) The gamma parameters, v and α .
- (iii) The length of the author’s publishing career to the current time, T .

For this investigation we will mainly be concerned with the time dependence of the g-index and how it correlates with Hirsch’s h-index and will illustrate these using particular examples. Specifically, we will look at publication rates of $\theta = 2, 5$ and 10 . Assuming a publication rate of two papers per annum might be thought of as typical in a field such as mathematics, but it would be viewed as low in fields such as biomedicine where there is much collaborative work. Taking the “moderate” publication rate of $\theta = 5$ papers per annum, this would certainly be high in mathematics while in scientometrics, it might be a reasonable value for several of the best-known contributors! Similarly, whether a publication rate of $\theta = 10$ is low, medium or high will depend very much on the discipline.

For the gamma parameters we will focus on the mean citation rate $\mu = v/\alpha$, taking as an illustrative value $\mu = 5$. This could be thought of as a low, medium or high citation rate, depending very much on the subject context. Burrell (2007a) found that the value of μ was much more important in calculations using the model rather than the individual values of α and v . For our illustrative examples we will base all calculations on the fixed values $\alpha = 1$ and $v = 5$.

(i) Determination of the g-index

For any given values of the parameters, we have to determine $g(T)$ according to (1). This first requires calculation of $N(n;T)$, $N(n;T)^2$ and $C(n;T)$ for a range of values of n using the formulae (4) and (5). Evaluation of these quantities is fairly straightforward in any statistical package, such as ExcelTM, that allows evaluation of the cumulative distribution function of the beta distribution. Rather than give this numerical determination, note that for the approximate determination of the g-index by graphical means, it is easy to plot $N(n;T)^2$ and $C(n;T)$ against n . Where these intersect, we can read off the value of $N(n;T)^2$ and hence find the appropriate $N(n;T) = g(T)$. We illustrate this as in Figure 1 (a)-(c) for $T = 5$ and various values of θ .

In Figure 1(a) with $\theta = 10$, the point of intersection appears to be at about $n = 13$, and we can read off the approximate value of the ordinate. In fact, calculations show that $N(13;5)^2 = 425.78 < C(13;5) = 436.82$ while $N(12;5)^2 = 504.33 > C(12;5) = 459.21$ so that $g(5) = N(13;5) = 20.63$.

In the case of Figure 1(b) the point of intersection is at about $n = 7$ and again we can read off the approximate value of the ordinate. Calculations now show that $N(7;5)^2 = 271.38 < C(7;5) = 279.40$ while $N(6;5)^2 = 311.04 > C(6;5) = 287.32$. Hence we have $g(5) = N(7;5) = 16.47$.

Note that in Figure 1(c) there is no point of intersection because here we have an example where $\theta < \mu/2$, in which case, from Proposition 2 we have that the g-index does not exist.

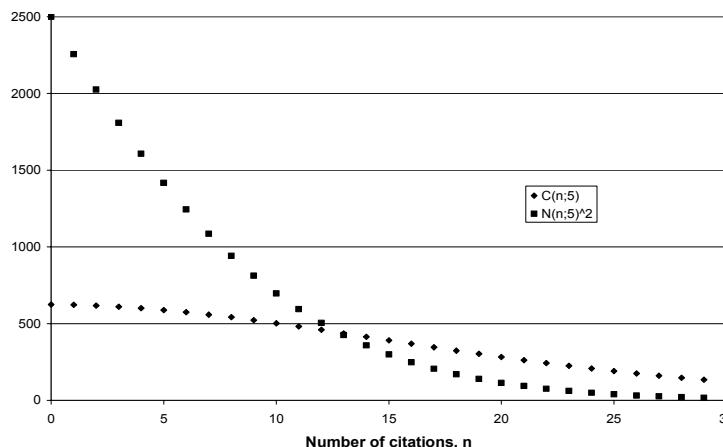
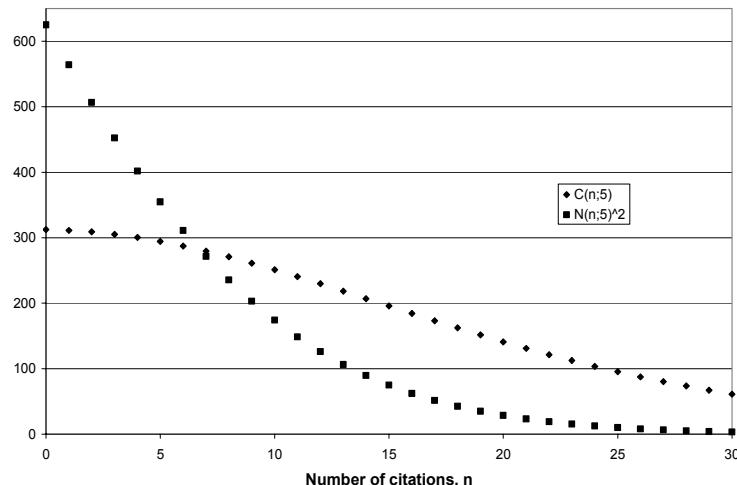
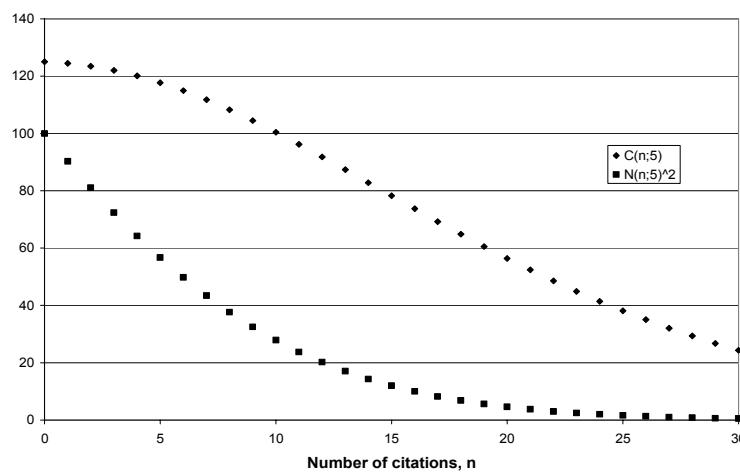
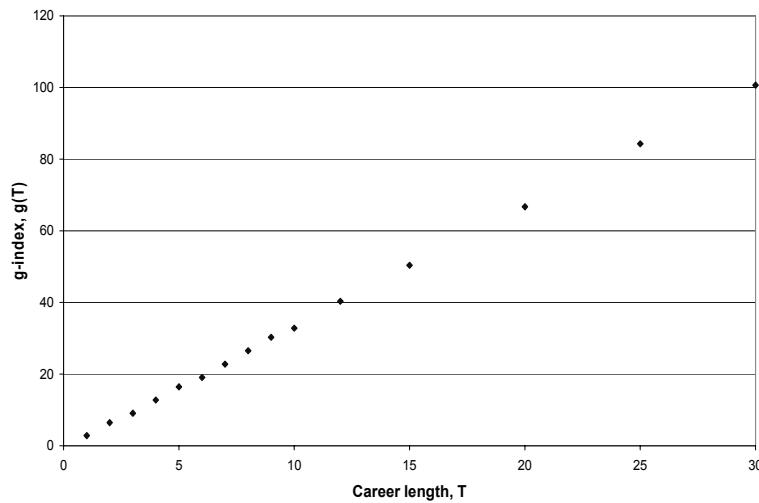


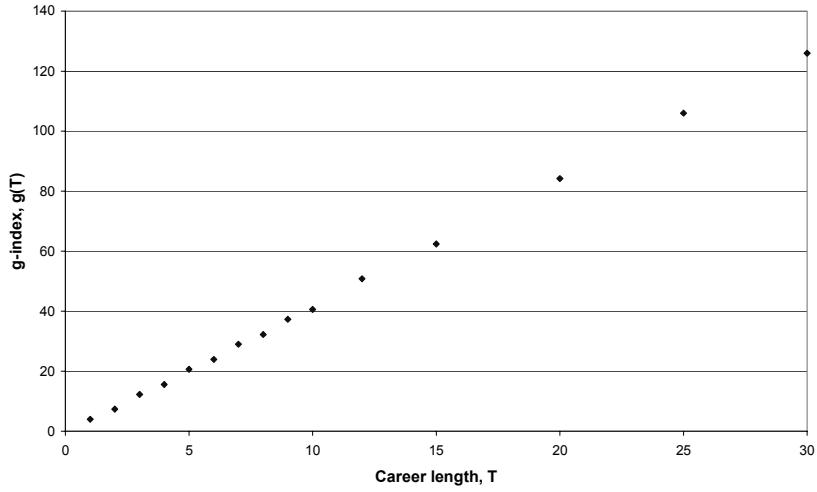
Figure 1(a). Graphical determination of $g(5)$, with $\theta = 10$, $\mu = 5$.

Figure 1(b). Graphical determination of $g(5)$, with $\theta = 5, \mu = 5$.Figure 1(c). Graphical determination of $g(5)$, with $\theta = 2, \mu = 5$.

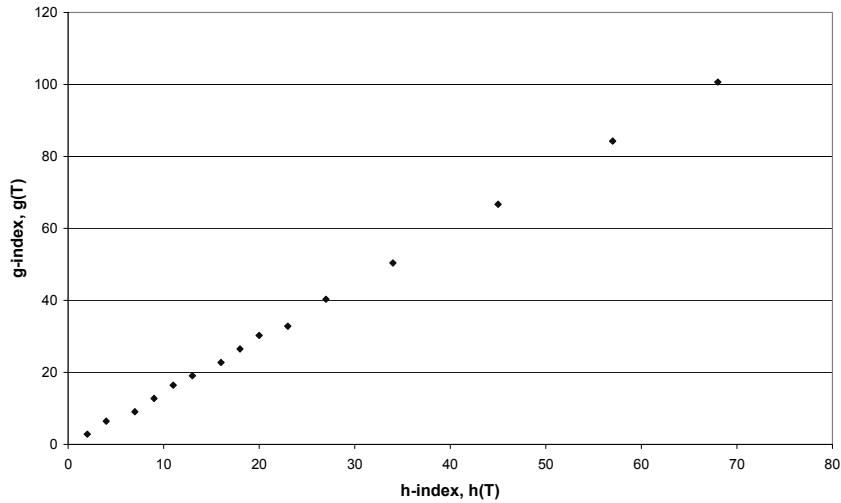
(ii) Time dependence of the g -index

In Figures 2(a, b) we give plots of g against time and note the almost perfect linearity, indeed proportionality. (In both cases we find $R^2 > 99\%$ for a regression line assumed to pass through the origin.)

Figure 2(a). The case $\theta = 5, \mu = 5$.

Figure 2(b). The case $\theta = 10, \mu = 5$.*(iii) Relationship with the h-index*

Burrell (2007a) found, using the same stochastic model, that the h-index is proportional to time, as originally speculated by Hirsch (2005). Given this, together with the result for the g-index, we would expect $g(T)$ to be directly proportional to $h(T)$. This is confirmed by Figures 3(a, b).

Figure 3(a). The relationship between $g(T)$ and $h(T)$ for $\theta = 5$.

It is clear that these plots suggest clear direct proportionality between the indexes. In fact, if we fit regression lines constrained to pass through the origin, a reasonable requirement, in both cases we find $R^2 > 99\%$. This direct proportionality is predicted in Egghe's (2006b) paper, using a non-stochastic model based upon an assumed Lotka form for the distribution of citations. Egghe (2006b) further suggests using the constant of proportionality $g(T)/h(T)$ as a possibly interesting measure to use when comparing the outputs of different scientists. (Note that this is just the slope of the plot as in Figures 3(a, b).) We agree that this would seem to be a possibly fruitful line of inquiry for empirical scientometric studies.

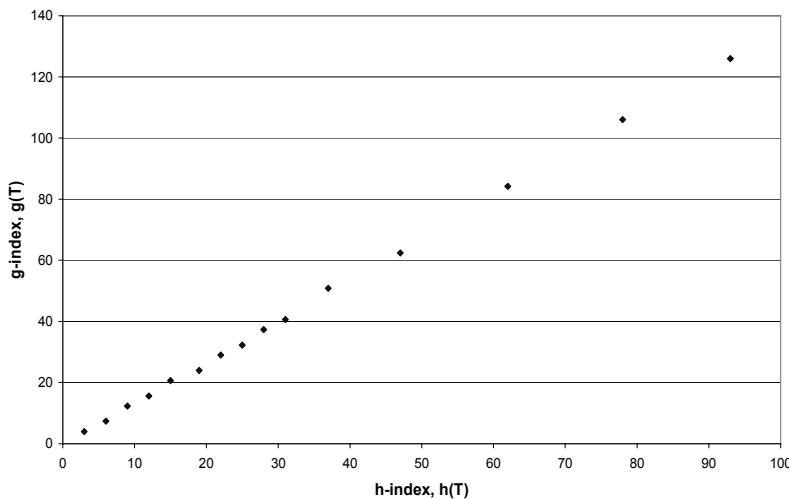


Figure 3(b). The relationship between $g(T)$ and $h(T)$ for $\theta = 10$.

Concluding Remarks

Within the confines of the model, we find that an author's g -index is directly proportional to the current career length and to the h -index, but note that this is based on selected numerical calculations rather than mathematical analysis and hence further theoretical as well as more extensive numerical work is required. From a practical point of view, which index is to be preferred? As we have that for any author, $g \geq h$ this means that accurate determination of the g -index requires more – and possibly very much more – work than does the h -index. On the other hand, Egghe (2006b) shows that in comparative empirical studies the two indexes can lead to different perspectives. It would seem that, at this stage, both measures are worth exploring.

References

- Banks, M. G. (2006). An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics*, 69(1), 161-168.
- Burrell, Q. L. (1992). A simple model for linked informetric processes. *Information Processing and Management*, 28, 637-645.
- Burrell, Q. L. (2006). Hirsch's h -index: a preliminary stochastic model. Book of Abstracts: 9th International Science and Technology Indicators Conference, 7-9 September 2006, Leuven, Belgium, 26-28. Katholieke Universiteit, Leuven.
- Burrell, Q. L. (2007a). Hirsch's h -index: A stochastic model. *Journal of Informetrics*, 1(1), 16-25.
- Burrell, Q. L. (2007b). On the h -index, the size of the Hirsch core and Jin's A- index. *Journal of Informetrics*, (to appear).
- Burrell, Q. L. (2007c). Hirsch index or Hirsch rate? Some thoughts arising from Liang's data. *Scientometrics*, 73(1), (to appear).
- Eghe, L. (2006a). An improvement of the h -index: the g -index. *ISSI Newsletter*, 2(1), 8-9.
- Eghe, L. (2006b). Theory and practice of the g -index. *Scientometrics*, 69(1), 131-152.
- Eghe, L. (2007). Dynamic h -index: the Hirsch index in function of time. *Journal of the American Society for Information Science and Technology*, 58(3), 452-454.
- Eghe, L. & Rousseau, R. (2006). An informetric model for the h -index. *Scientometrics*, 69(1), 121-129.
- Glänzel, W. (2006a). On the H-index – a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67(2), 315-321.
- Glänzel, W. (2006b). On the opportunities and limitations of the H-index. *Science Focus*, 1 (1), 10-11, (in Chinese).
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569-16572. (Also available in preprint form as arXiv:physica/0508113, accessible at <http://xxx.arxiv.org/abs/physics/0508025>.)
- Jin, BH (2006). H-index: an evaluation indicator proposed by scientist. *Science Focus*, 1(1), 8-9. (In Chinese.)
- Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original h -index. *ISSI Newsletter*, 2(3), 4-6.
- Liang, L. (2006). h -index sequence and h -index matrix: Constructions and applications. *Scientometrics*, 69(1), 153-159.

- Rousseau, R. (2006a). A case study: evolution of JASIS' Hirsch index. *Science Focus*, 1 (1), 16-17, (in Chinese).
 (English version available at E-LIS, code 5430.)
- Rousseau, R. (2006b). New developments related to the Hirsch index. *Science Focus*, 1(4), 23-25, (in Chinese).
 (English version available at E-LIS, code 6736.)

Appendix

(a) Proof of Proposition 1

$$\begin{aligned}
 (i) \quad f(n; T) &= E[\# \text{ papers receiving } n \text{ citations by time } T] \\
 &= E[\# \text{ papers by time } T] P(\text{paper receives } n \text{ citations}) \\
 &= \theta T P(X_T = n) \\
 &= \frac{\alpha\theta}{(v-1)} B\left(\frac{T}{\alpha+T}; n+1, v-1\right) \text{ for } n = 0, 1, 2, \dots
 \end{aligned}$$

In the above we have used standard results for the mean of a binomial distribution, the mean of a Poisson process and equation (1) of the Theorem.

$$\begin{aligned}
 (ii) \quad N(n; T) &= E[\# \text{ papers receiving at least } n \text{ citations by time } T] = \sum_{j=n}^{\infty} f(j; T) \\
 &= \frac{\alpha\theta}{(v-1)} \sum_{j \geq n} B\left(\frac{T}{\alpha+T}; j+1, v-1\right) \\
 &= E[\text{total # papers by time } T] \\
 &\quad - E[\# \text{ papers receiving less than } n \text{ citations by time } T] \\
 &= \theta T - \sum_{j=0}^{n-1} f(j; T) \\
 &= \theta T - \frac{\alpha\theta}{(v-1)} \sum_{j=0}^{n-1} B\left(\frac{T}{\alpha+T}; j+1, v-1\right) \\
 &= \theta T \left(1 - \frac{\alpha}{(v-1)T} \sum_{j=0}^{n-1} B\left(\frac{T}{\alpha+T}; j+1, v-1\right)\right)
 \end{aligned}$$

Here we have used the standard result for the mean of a Poisson process together with the result (3).

$$\begin{aligned}
 (iii) \quad C(n; T) &= E[\text{total # citations for those papers receiving at least } n \text{ citations each}] \\
 &= \sum_{j \geq n} j f(j; T) \\
 &= E[\text{total # citations for all papers}] \\
 &\quad - E[\text{total # citations for those receiving } < n \text{ citations each}] \\
 &= \frac{1}{2}\theta T^2(v/\alpha) - \sum_{j=0}^{n-1} j f(j; T) \\
 &= \theta T \left(\frac{vT}{2\alpha} - \frac{\alpha}{(v-1)T} \sum_{j=0}^{n-1} j B\left(\frac{T}{\alpha+T}; j+1, v-1\right)\right)
 \end{aligned}$$

This follows from the use of (5) and (2).

Is there a Role for Novel Citation Measures for the Social Sciences and Humanities in a National Research Assessment Exercise?

Linda Butler and Kumara Henadeera

linda.butler@anu.edu.au

Research Evaluation and Policy Project Australian National University, ACT 0200 (Australia)

Abstract

Australia is about to move to a new system of distributing government block grants for research among universities, with the introduction of a process similar to Britain's Research Assessment Exercise. In the Australian model, peer judgements will be informed by quantitative performance measures, including citation analysis. However, standard bibliometric measures are widely acknowledged to be inappropriate for most disciplines in the social sciences and humanities. In an attempt to identify a more suitable alternative, two recently completed pilot studies have trialled new approaches to bibliometrics, testing their applicability for the assessment of research in political science and history. The new methodology, which extended citation data to included citations to books, book chapters and journals outside the citation databases, was endorsed at discipline workshops by senior academics from the two disciplines. They found that this new approach to bibliometrics could be a valuable tool for both disciplines, with some caveats attached. The chief reservation related to the role of quantitative measures – there was consensus that they should be used to inform peer review, rather than drive the assessment process.

Keywords

citations to books; social sciences; humanities; research assessment.

Introduction

In November 2006, the Australian Department of Education, Science and Training (DEST) announced a major change to the way in which it will allocate block funding for research to universities (DEST, 2006a). Following an assessment to be undertaken in 2008, funding from 2009 will no longer be allocated through a formula constructed from data on higher degree students, publications and competitive grant income. Instead, it will be distributed through a Research Quality Framework (RQF) assessment, loosely modelled on the British Research Assessment Exercise (RAE), though with significant local modifications.

The framework retains the RAE's expert peer review process examining the four 'best' outputs nominated by research active staff to assess the quality of research groups. But in addition, panel deliberations will be "assisted by the inclusion of relevant and appropriate quantitative measures of research quality" (DEST, 2006a).

After extensive sector consultation through a Metrics Working Group (MWG) established by the RQF development team (DEST, 2006b), three types of measures were recommended for use in this context:

1. Ranked outputs — distribution of selected category(ies) of research output across a limited number of bands, based on predetermined discipline-specific rankings (refereed journals, professional journals, book publishers, conferences, performance venues, etc);
2. Citation data (judged against discipline-specific world benchmarks) — citations per publication; and the proportion of output falling in the top deciles for the discipline; and
3. Grant income data (judged against discipline-specific national benchmarks) — income from Australian competitive funding schemes and international peer reviewed sources.

While the MWG recommended that the full 'basket of measures' be used wherever possible, it was acknowledged that citation data, and in particular the measures listed in point 2 above, would not be appropriate for many disciplines in the social sciences, arts and humanities, or for a number of the applied sciences. Moed (2005) has recently examined this issue in detail and demonstrated their relatively poor coverage by Thomson Scientific's citation databases .

The MWG recommended that further research be undertaken to determine whether alternative bibliometric measures currently being developed might provide acceptable alternatives for some disciplines. These developments seek to extend the coverage of output from the disciplines by extracting citation counts to publications other than articles in ISI¹-indexed journals.

The feasibility of undertaking this type of analysis on a large scale had already been confirmed (Butler & Visser, 2006). However, further investigation was required to determine whether such measures resulted in robust results that could prove useful in a national research assessment exercise; and whether they could gain acceptance from a community understandably reluctant to embrace the more common citation measures usually proposed for the sciences in such exercises. As no analysis of this type and scale has ever been undertaken, the proposal to use it in the RQF is breaking new ground.

This paper reports the results of two pilot studies undertaken by the Australian Council for the Humanities, Arts and Social Sciences (CHASS) on behalf of DEST, in the disciplines of political science and history. CHASS had previously raised concerns about the propensity of analysts to assess research in their disciplines using science-oriented indicators (CHASS 2005), and was keen to explore alternatives more sympathetic to the modes of communication of the disciplines it embraced.

Methodology

While standard bibliometric measures only utilise the citations to publications in ISI-indexed journals, ISI's Web of Science (WoS) records many additional references to books, book chapters, articles in non-ISI journals, and other forms of publication (known collectively as 'non-source' items). There is growing recognition of the magnitude of this segment of their databases (Van Leeuwen, 2006). To this point in time, these data have not been utilised in research evaluation. However, in an attempt to overcome the limitations of standard bibliometric measures, their potential in this context is starting to be recognised. Research is being undertaken into 'mining' the WoS indexes for citations to these non-source items, and deploying them in the development of novel bibliometric indicators, particularly in the social sciences and humanities (Butler & Visser, 2006). The two different reference universes – that confined to ISI-indexed journals, and the wider domain encompassing all the references in the database – are represented diagrammatically in Figure 1.

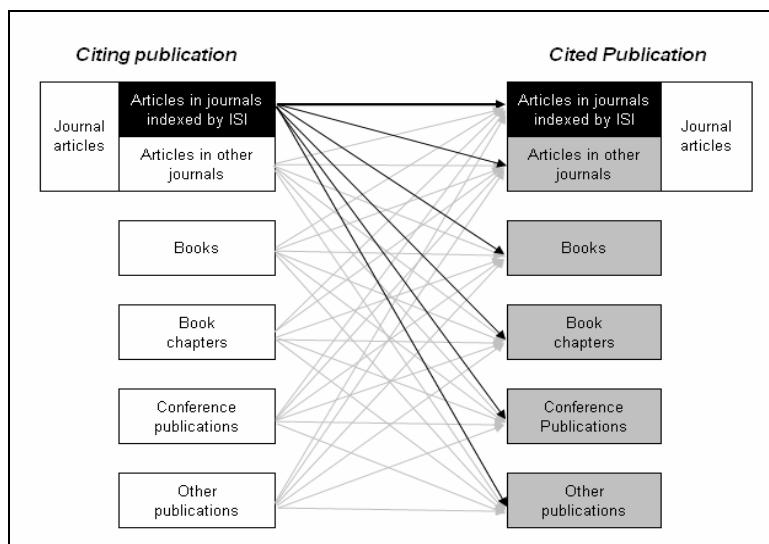


Figure 1. ISI coverage of publications and citations

¹ The citation indexes now compiled by Thomson Scientific were originally compiled by the Institute for Scientific Information (ISI). Thomson retains this abbreviation for a number of its products and I refer to the indexes as ISI throughout this paper.

The references indicated by the bold solid line between the top two boxes are those used in standard citation analysis. In this project, the references indicated by the first three regular lines have been added to the analysis. As can be seen, this new approach still does not cover references indicated by the grey lines, such as book to book citations, nor references from the indexed literature to conference publications and other publication types (indicated by the bottom two regular lines).

Citations to non-source items were identified using ISI's Cited Reference search facility, rather than the more commonly used General Search query page. A search using the Cited Reference query form returns all cited publications that meet the specified search criteria, whether they are in the indexed journals or in other types of output (i.e. those depicted in Figure 1 by all the black lines, bold and regular). Using the General Search query form would only have returned citation data on articles in the indexed journals (i.e. the solid black line in Figure 1)

Extracting non-source citations is a time-consuming process, as the references are not standardised in the way citations to ISI journal articles are, and only the first author of any publication is indexed. This necessitates access to full bibliographic details from either CVs or, as in this case, a list of publications supplied by institutions. Nevertheless, the methodology enables analysts to extract a body of data far greater than that resting solely on ISI journal publications, and for some disciplines this may enable a reasonably robust analysis to be undertaken.

Australian universities are currently required to annually report details of their output in four publication categories to DEST for incorporation into their current funding formula. The 'DEST categories', as they have come to be known, are books and book chapters from commercial publishers, peer reviewed journal articles, and peer reviewed conference papers. For this project, we approached all universities, and nineteen provided details of their DEST publications (excluding conferences) for political science and history. Not all universities had departments of the relevant disciplines, some found it impossible to extract the data in the short timeframe required, and others were apprehensive of the study's aims. To allay any concerns of those who did participate, anonymity in any published report was guaranteed.

We requested all data from departments in the two disciplines, irrespective of the field in which individual academics were publishing. Most data was provided on this basis, though a small number of universities sent details of all publications coded to political science and history. These were primarily from universities with no such departments, and where the publications came from multidisciplinary organisational units. The data cover a six year period: 2000-2005, with both publications and citations counted within this six year window. The period was chosen to replicate closely the likely length of time covered by the RQF (though not necessarily the exact period) and the expected citation window (i.e. the time frame in which publications could attract citations).

The citation data for all publications were extracted from the WoS in October 2006, and multiple instances of varying references to the same publication were aggregated to provide 'clean' counts. The data was then aggregated by discipline and institution, and a range of measures calculated:

- total citations – for all publications, and for each type of publication (books, book chapters, and journal articles);
- citations per publication – the total number of citations was divided by the number of publications reported by the institution. Calculations were undertaken for all publications, and for each type of publication; and
- ISI citation rates – citation per publication calculations were made for journal articles, limited to those in journals indexed by ISI (i.e. equivalent to a standard bibliometric measure).

CHASS convened two workshops in November with senior academics from the two disciplines. A set of tables was provided for each discipline and distributed, together with an overview of the methodology, prior to the workshops.

The main questions considered at the workshops were:

- Does the ‘picture’ painted by the data coincide with the participants’ knowledge of the relative strengths and weaknesses of the discipline in the participating universities?
- Where the data appear at odds with their knowledge of the discipline – are there any factors that immediately spring to mind that might explain this?
- Where the data reinforce their assessment – which measures are the most robust?
- Does the data provide useful additional information that could assist RQF panellists in assessing the field?

Results

Each workshop was provided with data for the three measures specified in the methodology. As discussions focussed almost entirely on just two of these – citation rates based on either all publications, or restricted to ISI journals only – only data relating to those two measures are presented in this paper.

Political Science

Table 1 shows the data for all publications in Political Science departments, and data for the same departments restricted to articles in ISI-indexed journals. In addition to publication numbers, and citation totals, a citation per publication rate is calculated (cpp) and universities are ranked on the basis of their cpp rates.

Table 1. Political Science citation analysis, 2000-2005

University Department	Limited to articles in ISI journals				All publications – books, chapters, articles			
	No. Pubs	No. Cites	cpp	rank	No. Pubs	No. Cites	cpp	rank
<i>Dept A</i>	90	398	4.42	2	585	887	1.52	1
<i>Dept E</i>	8	19	2.38	10	102	140	1.37	2
<i>Dept M</i>	35	83	2.37	11	294	339	1.15	3
<i>Dept I*</i>			2.96	7			0.98	4
<i>Dept D</i>	13	47	3.62	4	170	158	0.93	5
<i>Dept H</i>	7	23	3.29	6	139	114	0.82	6
<i>Dept K</i>	35	95	2.71	8	385	271	0.70	7
<i>Dept C*</i>			3.40	5			0.66	8
<i>Dept F</i>	2	4	2.00	12	58	32	0.55	9
<i>Dept N</i>	9	23	2.56	9	118	52	0.44	10
<i>Dept J</i>	8	12	1.50	14	107	41	0.38	11
<i>Dept G</i>	9	15	1.67	13	227	85	0.37	12
<i>Dept L*</i>					27	10	0.37	12
<i>Dept P</i>	1	5	5.00	1	42	14	0.33	13
<i>Dept B</i>	1	4	4.00	3	16	5	0.31	14
<i>Dept O</i>	2	3	1.50	14	42	13	0.31	14

* Note: Three departments were unable to provide details of their 2000 publications. This had a deflationary effect on citation per publication rates for the period and significantly affected their performance. To provide a more realistic measure of their relative standing in the discipline, estimates were made of the effect of the missing publications, based on the distribution of citations across the six year window for those departments that reported the full range of years.

Undertaking a standard bibliometric analysis, with data limited to articles in ISI-indexed journals, produced results that were of little value for an RQF-style process. Some university departments had no publications in the ISI literature for the period. Van Raan has proposed ten or 20 publications per year — the usual output of a research group in the sciences — as a sufficient basis for bibliometric calculations while rejecting those based on a few publications per year (Van Raan 2000). Using this criteria, only one department (Dept A) had sufficient output for analysis.

Incorporating non-ISI publications into the analysis resulted in a ten-fold increase in the number of publications, and a three-fold increase in the number of citations, adding significantly to the apparent robustness of the rankings. Now only four departments (Depts L, P, B and O) failed to satisfy van Raan's criteria. It overcame the absurd result which saw Dept P ranked at the top of the ISI analysis on the basis of a single publication. Dept E, which was ranked second when all publications were considered, was only in tenth position when the analysis was restricted to ISI journal articles. Dept A was the only institution with a reasonable number of ISI publications, and it ranked highly in both analyses.

The visibility of various publication types in the WoS was analysed, and is depicted in Figure 2.

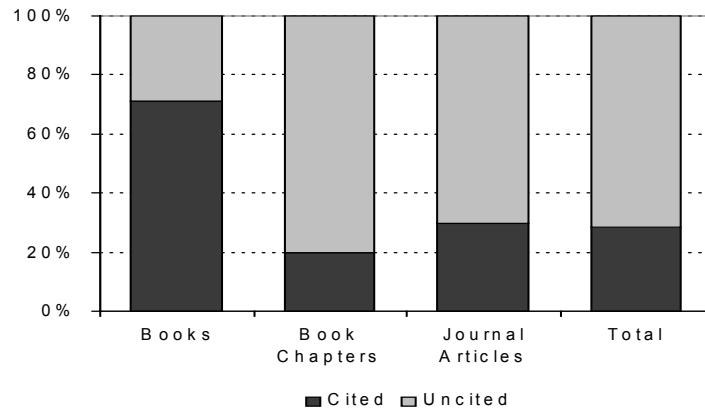


Figure 2. Percentage of publications cited by type, Political Science

The most important feature of Figure 2 is the proportion of books that are visible in the WoS (i.e. are cited in the indexed journal literature). This is particularly significant for a discipline that attaches great importance to the research monograph (AHRC ref?).

The political scientists rejected outright the ranking of universities on the basis of ISI publications only. They commented that the ranking did not accord with their knowledge of relative performance, and also voiced concerns about the small number of publications on which the assessment was based. They did however observe that the analysis based on the extended publication set concurred with their 'picture' of the relative strengths and weaknesses of the departments participating in the pilot study. Dept A was the acknowledged leader in the discipline, and Depts E and M the other two strongest departments. The ranks given to departments were deemed of little value in themselves, suffering from the problems identified by van Raan of creating quality difference where the data might reveal only marginal differences in the absolute numbers (van Raan 2005). However the relative performances based on the cpp rates were judged informative.

History

Historians participating in the workshop for their discipline were given similar data for perusal. Table 2 shows the data for all publications in history departments, and data for the same departments restricted to articles in ISI-indexed journals. In addition to publication numbers, and citation totals, a citation per publication rate is calculated (cpp) and universities are ranked on the basis of their cpp rates.

Table 2. History citation analysis, 2000-2005

University Department	Limited to articles in ISI journals				All publications – books, chapters, articles			
	No. Pubs	No Cites	cpp	rank	No. Pubs	No. Cites	cpp	rank
<i>Dept VII</i>	5	5	1.00	12	232	199	0.86	1
<i>Dept IV</i>	8	33	4.13	3	66	55	0.83	2
<i>Dept XV</i>	11	29	2.64	6	183	152	0.83	2
<i>Dept I</i>	26	82	3.15	5	443	364	0.82	3
<i>Dept V</i>	13	46	3.54	4	312	225	0.72	4
<i>Dept XI*</i>			1.54	10			0.60	5
<i>Dept XVI</i>	6	6	1.00	12	139	79	0.57	6
<i>Dept III*</i>			1.80	8			0.55	7
<i>Dept VI</i>	7	9	1.29	11	220	104	0.47	8
<i>Dept XII</i>	3	3	1.00	12	159	73	0.46	9
<i>Dept XVIII</i>	1	12	12.00	1	41	19	0.46	9
<i>Dept XVII</i>	1	9	9.00	2	37	15	0.41	10
<i>Dept XIV*</i>					27	10	0.37	11
<i>Dept XIII</i>	7	11	1.57	9	305	107	0.35	12
<i>Dept VIII</i>	5	5	1.00	12	164	51	0.31	13
<i>Dept IX</i>	1	2	2.00	7	30	7	0.23	14
<i>Dept II</i>	1	1	1.00	12	29	5	0.17	15
<i>Dept X</i>					18	3	0.17	15

* See note to Table 1.

The total number of publications for the history analysis was slightly larger than that for political science, yet limiting the data to articles in ISI journals restricted the analysis to a mere 106 publications. Only three universities reached double figures for the six year period, which rendered an analysis using standard bibliometric techniques untenable. The university that featured most strongly when all publications were included in the analysis (Dept VII) ranked only twelfth with just 5 publications in the more restricted analysis. Adding the additional publication types to the analysis had an even greater effect in history than it did in political science, with a six-fold increase in citation counts.

History departments did not demonstrate the same degree of discrimination in citation performance at the top end as their political science colleagues. Depts VII, IV, XV and I were all at a similar level. But as with their colleagues, the historians declared that this aligned with their knowledge of relative strengths and weaknesses in their discipline — those four departments were the ones they expected to see at the top of any assessment of research in their field.

The degree of visibility of the publication outlet they considered the most important in their discipline, the research monograph, was of particular interest to the historians. The proportion of each type of output that attracted at least one citation in the WoS is shown in Figure 3.

The historians, like the political scientists, rejected measures based solely on ISI journal articles for identical reasons. However the visibility of books in the approach increased the confidence of this initially-sceptical group in the new measures proposed.

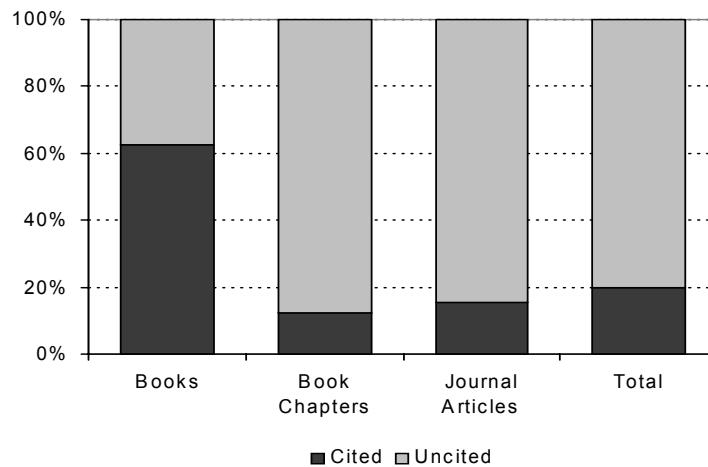


Figure 3. Percentage of publications cited by type, History.

Expert assessment of measures

The discussions for the two working groups were kept separate in the belief that the two disciplines might arrive at different positions regarding the potential use of these enhanced bibliometric measures. Surprisingly, there was almost total concurrence of their ideas and recommendations. Foremost among these was that standard bibliometrics are not considered appropriate measures of quality in the disciplines of political science and history because they capture less than 20% of output. Both panels noted that the institutional rankings produced by examining ISI-only data were inconsistent with their perceptions of research quality around the country.

Panellists agreed that the data painted a much more accurate picture when expanded to include books and book chapters. They added that such analysis would be more meaningful if accompanied by measures of staff productivity (publications per effective full time staff member) and employment status (seniority levels being particularly informative information). It would be enhanced even further if the distribution of citations across the members of the department being assessed was shown, demonstrating whether overall citations rates were determined by one high-impact researcher, or whether the impact was distributed more evenly across all staff.

Both panels stressed the importance of the role given to bibliometrics data in reaching their conclusions. They argued that such measures were only acceptable in the RQF if they served to inform peer/expert review, rather than being used to replace it. Their value would be further enhanced if an authoritative group could annotate the data before they were referred to RQF expert panels, to note any particular characteristics of a group's research portfolio that might influence its citation performance. They also recommended that in the social sciences and humanities, the assessment period should be expanded to seven years, allowing for a greater capture of citations and more robust data.

Any bibliometrics exercise needs to be mindful of a structural bias against local projects. Both panels noted that citations counts for writers on international topics were generally higher than similarly well-regarded scholars working on Australian and Pacific topics. They believed a similar effect occurs in other disciplines within the social sciences and humanities. This is potentially a very serious problem for an Australian national research funding system. It raises the possibility that Australian universities might be encouraged to run down their research strengths in locally relevant studies across the board.

They recommended that DEST commission further research into publishing and citation trends across publications not indexed by the ISI, which represent very high proportions of DEST-recognised scholarly publications in the disciplines of political science and history at Australian universities. Such

research should closely examine how citation patterns from books and book chapters compare to citation patterns from journal articles (see Cronin et al. 1997 for a study of this issue).

Members of the History group expressed a concern that citation counting is blind as to the ‘quality’ of a given citation. Thus an in-depth discussion of a particular piece of scholarship rates the same as a citation-only listing of it. It is hard to imagine how this limitation could be overcome, practically speaking, although it reinforces the imperative that these not be used as stand-alone measures.

Conclusions

The overall positive reception of these novel bibliometric measures by both historians and political scientists surprised the researchers undertaking this analysis. In trying to determine the reason, a number of factors were identified. Given the makeup of the RQF expert panels, participants were not confident that they would contain sufficient expertise to assess their whole disciplines. With only 12 to 15 members on panels that encompass many disciplines, both working groups anticipated they would be represented by no more than two experts from their own field. They desired additional, objective data that would, in the words of Paul Bourke, ‘trigger the recognition of anomalies’ (Bourke et al. 1999). Where the quantitative measures and the peer assessment agreed, there would be added confidence in the outcomes. Where they differed, further examination would be required to determine whether discipline characteristics rendered the data inaccurate, or whether gaps existed in the knowledge of the experts.

Another concern, also due to the relatively light representation of their discipline on the expert panels, related to the potential for reviewer bias. Where a large group of experts in a discipline is assembled, as in the case of the UK RAE, the potential for individual prejudices to sway assessments is minimised. Participants believed there was a real danger of this occurring in the Australian RQF when only one or two panellists represented their discipline. Several anecdotes purporting to detail recent concerns with peer assessment of Australian research were discussed at length.

In both sessions, participants responded that the data corresponded to their assessment of the relative strengths and weaknesses of the institutions, however in coming to this conclusion they had made specific assumptions in interpreting the data. This related particularly to their knowledge of the focus of research in each department, and the likely impact of this on the citation data. Specifically, they took into account research in areas likely to return low citation rates relative to other research of equal standing (e.g. research focussing on Australian or Pacific issues), and the proportion of early career researchers in the departments.

Their recommendations mirrored both their support for the measures, and their concerns that the data needed to be interpreted with sensitivity. Their support does not mean automatic inclusion of the measures in the RQF. A number of additional hurdles need to be overcome – a wider understanding and acceptance of the measures throughout the sector; solving the significant IT challenges that such measures present; and extending the analysis to additional disciplines. These novel measures proved informative in history and political science, but they may be less appropriate for other social science and humanities disciplines. While it has already been demonstrated that it is possible to extract citations to a large number of non-source publications (Butler and Visser, 2006), the magnitude of the effort required should not be underestimated, particularly when the RQF timelines are very tight.

References

- Bourke , P., Butler, L. & Biglia, B. (1999). A Bibliometric Analysis of Biological Sciences Research in Australia. DETYA No. 6307HERC99A. Commonwealth of Qustralia: Department of Education, Training and Youth Affairs, Higher Education Division.
- Butler, L. & Visser, M. (2006). Extending citation analysis to non-source items. *Scientometrics*, 66(2), 327-343.
- Council for Humanities, Arts and Social Sciences (CHASS). (2005). *Measures of quality and impact of publicly funded research in the humanities, arts and social sciences*. Retrieved November 29, 2006 from : <http://www.chass.org.au/op2.pdf>.

- Cronin, B., Snyder H. & Atkins, H. (1997). Comparative citation rankings of authors in monographic and journal literature: a study of sociology. *Journal of Documentation*, 53(3): 263-273.
- Department of Education, Science and Training (DEST). (2006a). *Research Quality Framework: Assessing the quality and impact of research in Australia*. Retrieved 5 March, 2007 from: http://www.dest.gov.au/NR/rdonlyres/7E5FDEBD-3663-4144-8FBE-AE5E6EE47D29/14867/Recommended_RQF_Dec2006.pdf
- Department of Education, Science and Training (DEST). (2006b). *Research Quality Framework: Assessing the quality and impact of research in Australia: Quality Metrics*. Retrieved 5 March, 2007 from: http://www.dest.gov.au/NR/rdonlyres/EC11695D-B59D-4879-A84D-87004AA22FD2/14099/rqf_quality_metrics.pdf.
- Moed, H.F. (2005). *Citation Analysis in Research Evaluation*. Netherlands: Springer.
- Van Leeuwen, T.N. (2006). The application of bibliometric analyses in the evaluation of social science research. Who benefits from it, and why it is still feasible. *Scientometrics*, 66(1): 133-154.
- Van Raan, A.F.J. (2000). The Pandora's Box of Citation Analysis: Measuring Scientific Excellence – The Last Evil? In B. Cronin and H. Barsky Atkins (eds.), *The Web of Knowledge* (pp. 301-319). Medford, New Jersey: Information Today, Inc.
- Van Raan, A.F.J. (2005). Fatal attraction: conceptual and methodological problems in the ranking of universities by bibliometric measures. *Scientometrics*, 62(1): 133-143.

Combining Mapping and Citation Network Analysis for a Better Understanding of the Scientific Development: The case of the Absorptive Capacity Field.

Clara Calero Medina and Ed C.M. Noyons

clara@cwts.nl, noyons@cwts.nl

Centre for Science and Technology Studies (CWTS), University of Leiden,
Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden (The Netherlands)

Abstract

The general aim of this paper is to show the preliminary results of a study where we are combining bibliometric map and citation network analysis to better understand the process of creation and transfer of knowledge through scientific publications. The novelty of this approach is the combination of both methods. The bibliometric co-word map will provide insight into the content of the publication. This will be used to the interpretation of groups of citations that may constitute the backbones of a research tradition or the future of the research.

Keywords

citation network analysis; bibliometric mapping; main research stream; emerging terms.

Introduction

In every scientific field there are key concepts that set the base for theoretical developments through the years. The objective of our study is analyzing the influence of the introduction of a new concept on a research field through the analysis of scientific publications. The purpose is being able to answer the following questions:

1. What and how is the diffusion rate of the concept through the literature?
2. Which are the terms and theories associated to the concept?
3. Which papers and theories are considered the main research streams of the field?
4. Which are the emerging terms? Which are their applications and impact on the research field? What is their degree of acceptance in the field?

The novelty of our approach is that to answer to these questions we will combine bibliometric mapping and citation network analysis. The bibliometric co-word map will provide insight into the content of the publication. This will be used to the interpretation of groups of citations that may constitute the backbones of a research tradition or the future of the research.

Data and Methods

Data

In the organization field the concept “Absorptive Capacity” is considered as one of the most important introduced in the last fifteen years. The paper published in 1990 by Cohen & Levinthal is generally accepted as the founding paper. This influential paper has received more than 1,500 citations (up to now) and, as Figure 1 shows, the attention is growing. Recently there have been two main efforts for reviewing the absorption of Absorptive Capacity notion in the literature of Organizational Theories (Foss, Lyles & Volberda and Lane, Koka & Pathak (2006)). Some of these experts are involved on the validation of the results of our study.

The data set are the 1213 publications citing Cohen and Levinthal (1990) up to 2005. The publications were extracted from journals processed by Thomson Scientific for the Web of Science. From now on we will name this set of publications as the Absorptive Capacity field.

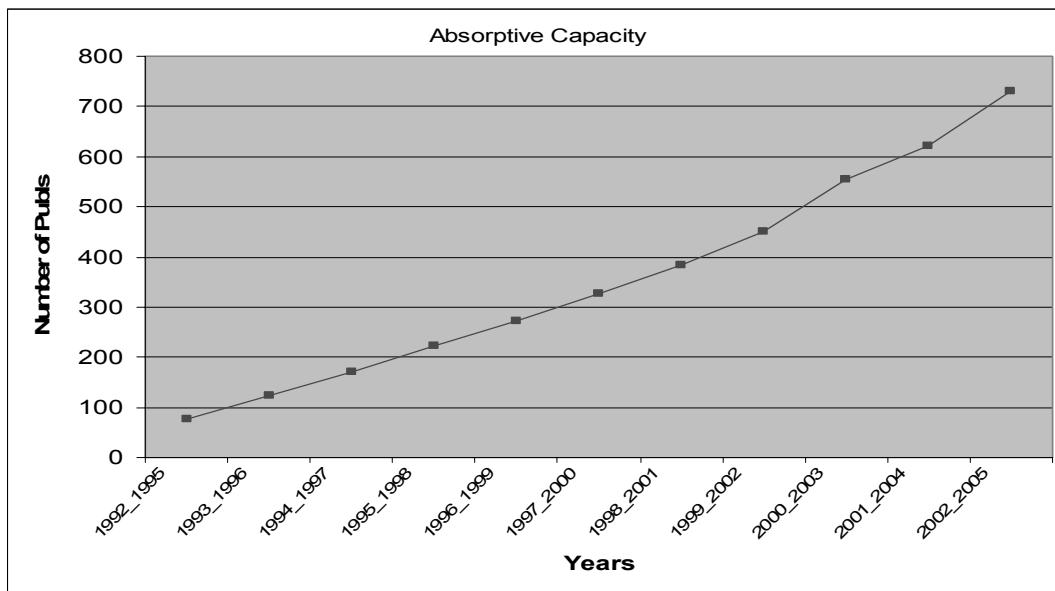


Figure 1. Number of Publications citing Cohen & Levinthal (1990) during the period 1992-2005
(Source: Web of Science)

Bibliometric Map Analysis

The first step is to get a general overview of the Absorptive Capacity field. We map the structure of all publications citing Cohen and Levinthal (1990) by applying a co-occurrence bibliometric mapping method. With this method we create a 2-dimensional graph with sub-domains representing topic clusters. The distances between sub-domains represent their mutual cognitive similarity. The closer they are in the map, the more similar. The topic clusters are created by applying a co-word analysis to the keywords in the citing publications (Noyons, 1999).

This part of the analysis is just giving us a first overview of the field. We can identify the sub-domains that attract more publications and their growth rate in terms of number of publications over the period. But we don't know anything about the publications behind these sub-domains.

Publication content labeling

Next was to label each of the 1213 publications citing Cohen & Levinthal (1990) with the sub-domain or sub-domains where they belong. Publications may be represented in more than one sub-domain. In this way we intend to give a theoretical content to the publication.

Citation Network Analysis

After classifying each article with the sub-domain/s, we create a citation network based on the citation ties that exist among the 1213 papers. The citation network analysis began with the study by Garfield, Sher & Torpie (1964) of Asimov's history of DNA. The analysis showed that there is "a high degree of coincidence between an historian's account of events and the citational relationship between these events". We carry out a citation network analysis to study the processes of the diffusion of the concept of Absorptive Capacity and the theories around it. The citation analysis allows us to view the structure of part of the Absorptive Capacity literature that has emerged from current citation practices and shows how this emergent structure elevates certain viewpoints and approaches and marginal others. In this context, following Small (1978), a citation stands for a concept. A citation often cited may be seen as a "concept symbol" that represents an author's orientation to a community of scientists or an approach to a topic (Moed, (2005)).

Main Path

If knowledge flows through citations, a citation that is needed in paths between many articles is more important than a citation that is barely needed for linking articles. The most important citations constitute one or more main paths, which are the backbones of a research tradition (Hummon &

Doreian (1989, 1990), Hummon & Carley (1993), Batagelj (2003) and De Nooy, Mrvar & Batagelj (2005)). The main path analysis presents a longitudinal perspective on how a research field has progressed according to its citation patterns. In our case, and this is the novelty of our approach, the link of the publications with the sub-domains will allow us to follow through the time the spreading of the theories and concepts. The main path analysis is conducted with the software package PAJEK.

For extracting from the citation network the main path we compute the traversal weights. The traversal weight measure the number of times that a tie or link between articles is involved in connecting other articles in a citation network. Second, starting from the Cohen and Levinthal (1990) paper, the main path algorithm chooses the next link in the path as the outgoing link with the highest traversal weight. By repeatedly applying this choice rule, we define a path through the network that follows a structurally determined most used path. The main path, chosen on the basis of the most used path will identify the main stream of the Absorptive Capacity literature.

Emerging publications

We have identified the publications that define the main stream of the field. The next goal is to identify the main publications that are introducing new concepts on the field. These publications do not belong to the main stream of the field because they are too recent (no time to take part of the main theories) or because they are just not related with the main core of the field.

At the moment we are working on this part of the analysis. We will make a separate analysis of these publications focused on the most recent years. We will identify the set of publications related with the small fast growth sub-domains from the bibliometric map. We will work not only the traversal weights but other centrality measures from network theory. Additionally we may consider the citations out of the field because if the fast growing sub-domain is also positioned far away from the others may indicate 'imported' concepts from other fields.

Preliminary Results

The Absorptive Capacity Bibliometric Map

Figure 2 shows the map of the Absorptive Capacity Field. It is the result of grouping the keywords into clusters (sub-domains) and maps those sub-domains in a two-dimensional figure, with the size of each sub-domain indicating the number of publications represented and the colour of each sub-domain indicating the growth in the number of publications until 2005 (black: fast growth; grey: growth around average; white: growth below average). Sub-domains that are closer to one another co-occur more often than sub-domains that are further apart.

Most of the studies in Absorptive Capacity are focused on R&D rates in various industries (sub-domain 9), inter-organizational and managerial antecedents (sub-domain 1 and 2). Fast growth areas of Absorptive Capacity (the black circles) seem to be studies on knowledge flows and dynamic capabilities (sub-domain 6), the impact of Absorptive Capacity on technological innovation and firm performance (sub-domain 10), and the effects of relational (trust) versus formal governance modes (sub-domain 7) on Absorptive Capacity. Figure 2 also shows that organizational innovation (sub-domain 11) and realized Absorptive Capacity (sub-domain 3) have been underrepresented.

Citation Network Analysis

Main Path

The main path (Figure 3) shows the main track followed by the researchers in this field to explain the firm's innovative processes. The nodes (circles) of the graph represent the publications, the lay out is ordered in time from top to bottom (from 1990 until 2003), the colors (grey scale) represent the year of the publication and the thickness of the lines relate to the traversal weights. The publications are labeled with the first author's name, publication year and between parenthesis the number of their sub-domain/s.

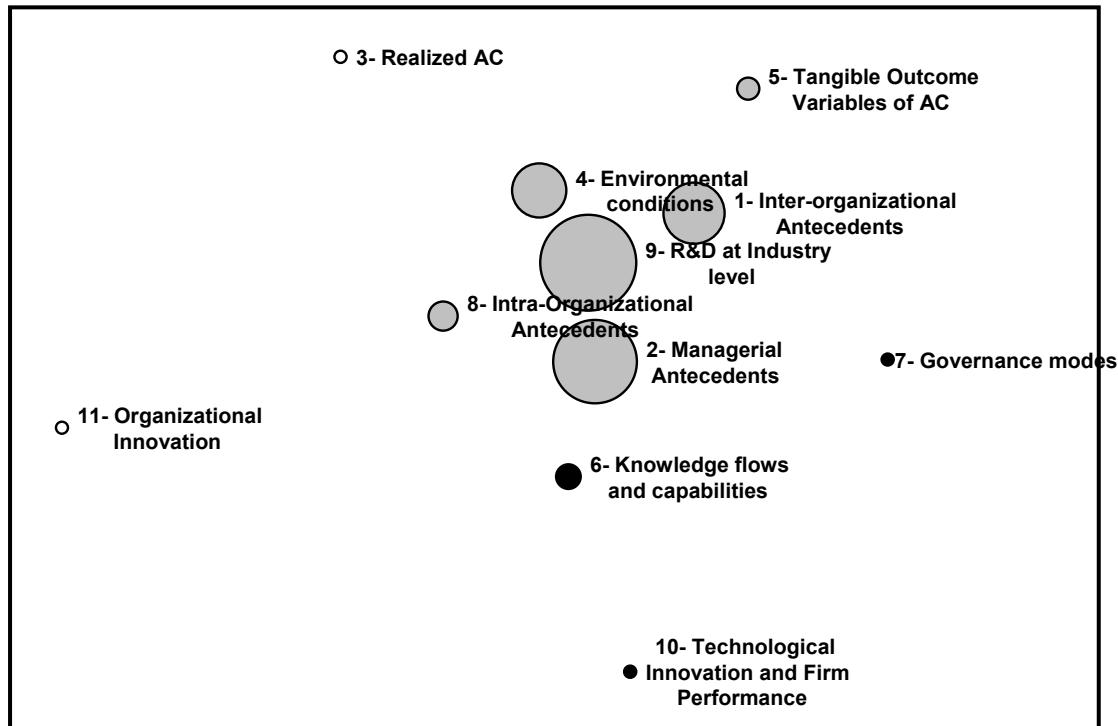


Figure 2. Bibliometric map of the field of Absorptive Capacity.

Figure 3 illustrates that there are just a few publications that constitute the main stream on the Absorptive Capacity literature. As we can see on the graph, these papers are very focused on the main sub-domains of the map (1, 2, 4, 9). However, starting with the paper from Zahra SA (2002), sub-domain 5 becomes part of the main path. This means that this small sub-domain during the last years is starting to take part of the main research stream of the field.

Emerging Publications

As we mentioned above at the moment we are working on the algorithm to detect the emerging publications. The interest is on publications from the small fast growing sub-domains 6, 7 and 10. Special attention will be paid to the publications from sub-domain 10 as an emerging and far away sub-domain.

Concluding remarks and follow-up research

We think that these preliminary results show a lot of potential for going a step further in unraveling the pattern behind a set of publications that represent a field. These first results have been already validated by experts of this particular field. Combining their comments with our own research interest we are going to keep working on this direction during the following months.

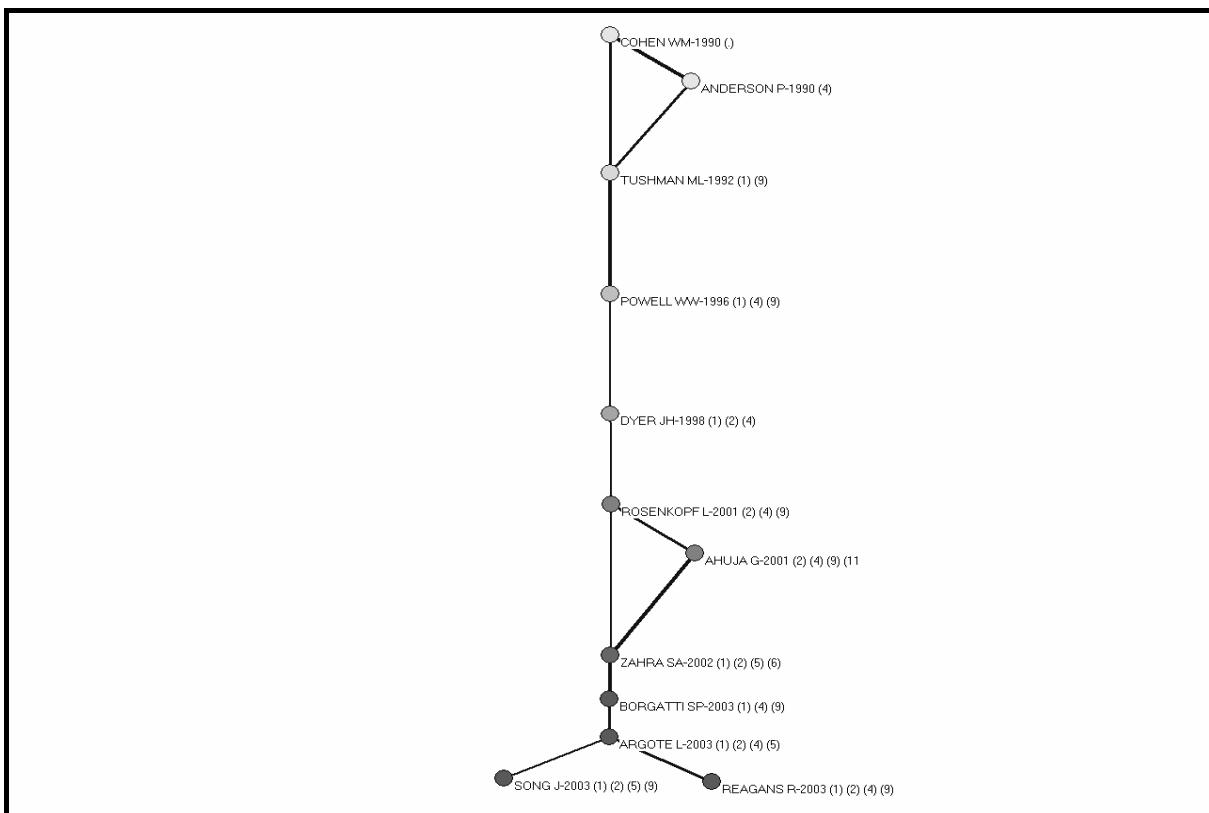


Figure 3. Main Path of the Absorptive Capacity Field

References

- Asimov, I. (1963). The Genetic Code. New York: New American Library.
- Batagelj, V. (2003). Efficient Algorithms for Citation Network Analysis. Preprint Series. Univ. Ljubljana, Inst. Math., 41 (897), 1-29.
- Batagelj, V. & Mrvar, A. (2006). Pajek: Program Package for Large Network Analysis, University of Ljubljana, Slovenia. <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>.
- Cohen, W.M. & Levinthal, D.A. (1990). Absorptive Capacity – A New Perspective on Learning and Innovation. *Administrative Science Quarterly*, 35 (1), 128-152.
- De Nooy, W., Mrvar, A. & Batagelj, V. (2005). Exploratory Social Network Analysis with Pajek. New York: Cambridge University Press.
- Foss, N.J., Lyles, M.A. & Volberda H.W. Absorbing the Concept of Absorptive Capacity: How to Realize its Potential in the Organization Field. To be published in *Organization Science*.
- Garfield, E., Sher, I.H. & Torpie, R.J. (1964). The Use of Citation Data in Writing the History of Science. Philadelphia: Institute for Scientific Information.
- Hummon, N. & Carley, K. (1993). Social networks as normal science. *Social Networks*, 15, 71–106.
- Hummon, N. & Doreian, P. (1989). Connectivity in a citation network: the development of DNA theory. *Social Networks*, 11, 39–63.
- Hummon, N. & Doreian P. (1990). Computational methods for social network analysis. *Social Networks*, 12, 273–88.
- Lane, P.J., Koka, B.R. & Pathak, S. (2006). The reification of absorptive capacity: A critical review and rejuvenation of the construct. *Academy of Management Review*, 31 (4), 833-863.
- Moed, H. F. (2005). Citation Analysis in Research Evaluation. Dordrecht: Springer.
- Noyons, E.C.M. (1999). Bliometric Mapping as a science policy and research management tool. Thesis Leiden University. Leiden: DSWO Press.
- Small, H.G. (1978). Cited documents as concept symbols. *Social Studies of Science*, 8, 327-340

A Research Productivity Index to Account for Different Scientific Disciplines

Mônica G. Campiteli, Pablo D. Batista and Alexandre S. Martinez

mocampiteli@gmail.com, pablosfisica@yahoo.com, asmartinez@ffclrp.usp.br

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto – Universidade de São Paulo
Av. Bandeirantes, 390 14040-901 Ribeirão Preto – SP (Brazil)

Abstract

The number h of papers with at least h citations has been proposed to evaluate individual's scientific research production. This index is robust in several ways but yet strongly dependent on the research field. We propose a complementary index $h_I = h^2/N_a^{(T)}$, with $N_a^{(T)}$ being the total number of authors in the considered h papers. A researcher with index h_I has h_I papers with at least h_I citation if he/she had published alone. We have obtained the rank plots of h and h_I for seven German scientific communities. Contrasting to the h -index curve, the h_I index present a perfect data collapse into a unique curve allowing comparison among different research areas.

Keywords

scientometrics; h-index; co-authorship

Introduction

New proposals for the scientific research output evaluation have been suggested recently (Hirsch, 2005, Batista *et al*, 2006, Popov, 2006, Taber, 2005). In particular, Hirsch (2005) has proposed a new scalar index h to quantify individual's scientific research output. A researcher with index h has h papers with at least h citations. This index has several advantages: (i) it combines productivity with impact, (ii) is automatically calculated in Thompson ISI Web of Science database, (iii) it is not sensitive to extreme values, (iv) it is hard to inflate, (v) automatically samples the most relevant papers concerning citations, etc. This index is related to extremal statistics, which is dominated by exponential density distributions, meaning that high h values are difficult to achieve.

Among the constraints of the h index, the sensitiveness to the research field remains a concern. The maximum h values for the different disciplines are very distinct disabling comparisons among scientists of distinct areas. Even inside a given discipline, say Theoretical and High Energy Physics, it would be hard to compare scientific research output. The idealization of an index that could account for such comparisons is of major importance.

In recent papers, it has been shown that the number of citations a paper receives can be influenced by the number of authors (Glanzel & Thijs, 2004). Since: (i) the greater the number of authors, the greater the number of self-citations and (ii) the co-authorship behavior is characteristic of each discipline, we have proposed a complementary index h_I to quantify an individual's scientific research output valid across disciplines. The statistics of h and h_I are presented for the fundamental research fields in Germany. Contrasting to h rank plots, we have shown that the relative h_I rank plots collapse into a single unique curve. This universal behavior suggests that it could be used to compare scientific research output performance in different research fields.

Methodology

From Thomson ISI Web of Science database, we have considered the German scientific research output in seven different fields: Physics, Chemistry, Biology/Biomedical, Mathematics, Engineering, Humanities and Medicine. The database has been compiled from the database of the Institute for Scientific Information (ISI). The search has been conducted using the query “Germany OR Deutschland” in the address field. This means that it has been accounted all the documents with at least one German address with citations till October 2005. Researcher nationality and researches done by Germans abroad (foreigner address) are disregarded in the considered database. We have considered all documents published from 1945 to 2004. The search has been performed separately for each year.

Our database contains information of about 1,215,059 bibliographical references and a total of 11,808,445 citations. The database includes type of publication, citations received, authors' names and addresses, including the institutions, cities, states and country. For the statistical analysis it is reasonable to consider only Articles, Meeting Abstracts, Reviews and Notes equivalent to 85.24% of the documents and 92.30% of citations. Documents have been classified into the research fields using the tag subject. Seven lists have been compiled containing author name, publication number, times cited and number of authors. Notice that a given researcher can appear in more than one list.

It is worth mentioning that no corrections have been made about homonyms and the use of different publication names by some authors. This practice can inflate the rate of citations of some authors. However, here we are dealing with whole distributions and comparisons among different fields instead of absolute values and we assume that the same bias will be present with the same weight in each discipline. Thus, we believe that this negligence would not jeopardize the results and they remain reliable.

Results

Figure 1 shows the histogram of total number of citations and total number of papers per author for the seven different disciplines. One can notice that each discipline has its own pattern of scientific production and citations with Mathematics and Humanities being the disciplines with less productivity and less citations rate and Physics, Medicine and Biology the disciplines with the larger rates. Figure 2 corroborates this result showing the number of researchers with index h for each discipline.

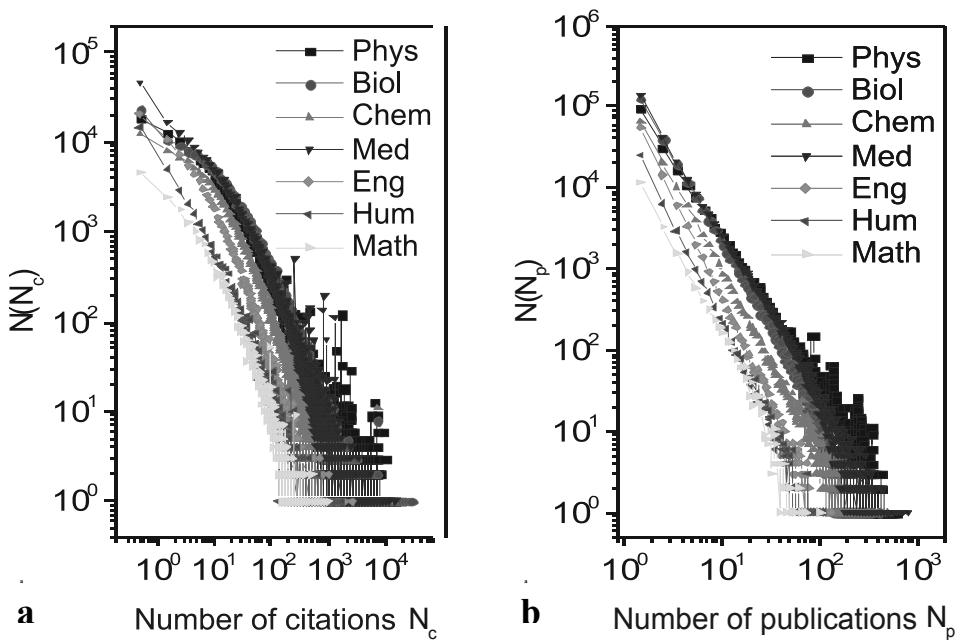


Figure 1. Histogram of number of citations per author (a) and number of publications per author (b).

These results are in great accordance with the results obtained from the Brazilian scientific community (Batista *et al*, 2006). The distributions of papers with k authors are shown in Figure 3. Again, each discipline has a specific pattern of co-authorship. One sees that the maximum of the distributions is at $k_{max} = 2$ for most of the disciplines except for Maths and Humanities whose publications are mostly single-author papers and Medicine that shifts its peak to four authors. Physics have several papers with more than 50 authors. These papers probably reflect collaborations with large international teams (Lehmann, Lautrup & Jackson, 2003).

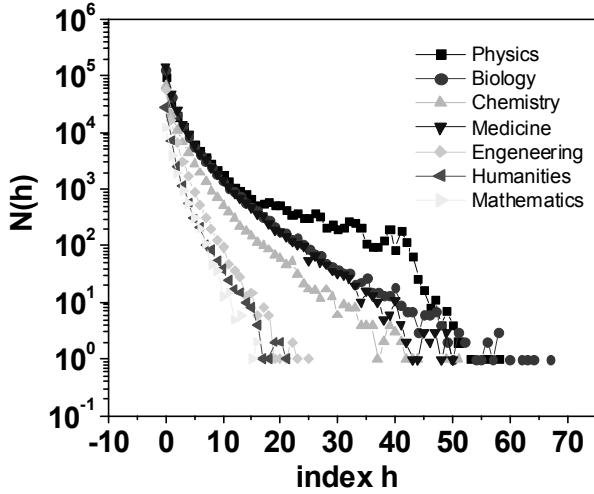


Figure 2. Number of researchers with h index in the seven different research fields investigated in Germany.

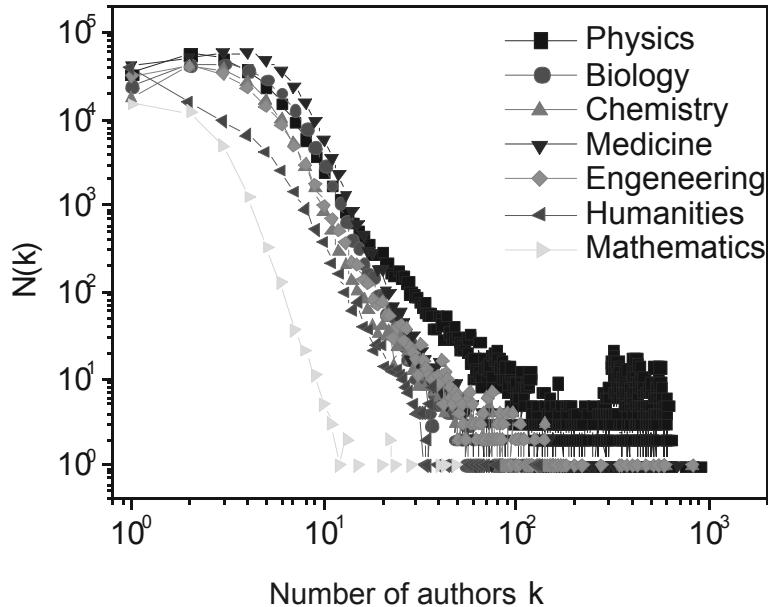


Figure 3. Number of publication with k authors per article in seven research fields in Germany.

From the data showed, one sees that it is very difficult to compare researchers from different fields. However, we have noticed a strong correlation between h and the number of authors that sign the top h publications (data not shown).

To account for the co-authorship effect, divide h by the mean number of researchers in the h publications $\langle N_a \rangle = N_a^{(T)} / h$, where $N_a^{(T)}$ is the total number of authors (author multiple occurrences are allowed) in the considered h papers. Thus, we obtain a new index:

$$h_I = h / \langle N_a \rangle = h^2 / N_a^{(T)} \quad (1)$$

which gives further information about the research output.

The rationale for this procedure is that we want to measure the effective individual average productivity. More authors could produce more future self-citations which may produce statistical biases. If a given researcher is the only author in his/her h papers, then $N_a^{(T)} = h$ and $h_I = h$. The h_I

index indicates the number of papers a researcher would have written alone along his/her carrier with at least h_I citations. Once h has been computed, the h_I index is also easy to compute from the Thompson ISI Web of Science. The rank plots of h (Fig. 4a) and h_I (Fig. 4b) are strongly different.

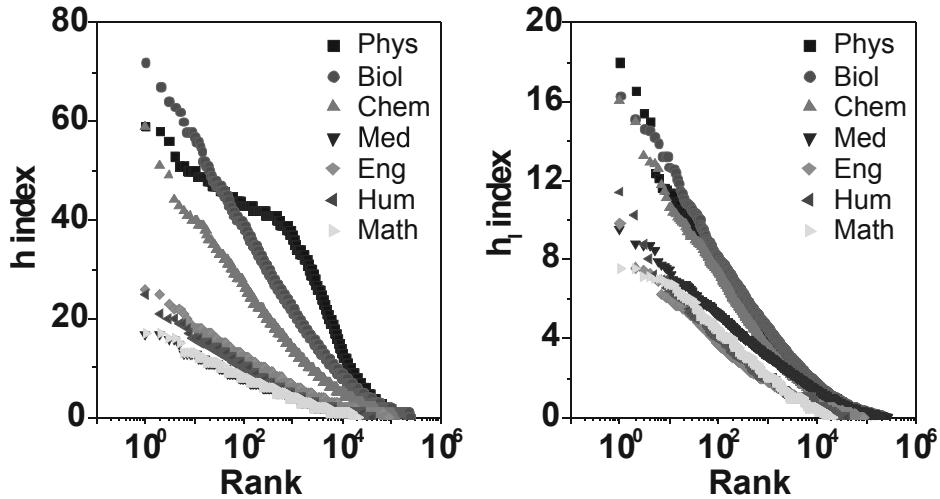


Figure 4. The index h (a) and h_I (b) as a function of the ranking R for the German research fields. The h_I curves, in contrast to h curves, have the same functional shape.

Figure 4 presents the h and h_I indices in a decreasing ranking. Physics rank plot is drastically different from the rank plot of other considered fields for the h index. The h_I rank plot is much smoother and, importantly, all the distributions are more similar, being close to stretched exponentials (straight line in the linear-log plot) (Laherrère & Sornette, 1998). The disciplines split in two well defined major groups in the h_I rank plot. The first one concerning Physics, Biology and Chemistry. Second one concerns Medicine, Engineering, Humanities and Mathematics. Notice that Medicine has dropped to the lower group after the normalization for author number because of its characteristics of a large mean number of authors per paper (Figure 3). The smoothness of the h_I curves displays the emergence of a universal behavior.

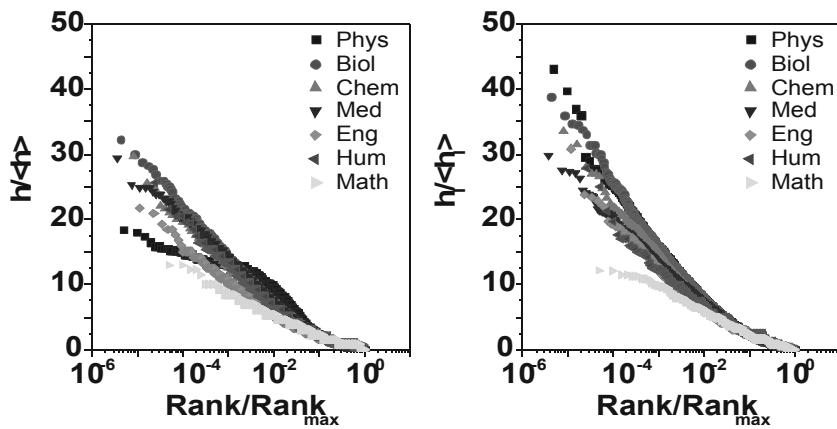


Figure 5. The indexes $h/\langle h \rangle$ (a) and $h_I/\langle h_I \rangle$ (b) as a function of the ranking R/R_{\max} in the different research fields in Germany.

(A single unique curve is found for index h_I permitting comparisons among different research fields. Data collapse is not obtained for h curves because of the co-authorship effects in Physics)

The functional similarity of the h_I rank plots has motivated us to scale the variables. With this aim, we have divided each h_I curve by its mean values and the ranks have been divided by the size of the community (maximum rank). The scaled variables are plotted in Figure 5, where the data collapse is

shown by a single unique curve. This universal curve is not observed for the relative h index (Fig.5b) since the co-authorship effects exclude Physics.

The use of the mean value in the definition of h_I index could penalize authors with eventual papers with large number of authors, since the mean is a measure very sensitive to extremum values. A possible correction to this factor is to consider the median or harmonic mean instead of the mean value. In fact, we have observed a strong correlation ($r = 0.93$) between the rankings using the mean value and median measures.

Conclusion

The index h_I is complementary to h and indicates the number of papers a researcher would have written along his/her carrier with at least h_I citations if he/she has worked alone. It diminishes the h degenerescency and has the advantage of being less sensitive to different research fields. This allows a less biased comparison due to the consideration of co-authorship effects. The h ranking studied takes into account publications that have at least one author with German address and presented strong differences in functional form between fields say, Physics and Humanities. Such differences are due to intrinsic differences in production and co-authorship behaviors among the disciplines and are softened for h_I , where data colapse has been found with the appropriate scaling. This universal behavior allows comparisons among different fields.

It worths mentioning that collaboration is not the only factor regarding the differences in the citation patterns among the disciplines. Further modifications of the h -index should be considered, such as publication periodicity and delays. It may also be interesting to perform this study for other countries and other instances as department evaluations and periodic publications.

References

- Batista, P.D. Campiteli, M. G. Kinouchi, O. and Martinez, A. S. *Is it possible to compare researchers with different scientific interests?*, Scientometrics 68 (1) 179-189 (2006).
- Hirsch, J. E. *An index to quantify an individual's scientific research output*, PNAS 102, 16569 - 16572 (2005).
- Glanzel, W. and Thijs, B. *Does co-authorship inflate the share of self-citations?*, Scientometrics 61, 395 - 404 (2004).
- Laherrère, J. and Sornette, D. *Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales*, Eur. Phys. J. B 2, 525-539 (1998).
- Lehmann, S. Lautrup, B. and Jackson, A. D. *Citation networks in high energy physics*, Phys. Rev. E 68, 026113 (2003).
- Popov, S. B. *A parameter to quantify dynamics of a researcher's scientific activity*, physics/0508113. Taber, D. F. *Quantifying publication impact*, Science 309, 4 (2005).

Trends in Conceptual Modeling: Citation Analysis of the ER Conference Papers (1979-2005)

Chaomei Chen, Il-Yeol Song and Weizhong Zhu

chaomei.chen@cis.drexel.edu, song@drexel.edu, wz32@ischool.drexel.edu

College of Information Science and Technology, Drexel University, Philadelphia, PA 19104-2875 (USA)

Abstract

We analyze thematic trends and challenging issues in conceptual modeling based on the metadata of 943 research papers published in a series of conferences on conceptual modeling (known as the ER conferences) between 1979 and 2005. We specifically address 1) all-time prominent challenges in conceptual modeling, 2) current challenges and emerging trends, and 3) the structure and dynamics of the conceptual modeling community. We utilize CiteSpace, a progressive domain visualization tool, to identify and visualize the movement of research fronts and intellectual bases, persistent clusters of papers, critical paths connecting these clusters, and the evolution of co-authorship networks as well as citation networks. The work contributes an in-depth analysis of a major forum of conceptual modeling and a practical method that one can use as frequently as needed to keep abreast of the state of the art of conceptual modelling.

Keywords

progressive knowledge domain visualization; citeSpace

Introduction

Conceptual modeling represents real-world phenomenon using semantic primitives. It is a basis for understanding the phenomenon and for creating a requirement specification that can be incorporated into an information system. Historically, conceptual modeling served as a framework for database design, database integration, visual query paradigms, and information system development.

The output of conceptual modeling is a conceptual model. The most influential conceptual model in the database community is the entity-relationship (ER) model (Chen, 1976). Since the introduction of the ER model, the original ER has been extended to achieve more semantic powers. Many versions of extended ER models have been developed and widely researched such as the Extended ER model (Teorey et al., 1986), the E²R model (Embley & Ling, 1989), the hierarchical ER model (Thalheim, 2000), Temporal ER models (Gregersen & Jensen, 1999), and the starER model (Tryfona et al., 1999).

ER conferences refer to a series of conferences on conceptual modeling, namely the International Conference on Conceptual Modeling and its predecessor – the International Conference on Entity-Relationship. This series started in 1979. Up to 1985, the conferences were held every alternative year, and then became annual conferences from 1986. The ER proceedings collect several types of papers: regular research papers, industrial papers, abstracts from keynote speeches or tutorials, a panel statement, and editorials. In this study, our analysis includes all the research papers in the proceedings. Several workshops are also held with the main ER conferences, but we have not included any paper from the ER workshops.

The online computer science bibliography DBLP¹ contains more than 700,000 bibliographic records in major computer science subjects, including the ACM SIGMOD Anthology. We retrieved bibliographic records of ER conference papers from DBLP's metadata². In DBLP, citation information is available for some of the records, but not for all of them. Abstracts are not available in the DBLP data. In particular, citation data are available only between 1979 and 1999.

¹ <http://www.informatik.uni-trier.de/~ley/db/>

² <http://dblp.luni-trier.de/xml/>

Since DBLP records after 2000 do not contain citation links, we retrieved citation data between 2000 and 2005 from the Web of Science³ (WoS) instead, including Science Citation Index (SCI), Social Science Citation Index (SSCI), or Art & Humanities (A&H). A bibliographic record in WoS contains fields such as author, title, and abstract. It also has a cited reference field (CR), which contains references cited by the corresponding paper. However, it appears that WoS does not have any records for the ER 2001 conference. As a result, our dataset includes citations made by all ER conference papers except the ones cited by the ER 2001 conference.

For the citation analysis, we used CiteSpace (Chen, 2004; Chen, 2006). CiteSpace is a Java application for analyzing and visualizing co-citation networks. Its primary goal is to facilitate the analysis of emerging trends in a knowledge domain. It allows the user to take a time series of snapshots of a domain and subsequently merge these snapshots. The initial version of CiteSpace was used to reveal turning points in superstring revolutions in physics. However, several issues remained unresolved when we implemented the first version of CiteSpace. The most distinctive new feature is the combination of computational metrics and visual attributes of pivotal points. The motivation is to substantially reduce the user's cognitive burden as they search for pivotal points in a knowledge structure.

In CiteSpace, users can identify pivotal points by visually scanning a visualized network for nodes that connect different clusters. One of the advantages of this approach is that no additional computing is required. CiteSpace also allows users to identify pivotal points in terms of high betweenness centrality (Freeman, 1979). Pivotal points are computationally identified and rendered so that they become preattentive, or pop-out, in the visualized network. Pivotal points are highlighted in the display with a purple ring so that they stand out in a visualized network. Graph-theoretically identifiable pivotal points allow us to reduce network-wide operations to the subset of pivotal nodes only so as to improve the interpretability of the network.

Using CiteSpace, we analyze all the papers of the ER proceedings. We present several citation statistics such as most frequently cited papers in the ER conferences and most frequently cited ER papers as well as co-citation maps and their interpretations.

The rest of this paper is organized as follows. Section 2 discusses data collection and analysis procedure. Section 3 presents several citation statistics of ER papers up to 1999. Section 4 presents high level clustering of research areas addressed in the ER conferences. Section 5 analyzes the conceptual modeling community. Section 6 concludes our paper.

Data Collection and Analysis Procedure

The trend analysis and visualization based on citation networks consists of nine steps:

1. Identify a knowledge domain. In this study, the knowledge domain of conceptual modeling is defined by full papers published in the ER conference series between 1979 and 2005.
2. Data collection. We collect ER conference papers from two sources, namely, DBLP and the Web of Science (WoS) as follows:
 - All the ER bibliographic data including paper titles and authors were retrieved from DBLP from 1979 to 2005.
 - All the reference data from 1979 to 1999 were retrieved from DBLP.
 - All the reference data including abstracts of the ER papers between 2000 to 2005, except for ER2001, were retrieved from the WoS.
 - DBLP records do not contain abstracts. Neither the WoS nor DBLP contain citation data for the 2001 ER conference.
3. Extract research front terms. Extract phrases, or terms, from titles, abstracts, descriptors, and identifiers of citing articles in the dataset retrieved from the WoS (2000-2005), except 2001. For ER conferences between 1979 and 1999, the extraction is limited to title words only because DBLP does not provide abstracts. Extracted terms are further filtered based on the so-called burst

³ <http://www.isinet.com>

- rates, which measure significant increases or decreases of frequencies over a given time interval. Burst terms are used to capture fast-growing interests.
4. Time slicing. Specify the range of the entire time interval and the length of a single time slice.
 5. Threshold selection. CiteSpace allows users to specify three sets of threshold levels for citation counts, co-citation counts, and co-citation coefficients. Citation counts are the number of times a publication is cited by the ER conference papers in the combined dataset. Two publications are called co-cited if a paper cites both of them. Co-citation counts for a given pair of publications are the number of ER conference papers in our dataset that cite the pair. Co-citation coefficients are normalized co-citation counts over each time slice. The specified thresholds are applied to three time slices, namely, the earliest slice, the middle one, and the last one. Linear interpolated thresholds are assigned to the rest of slices. In this study, most of our networks contain two types of vertices and three types of edges. Vertices could be authors, papers, journals, and burst terms, whereas edges may represent co-occurrence, co-citation, or referential links.
 6. Pruning and Merging. Pathfinder network scaling is the default option in CiteSpace for network pruning (Chen, 2004; Schvaneveldt, 1990). Users choose whether or not to apply the scaling operation to individual networks. Pathfinder network scaling is an asymptotically expensive algorithm. CiteSpace implements a concurrent version of the algorithm to process multiple networks simultaneously, which substantially reduces the overall waiting time. CiteSpace merges individual networks by taking a set union of all the vertices and selecting links that do not violate a triangle inequality condition in overlapping areas between networks. Users can choose whether or not to prune the merged network as a whole.
 7. Layout. CiteSpace supports a standard graph view and a time-zone view.
 8. Visual inspection. CiteSpace enables users to interact with the visualization of a knowledge domain in several ways. The user may control the display of visual attributes and labels as well as a variety of parameters used by the underlying layout algorithms.
 9. Verify pivotal points. The significance of a marked pivotal point can be verified by asking domain experts, for example, the authors of pivotal-point articles, and/or examining the literature, such as passages containing citations of a pivotal-point article. A particularly interesting direction of research is the development of tools that can automatically summarize the value of a pivotal point. Digital libraries, automated text summarization, machine learning, and several other fields are among the most promising sources of input.

Most Cited Papers

We present most cited papers in several groups. If paper A cites paper B, then paper A is called the source of the citation and paper B is called the target of the citation. DBLP provides a convenient way to locate an ER paper from a paper ID. For example, a paper with ScheuermannSW79 as its ID can be uniquely identified in DBLP as conf/er/ScheuermannSW79.

Table 1 is a list of ER papers that are frequently cited by all conference papers in DBLP, including ER conferences. One column lists citations made by ER papers. Another column lists citations made by papers from conferences other than ER. Table 2 is a list of ER papers that are frequently cited by journal papers indexed in DBLP. Non-ER papers refer to papers that appeared in conferences other than ER conferences. Table 3 lists non-ER papers frequently cited in ER papers up to 1999, whereas Table 4 contains non-ER papers cited in ER papers between 1979 and 2005, except 2001. Source names in Table 4 follow the journal abbreviations used by the Web of Science. See Table 5 for a list of the most popular ones. Table 4 is predominated by journal papers, except the two ER papers by Scheuermann and Santos. Table 5 lists most frequently cited journals in ER papers between 1979 and 1999.

Table 1. Most cited ER papers. Citation source: Papers from all conferences in DBLP up to 1999.

Reference ID (conf/er)	Title	Citations By ER papers only	Citations by non-ER papers	Total Citations
<i>ScheuermannSW79</i>	Abstraction Capabilities and Invariant Properties Modelling within the Entity-Relationship Approach	41	12	53
<i>SantosNF79</i>	A Data Type Approach to the Entity-Relationship Approach	29	11	40
<i>ElmasriW81</i>	GORDAS: A Formal High-Level Query Language for the Entity-Relationship Model	21	11	32
<i>Klopprogge81</i>	TERM: An Approach to Include Time Dimension in the Entity-Relationship Model	8	16	24
<i>Ling85</i>	A Graphical Query Language for Entity-Relationship Databases	21	2	23
<i>ZhangM83</i>	A Normal Form For Entity-Relationship Diagrams.	18	5	23
<i>DavisA87</i>	Converting A Relational Database Model into an Entity-Relationship Model	18	3	21
<i>AtzeniC81</i>	Completeness of Query Languages for the Entity-Relationship Model	15	6	21
<i>WiederholdE79</i>	The Structural Model for Database Design	6	15	21
<i>NavatheA87</i>	Abstracting Relational and Hierarchical Data with a Semantic Data Model	16	3	19

Table 2. Most cited ER papers. Citation source: Journal papers in DBLP up to 1999.

Reference ID (conf/er)	Title	Citations
<i>WiederholdE79</i>	The Structural Model for Database Design.	7
<i>RoseS91</i>	TOODM - A Temporal Object-Oriented Data Model with Temporal Constraints.	7
<i>ZhangM83</i>	A Graphical Query Language for Entity-Relationship Databases.	7
<i>Klopprogge81</i>	TERM: An Approach to Include Time Dimension in the Entity-Relationship Model.	6
<i>NavatheA87</i>	Abstracting Relational and Hierarchical Data with a Semantic Data Model.	5
<i>LipeckN86</i>	Modelling and Manipulating Objects in Geoscientific Databases.	4
<i>Lien79</i>	On the Semantics of the Entity-Relationship Data Model.	4
<i>SuL79</i>	A Semantic Association Model for Conceptual Design.	4
<i>WongK79</i>	Logical Design and Schema Conversion for Relational and DBTG Databases.	4
<i>RosenthalR87</i>	Theoretically Sound Transformations for Practical Database Design.	3

Table 3. Non-ER papers frequently cited by ER conference papers up to 1999.

DBLP Reference	Conference Paper Titles	# Citations
<i>afips/Chen77</i>	The Entity-Relationship Model - A basis for the Enterprise View of Data.	26
<i>ds/Abrial74</i>	Data Semantics.	25
<i>sigmod/BanerjeeKKK8</i>	Semantics and Implementation of Schema Evolution in Object-Oriented Databases.	15
<i>dood/AtkinsonBDDMZ8</i>	The Object-Oriented Database System Manifesto.	15
<i>sigmod/HammerM78</i>	The Semantic Data Model: A Modelling Mechanism for Data Base Applications.	15
<i>vldb/BachmanD77</i>	The Role Concept in Data Models.	15
<i>ds/HallOT76</i>	Relations and Entities.	14
<i>sigmod/LuskOP80</i>	A Practical Design Methodology for the Implementation of IMS Databases, Using the Entity-Relationship Model.	14
<i>db-workshops/BrodieR82</i>	On the Design and Specification of Database Transactions.	13
<i>db-workshops/YaoNW78</i>	An Integrated Approach to Database Design.	13

Table 4. Non-ER papers cited by ER conference papers (79-05), except 2001.

Cites	Burst	Centrality	Authors	Year	Source	Vol.	Page
308	0.00	0.80	CHEN PP	1976	TODS	1	9
95	8.45	0.22	SMITH JM	1977	TODS	2	105
55	6.78	0.25	TEOREY TJ	1986	CSUR	18	197
49	8.27	0.11	CODD EF	1979	TODS	4	397
44	6.89	0.09	CODD EF	1970	CACM	13	377
42	3.59	0.10	HULL R	1987	CSUR	19	201
40	6.22	0.17	HAMMER M	1981	TODS	6	351
40	7.35	0.03	ELMASRI R	1985	DKE	1	75
38	10.16	0.04	SCHEUERMANN P	1979	ER	0	121
33	14.81	0.27	RUMBAUGH JE	1991	PH	0	0
33	14.60	0.31	BATINI C	1992	BC	0	0
29	8.93	0.05	BATINI C	1986	CSUR	18	323
28	8.40	0.04	SANTOS CSD	1979	ER	0	103
25	3.73	0.03	SHIPMAN DW	1981	TODS	6	140
25	6.81	0.02	ELMASRI R	1989	BC	0	0

Table 5. Most cited Journals in ER papers up to 1999.

Journal Title	# Cites
ACM Transactions on Database Systems	886
Communications of the ACM	278
ACM Computing Surveys	265
IEEE Transactions on Software Engineering	235
Information Systems	225
Data & Knowledge Engineering	162
SIGMOD Record	96
ACM Transactions on Information Systems	93
IEEE Computer	90
IEEE Transactions on Knowledge and Data Engineering	83
Journal of the ACM	49
Computer Journal	44
VLDB Journal	37
IBM Systems Journal	35
Artificial Intelligence	34
IEEE Database Engineering Bulletin	31
Information Science	30
Distributed and Parallel Databases	27
IEEE Software	20
Journal of Intelligent Information Systems	20

All-time Prominent Challenges in Conceptual Modeling

In addition to the citation counts, we are interested in prominent research issues in conceptual modeling and how they have been addressed by the community and, in particular, by papers published in the ER conferences. Co-citation analysis of scientific literature aims to identify emergent patterns in scholarly publications derived from how scientists collectively attribute their work to prior published work. Specifically, the goal of co-citation analysis is to identify clusters of papers that are frequently cited together. Therefore, citations are seen as a filtering mechanism that selects the intellectual work that is valued by peer researchers collectively.

Prominent Co-Citation Clusters between 1979 and 2005

In this section, we present two co-citation networks. The first is pruned by Pathfinder network scaling algorithm to represent the most salient structure, whereas the second is not pruned so that the network retains more details than the pruned one.

Figure 1 shows a paper co-citation network of 760 citation links of top 487 papers cited by the ER papers between 1979 and 2005, except 2001. As shown in Figure 1 and Table 4, Chen's 1976 paper in *ACM Transactions on Database Systems* (TODS) has the largest citation. Thus, the network is focused on this seminal paper and all other papers linked to the paper.

Figure 2 shows four major co-citation clusters generated from top 548 papers and 4697 citation links between 1979 and 2005, except 2001. Again the focal point of the network is Chen's TODS paper. The term 'conceptual models' in the image is a burst term. A burst term means that the term was

associated with a sudden increase of popularity. In addition, we showed four rectangles, representing major co-citation clusters of research topics. The four major co-citation clusters of papers emerged from the network are: ER, UML, Design Patterns, and Ontology. We name these clusters based on the most cited members. Table 6 shows the list of top 20 most cited articles in each cluster.

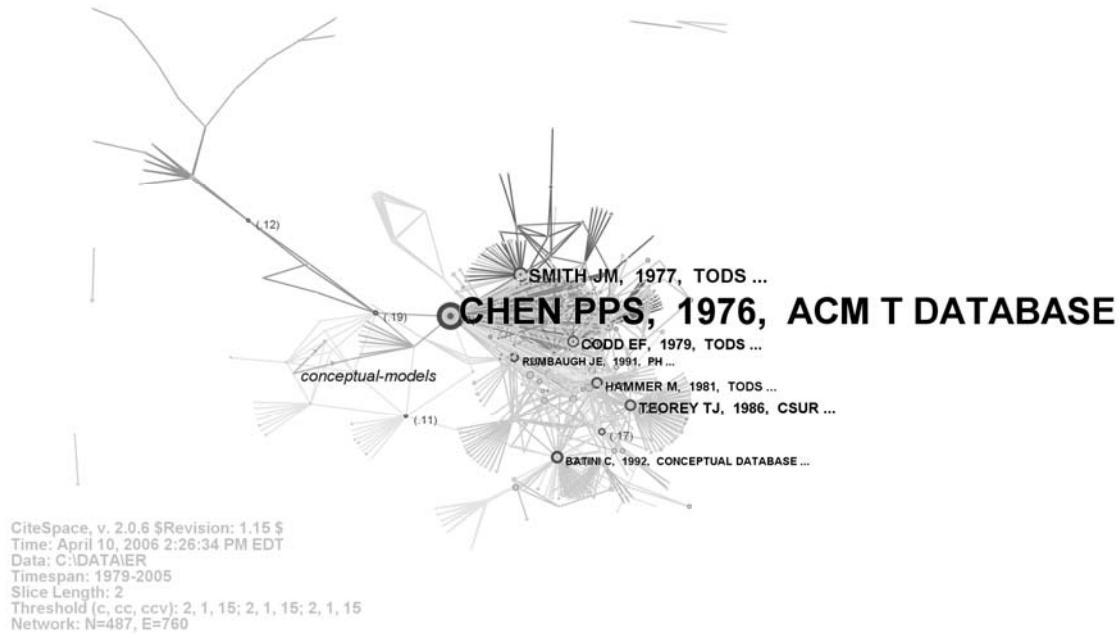


Figure 1. A document co-citation network of conceptual modeling derived from citations made by ER papers (1979-2005), except 2001.
(This network consists of 487 papers and 760 salient co-citation links. CiteSpace threshold values: c=2, cc=1, ccv=15)

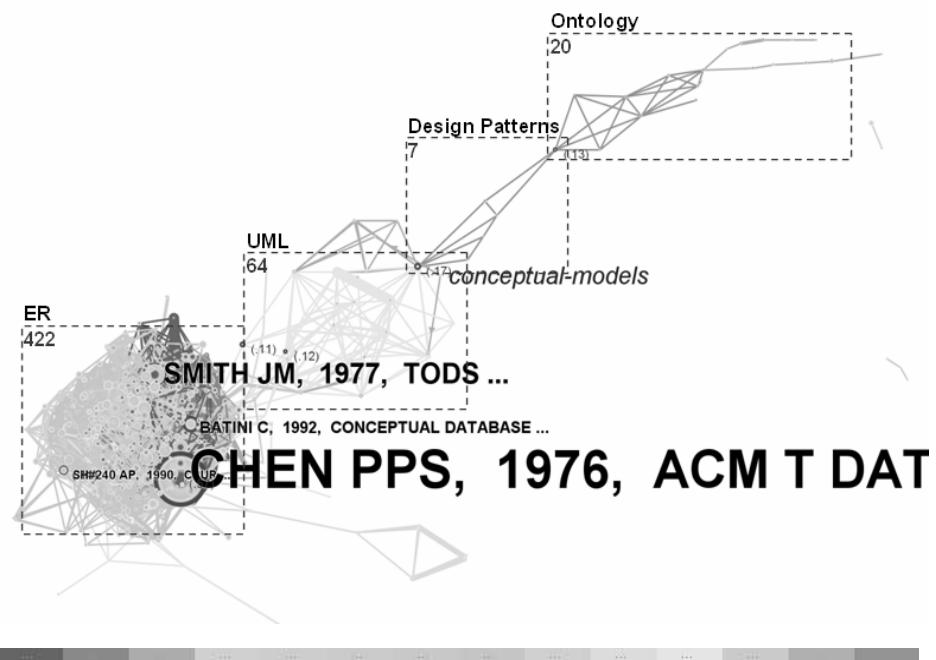


Figure 2. An un-pruned network of documents cited and co-cited by ER papers (1979-2005, two-year slices), containing 548 nodes and 4,697 links. CiteSpace threshold values: c=2, cc=1, ccv=25.

The ER cluster contains 422 papers. It appears to have two subcomponents (blue and green), but their connections are so tight that we regard them as one component. The most prominent papers in this cluster include the 1976 ER paper from Chen PPS, which was cited 308 times, a 1977 paper by Smith (cited 95 times),, followed by Teorey's Paper (cited as 55 times).

The Unified Modeling Language (UML) cluster contains approximately 64 papers. The most cited publication in this cluster is the 1999 book by Rumbaugh on UML, which is cited 9 times. The Design Pattern clusters contain 7 papers, while *the Ontology cluster* contains 20 papers. The 1995 book by Gamma et al. on Design Patterns is also cited 9 times.

The orange colors of the other two clusters indicate that they are more recent than the ER and the UML clusters. UML is linked to a 7-paper cluster on Design Patterns. It is in turn connected to a 20-paper cluster, containing papers on topics such as XML, automatic schema matching, mapping ontologies on the semantic web, and generic schema matching.

Figure 3 shows a co-citation network of publication sources of all the papers cited by the ER papers published between 1979 and 2005, except 2001. The majority of citations received by TODS of 394 citations are due to Peter Chen's ground-breaking paper in 1976. ER itself is the second largest source, cited 333 times. VLDB in the third place is cited 283 times, followed by SIGMOD 259 times, and by *Communications of the ACM* (CACM) 178 times. Nodes in the top half of the network are more recent than those in the lower half. Some of the nodes are actually the same source with different abbreviations, for example, *Lecture Notes of Computer Science* (LNCS) and the *ACM Transactions on Database Systems* (TODS). CiteSpace supports a function to merge such nodes to a unique node. Figure 3 shows the visualization without merging these nodes.

Table 6. Four major co-citation clusters of publications cited by ER papers.

Cluster	Cites	Centrality	Authors	Year	Source	Volume/Page
ER	308	0.57	CHEN PP	1976	TODS	1, 9
	95	0.11	SMITH JM	1977	TODS	2, 105
	55	0.06	TEOREY TJ	1986	CSUR	18, 197
	49	0.08	CODD EF	1979	TODS	4, 397
	44	0.06	CODD EF	1970	CACM	13, 377
UML	9	0.04	RUMBAUGH J	1999	UNIFIED MODELING	Book
	9	0.17	GAMMA E	1995	DESIGN PATTERNS	Book
	8	0.06	LINDLAND OI	1994	IEEE SOFTWARE	11, 42
	8	0.00	LIEN YE	1979	ER	155
	7	0.02	RUMBAUGH J	1991	OBJECT ORIENTED	Book
Design	9	0.17	GAMMA E	1995	DESIGN PATTERNS	Book
	5	0.07	VANDERAALST	2003	DISTRIB PARALLEL	14, 5
	3	0.13	OPDAHL AL	2002	SOFTWARE SYSTEMS	1, 43
	3	0.07	DUMAS M	2001	LNCS	V2185
Ontology	6	0.01	ABITEBOUL S	1995	FDN DATABASES	Book
	5	0.06	GRUBER TR	1993	KNOWL ACQUIS	5, 199
	5	0.03	RAHM E	2001	VLDB J	10, 334
	4	0.07	DOAN A	2002	Proc. 11th WWW	662-673

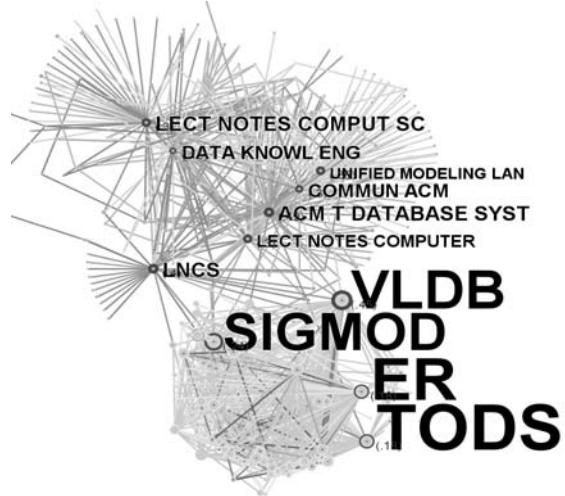


Figure 3. A Pathfinder pruned co-citation network of sources, i.e. journals and books as well as conferences derived from citations made by ER papers (1979-2005). (CiteSpace (1-year slices) threshold values: $c=2$, $cc=1$, and $ccv=5$. The network contains 311 sources and 791 co-citation links)

The Conceptual Modeling Community

In this section, we present the co-authorship map of 1,349 authors and 2,125 co-authoring links (1979-2005, slice length=3 years) in Figure 4. The ER conference co-authorship map depicts a social network of authors who have joint publications in the ER conferences. The map contains two types of vertices: authors who have published in the ER conferences and key phrases that appeared in the metadata of ER conference papers such as titles and abstracts. The size of a vertex represents the number of papers an author has published in the ER conferences. The larger the rings are, the more papers they represent. The color of each ring corresponds to the year of an ER conference in which their papers are published. The network is a hybrid network of directed and undirected graphs. Links between authors are co-authorship, which is undirected, whereas links between key phrases and authors are directed, meaning the authors used key phrases in their papers' titles and/or abstracts.

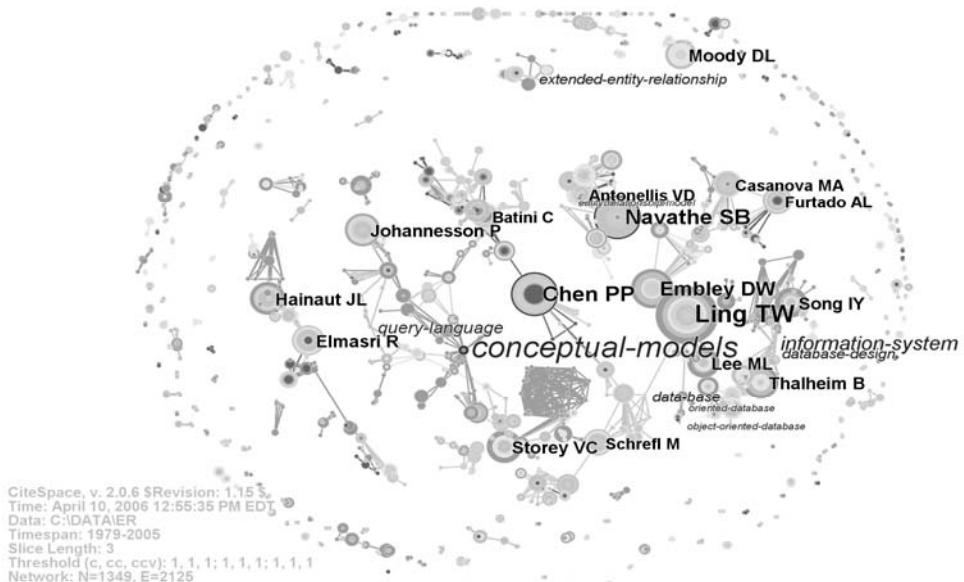


Figure 4. An ER conference co-authorship map of 1,349 authors and 2,125 co-authoring links (1979-2005, slice length=3 years). Red circles indicate burst of productivity during the entire interval.

The most productive authors who have published 8 or more ER full papers are listed in the following Table 7. For example, Tok Wang Ling published 18 papers in the ER conferences. Peter P. Chen and Shamkant B. Navathe both published 14 papers; they are also associated with a high burst rate of 6.01 and 6.71 respectively. We examined the history of each of the two authors and found that the high burst rate for Chen was due to an early peak of the number of papers in the ER conference, whereas Navathe's burst rate was due to an episode of an increasing number of papers, including 6 papers in a 3-year period starting 1991.

Table 7. Authors who have published 8 or more ER research papers (1979-2005).

# Papers	Burst Rate	Centrality	Authors
18	0.00	0.02	Tok Wang Ling
14	6.01	0.01	Peter P. Chen
14	6.71	0.00	Shamkant B. Navathe
12	0.00	0.01	David W. Embley
10	0.00	0.00	Bernhard Thalheim
10	0.00	0.00	Jean-Luc Hainaut
10	0.00	0.00	Paul Johannesson
10	0.00	0.00	Veda C. Storey
9	0.00	0.00	Daniel L. Moody
9	0.00	0.00	Il-Yeol Song
9	0.00	0.00	Mong-Li Lee
9	0.00	0.00	Ramez Elmasri
8	0.00	0.00	Antonio L. Furtado
8	0.00	0.01	Carlo Batini
8	0.00	0.00	Marco A. Casanova
8	0.00	0.01	Michael Schrefl
8	0.00	0.00	Valeria De Antonellis

Figure 5 shows a hybrid Pathfinder network of co-cited (first) authors and burst terms found in citing papers between 1979 and 1999. Most cited authors are listed in Table 8 along with their citation counts and centrality scores. For example, Chen PP is the most cited and has the highest centrality, indicating his predominant influence to the ER community. The only visible burst term in this period is the term *database*.

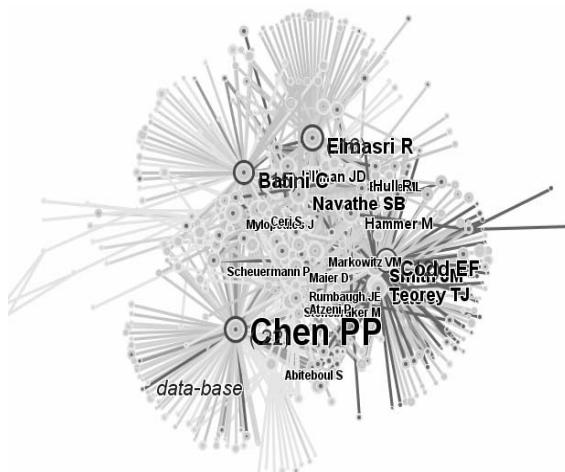


Figure 5. A hybrid Pathfinder network of author co-citations and burst term co-occurrences representing ER papers' citing behaviour between 1979 and 1999 (first authors only).

Table 8. Most cited authors in the author co-citation network shown in Figure 5 (first authors only).

Cites	Centrality	Authors
308	0.22	Chen PP
109	0.16	Elmasri R
108	0.15	Batini C
99	0.12	Codd EF
97	0.08	Smith JM
73	0.08	Teorey TJ
73	0.03	Navathe SB
64	0.06	Date CJ
61	0.02	Hammer M
55	0.02	Hull R

Figure 6 shows a hybrid network between 2000 and 2005, except 2001. Most cited authors are listed in Table 9. During this period, Abiteboul and Booch are most cited authors. Abiteboul is also significant in terms of its rate of citations. Several burst terms appear in this period, including terms such as *conceptual models*, *information system*, and *query language*. These burst terms identify potentially fast-growing research topics.

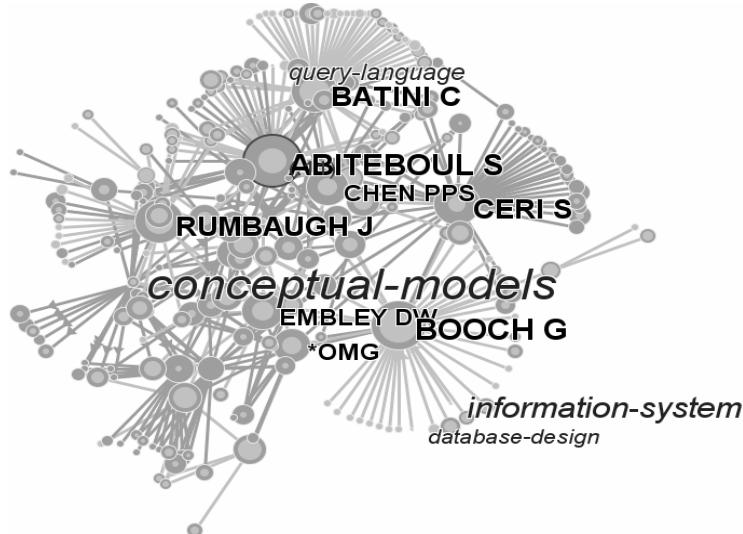


Figure 6. A hybrid Pathfinder network of author co-citations and burst term co-occurrences representing ER papers' citing behaviour between 2000 and 2005, except 2001 (first authors only).

According to a summary⁴, the Booch methodology introduced by Booch in 1991 was a widely used method in object-oriented analysis and design. The third place is James Rumbaugh, who developed the object-modeling technique (OMT), which is one of the precursors to UML. In 1994 Grady Booch and Jim Rumbaugh worked together to unify the Booch and OMT methods, which are regarded as the predecessor of UML. Carlo Batini's name appears as the fourth most cited author. Carlo Batini and Maurizio Lenzerini in their 1983 ER paper introduced a methodology for data schema integration in the ER model. In 1989, David W. Embley and Tok Wang Ling proposed what is known as the E²R model in their paper "Synergistic Database Design with an Extended Entity-Relationship Model" that solved two major problems of ER models. In their E²R model, designers no longer have to distinguish between attributes and entities and it also supports the normalization at the model level.

⁴ <http://cs-exhibitions.uni-klu.ac.at/index.php?id=447>

Table 9. Most cited authors from the author co-citation network shown in Figure 6 (first authors only).

Cites	Centrality	Authors
20	0.11	ABITEBOUL S
20	0.09	BOOCH G
17	0.03	RUMBAUGH J
16	0.09	BATINI C
16	0.08	CERI S
14	0.03	CHEN PPS
14	0.02	EMBLEY DW
13	0.02	*OMG
11	0.01	AGRAWAL R
11	0.00	FOWLER M
11	0.00	JACOBSON I
11	0.01	THALHEIM B

Conclusion

In this paper, we have presented citation analysis of all the papers published in regular ER conferences from 1979 to 2005. In some of statistics, 2001 citation data are missing as it was available neither in DBLP nor the Web of Science. We presented several citation statistics and visualizations of co-citation networks and social networks of collaborating authors. Our analysis indicates that bibliographic data can be used to identify key research focuses of conceptual modeling at various periods of time. The four identified co-citation clusters represent a trend of progression from the pioneering ER modeling, to object-oriented UML, to Design patterns, and to the more recent ontology. Finally, we showed a community of conceptual modeling in terms of the most publishing authors and most cited authors. On the other hand, we study has also revealed that the available bibliographic resources are still not readily available, for example the absence of abstracts in DBLP and missing earlier ER conferences in the ISI's Web of Science.

We believe our work contributed an in-depth analysis of a major forum of conceptual modeling and a systematic and streamlined method that one can use as frequently as needed to keep abreast of the history and the state of the art of conceptual modeling.

Notes

CiteSpace is available at <http://cluster.cis.drexel.edu/~cchen/citespace>.

Color versions of the figures are available at <http://cluster.cis.drexel.edu/~cchen/papers/issi2007/>.

References

- Chen C. 2004. Searching for intellectual turning points: Progressive Knowledge Domain Visualization. *Proc. Natl. Acad. Sci. USA* 101: 5303-10
- Chen C. 2006. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology* 57: 359-77
- Chen PP. 1976. The Entity-Relationship Model: Toward a Unified View of Data. *ACM Transactions on Database Systems* 1: 9-36
- Embley DW, Ling TW. 1989. *Synergistic Database Design with an Extended Entity Relationship Model*. Presented at Proc. of the 8th International Conf. on Entity-Relationship Approach, Toronto, Canada
- Freeman LC. 1979. Centrality in social networks: Conceptual clarification. *Social Networks* 1: 215-39
- Gregersen H, Jensen CS. 1999. Temporal Entity-Relationship Models: A Survey. *IEEE Transaction on Knowledge and Data Engineering* 11: 464-97
- Schvaneveldt RW, ed. 1990. *Pathfinder Associative Networks: Studies in Knowledge Organization*. Norwood, New Jersey: Ablex Publishing Corporations
- Teorey TJ, Yang D, Fry JP. 1986. A Logical Design Methodology for Relational Databases Using the Ex-tended Entity-Relationship Model. *ACM Computing Surveys* 18: 197-222
- Thalheim B. 2000. *Entity-Relationship Modeling: Foundations of Database Technology*: Springer
- Tryfona N, Busborg F, Christiansen JGB. 1999. *StarER: A Conceptual Model for Data Warehouse Design*. Presented at Proc. of ACM Second Int'l Workshop on Data Warehousing and OLAP (DOLAP '99), Kansas City, MO

A Comparative Study between International and Domestic Interdisciplinary Journals and Specialty Journals: A Trial Analysis of Medical Journals, Philosophy Journals and Journals in Philosophy of Medicine¹

Chen Li*, Pan Yuntao**, Ma Zheng **, Su Cheng ** and Wu Yishan **

**Chenli313@vip.sina.com*

Institute of Scientific and Technical Information of China; Beijing Jiaotong University (CHINA)

***wuyishan@istic.ac.cn*

Institute of Scientific and Technical Information of China, 15 Fuxing Road, Beijing 100038 (CHINA)

Abstract

Through the analysis of 17 international and domestic journals, this paper is to find the development trend of philosophy of medicine by using scientometric methods and visualization tool. The 17 journals include medical journals, philosophy journals, as well as journals bridging medicine and philosophy. The analysis involves such indicators as the Citing Half-Life and author affiliation. From the citation network maps derived from citation matrix, one can observe the development trend in philosophy of medicine, and the changing role played by philosophy or medicine in the development of an interdisciplinary field, namely philosophy of medicine.

Keywords

interdisciplinary science; comparative study; journal; visualization; citation network; philosophy of medicine

Introduction

The integration of available methods, theories, concepts and disciplines is always essential for achieving great academic success. The mutual permeation, interaction, and combination among different disciplines have become a norm in the development of science and technology. This paper is our attempt to do something to help the development of interdisciplinary research—to figure out a new way for analyzing an interdisciplinary science, namely philosophy of medicine.

Data

The data used in this research are derived from 3 databases. They are: Chinese Social Sciences Citation Index (CSSCI), which is built by Nanjing University; Chinese Scientific and Technical Papers and Citations Database (CSTPCD), which is built by the Institute of Scientific and Technical Information of China (ISTIC); and Social Sciences Citation Index (SSCI) produced by Thomson Scientific.

In order to compare journals in different fields and in different languages, the data used in this research include two parts: the data drawn from international journals and the data from domestic journals. Each part has 3 kinds of representative journals: philosophy of medicine journals, medical journals, as well as philosophy journals, see Table 1. The time span chosen for our data sample is from 2000 to 2004.

Statistical Analysis of International Data

Citing Half-life of International Journals

Table 2 presents the citing half-life of 8 international journals. As could be seen from Table 2, the citing half-life of philosophy of medicine journals happens to lie between medical journals and philosophy journals.

¹.This study was supported by a grant (No. 70673019) from the National Natural Science Foundation of China (NSFC)

Table 1. The sample of international journals and domestic journals.

	International Journals	Domestic Journals
Medicine	<i>Journal of The American Medical Association (JAMA)</i> <i>New England Journal of Medicine(NEJM)</i>	ZHONGYI ZAZHI (<i>Journal of Traditional Chinese Medicine</i>)(m1) BEIJING ZHONGYIYAO DAXUE XUEBAO (<i>Journal of Beijing University of Traditional Chinese Medicine</i>) (m2) ZHONGHUA YIXUE ZAZHI (<i>National Medical Journal of China</i>) (m3) ZHONGGUO YIXUE KEXUEYUAN XUEBAO (<i>Acta Academia Medicine Sinica</i>) (m4)
Philosophy	<i>Philosophy of Science(PS)</i> <i>Philosophy & Public Affairs(PPA)</i>	ZIRAN BIANZHENGFA YANJIU (<i>Studies in Dialectics of Nature</i>) (p1) ZHEXUE DONGTAI (<i>Philosophical Trends</i>) (p2) ZHEXUE YANJIU (<i>Philosophical Research</i>) (p3)
Philosophy of Medicine	<i>Bioethics(BIOE)</i> <i>Theoretical Medicine and Bioethics(TMB)</i> <i>Journal of Medicine and Philosophy(JMP)</i> <i>Journal of Medical Ethics(JME)</i>	YIXUE YU ZHEXUE (<i>Medicine and Philosophy</i>)(mp1) ZHONGGUO YIXUE LUNLIXUE (<i>Chinese Medical Ethics</i>) (mp2)

Table 2. Citing half-life of the international journals.

Discipline	Medicine		Philosophy		Philosophy of Medicine			
<i>Journal</i>	NEJM	JAMA	PS	PPA	BIOE	TMB	JMP	JME
<i>Citing Half-Life(year)</i>	4.31	5.53	12.71	9.19	7.14	7.14	6.73	5.64
<i>Journal Average(year)</i>	4.92		10.95		6.66			

Citation Network Maps for International Journals

Figures 1-5 illustrate the citation network maps for our chosen international journals over 2000-2004 periods. Each point in the figure represents a journal. A line between two points indicates that these two journals have citation relations. The arrow is pointed to the cited journal, and the other direction is the citing journal. The thickness of the line represents the link strength between two journals (or how many times they cite each other). The size of the arrows is in direct proportion to the thickness of the corresponding lines. Therefore, if the link strength is very weak, the corresponding arrows might be hidden by big arrows.

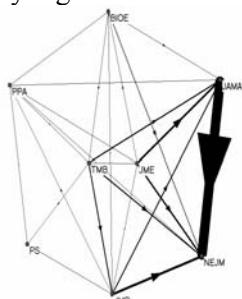


Figure 1. Citation network of 8 international journals in 2000.

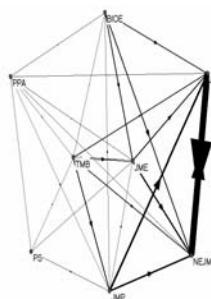


Figure 2. Citation network of 8 international journals in 2001.

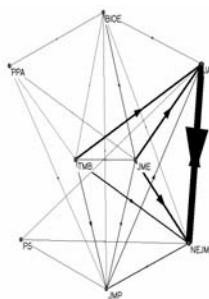


Figure 3. Citation network of 8 international journals in 2002.

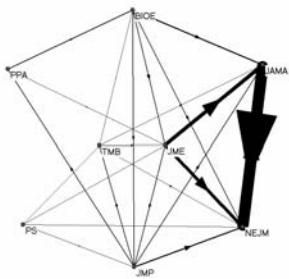


Figure 4. Citation network of 8 international journals in 2003.

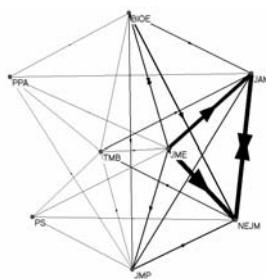


Figure 5. Citation network of 8 international journals in 2004.

From Figure 1 through to Figure 5, we find out that the lines between journals in philosophy of medicine become thicker, which means the relation among the journals gets much closer.

Author Affiliation Analysis

Our analysis on author affiliation is based on the data derived from two journals—JMP and JME. The affiliations are divided into 4 groups, including medical institutions, philosophy-related institutions, philosophy of medicine institutions and “others”. In dealing with co-authored papers, we only count the first author’s affiliation.

Table 3. Distribution of Author Affiliation for *JMP* and *JME*.

Journal	Affiliation	Medicine	Philosophy	Philosophy of Medicine	Other	Total
<i>JMP</i>	<i>Number of Papers</i>	13	11	9	9	42
	<i>Share(%)</i>	31.0	26.2	21.4	21.4	100
<i>JME</i>	<i>Number of Papers</i>	42	8	34	103	187
	<i>Share(%)</i>	22.4	4.3	18.2	55.1	100

Statistical Analysis of Domestic Data

For convenience, here we use some codes to stand for different Chinese journals, see Table 1.

Citing Half-Life of Domestic Journals

Table 4 presents the citing half-life of 9 domestic journals. From Table 4, we see that the citing half-life of philosophy of medicine journals is shortest among the three kinds of journals.

Table 4. Citing half-life of the 9 domestic journals.

Discipline	Medicine				Philosophy			Philosophy of Medicine	
<i>Journal</i>	m1	m2	m3	m4	p1	p2	p3	mp1	mp2
<i>Citing Half-Life(Year)</i>	6.28	6.84	4.72	4.96	7.64	11.63	8.54	4.40	3.45
<i>Average(Year)</i>	5.61				9.27			3.93	

Citation Network Maps for Domestic Journals

Figures 6-10 illustrate the citation network maps of domestic journals from 2000 to 2004. Since *Chinese Medical Ethics* was not included in CSTPCD until 2002, there are only 8 journals in the citation network map in 2000 and 2001.

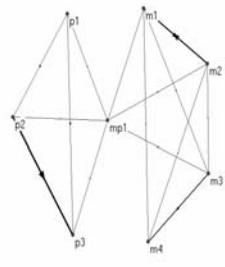


Figure 6. Citation network of 8 domestic journals in 2000.

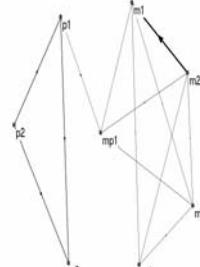


Figure 7. Citation network of 8 domestic journals in 2001.

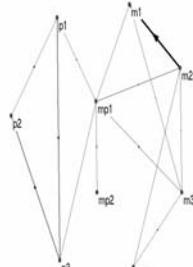


Figure 8. Citation network of 9 domestic journals in 2002.

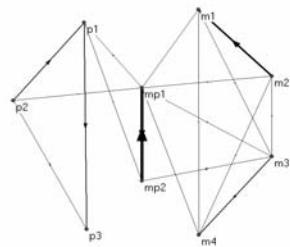


Figure 9. Citation network of 9 domestic journals in 2003.

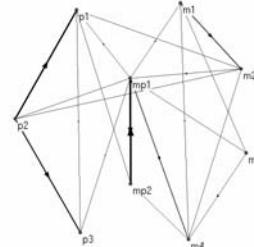


Figure 10. Citation network of 9 domestic journals in 2004.

From the figures above we can find out that:

The relation between philosophy of medicine and philosophy is getting closer, and the relationship between philosophy of medicine and medical science also witnesses an increasing trend, which means that through strengthening communication with specialty disciplines, philosophy of medicine is keeping improving itself.

The relations between domestic journals in philosophy of medicine and domestic journals of modern medicine become closer.

For a better discussion about the development trend of philosophy of medicine, and the changing role played by philosophy or medicine in the development of philosophy of medicine, we analyzed separately the share of philosophy literature, medicine literature, and philosophy of medicine literature in reference total of the journal “*Medicine and Philosophy*”, see Table 5.

Table 5. Share of philosophy literature, medicine literature, and philosophy of medicine literature in reference total of the journal “*Medicine and Philosophy*” 2000, 2002 and 2004.

Year	Share in All References(%)		
	Philosophy	Medicine	Philosophy of Medicine
2000	5.5	6.6	87.9
2002	7.4	4.9	87.7
2004	9.6	4.3	86.1

It is well known that the references of an article represent the inheritance of the forerunners' research achievements. So, to some extent, the changing share of different discipline's references in total references in a given research area can reflect the changing role played by different disciplines in that research area. Thus, from the data presented in Table 8, we find out that the importance of philosophy in philosophy of medicine increased, while that of medicine decreased. Judging by the present situation, philosophy is going to be more and more important in the development of philosophy of medicine.

Author Affiliation Analysis

Table 6 presents the distribution of author affiliation for two domestic philosophy of medicine journals, *Medicine and Philosophy* and *Chinese Medical Ethics*.

Table 6. Distribution of Author Affiliation for *Medicine and Philosophy* and *Chinese Medical Ethics*.

Journal	Affiliation	Medical Institution	Others	Total
<i>Medicine and Philosophy</i>	Number of Papers	251	158	409
	Share(%)	61.4	38.6	100
<i>Chinese Medical Ethics</i>	Number of Papers	95	97	192
	Share(%)	46.9	53.1	100

Comparison Between National and International Statistics

Based on the comparative analysis over international and domestic data above, we come to some conclusions as follows:

According to the citation network maps, the development trends for international and domestic journals in philosophy of medicine are almost the same. No matter at home or abroad, the relation between philosophy of medicine and medical science is closer. There are seldom direct connection between medicine journals and philosophy journals. The philosophy of medicine functions as a bridge, connecting medicine to philosophy.

We can see from Table 7 that the philosophical literature shares for both international and domestic philosophy of medicine journals are low; the share of philosophy of medicine in total references remains stable. The difference is that among international references, the share for medical reference is very high; but for domestic journals, the share for medical references is comparatively low, 4.3% in 2004, which is less than a half of the share for philosophy references. Besides, for domestic philosophy of medicine journals, the share of philosophy of medicine in references total is higher than international journals. This might give a warning: too high a self-citation rate for domestic philosophy of medicine journals is not conducive to new ideas and innovations.

Table 7 Reference Shares of Different Disciplines for International and Domestic Philosophy of Medicine Journals (%)

Referente Share	Philosophy		Medicine		Philosophy of Medicine	
	Year	Domestic	International	Domestic	International	Domestic
2000	5.5	5.6	6.6	42.4	87.9	52.0
2001	4.9	7.0	6.6	38.6	88.5	54.4
2002	7.4	7.6	4.9	39.1	87.7	53.3
2003	3.0	7.1	4.4	39.6	92.6	53.3
2004	9.6	3.6	4.3	43.3	86.1	53.1

The medical affiliation share of *Medicine and Philosophy* (domestic journal)'s authors is higher than that of JMP. It is partly because we can not separate "philosophy of medicine" institution category from medical institution in our raw data, while for international data, we could count philosophy of science institutions and medical institutions separately.

Conclusions

Based on the analysis above, in combination with what we find from literature review, we identify some development trends and would like to provide a few recommendations for China's philosophy of medicine.

The community of philosophy of medicine becomes bigger and bigger.

Each discipline's development is based on a powerful and highly qualified scientific community, and philosophy of medicine is not an exception. The average number of papers in domestic journals for philosophy of medicine in 2002 is 262, average number of authors per article is 1.91, co-authorship share is 58.0%; and the average number of papers in philosophy of medicine journals in 2004 reaches 301, average number of authors per article increasing to 2.11, while co-authorship share further growing to 63.7%. The increase in number of papers, average number of authors per article and co-authorship share all proved that there are more and more researchers joining in the philosophy of medicine community(Chen, 2006).

Philosophy is going to play a more and more important role in the development of philosophy of medicine.

According to the analysis above, we find out that the share for philosophy references in the reference total of philosophy of medicine journals increased annually, while the share for medical references dropped. One interpretation of this is that the philosophical issues in medical research are attracting more and more attention. Another likely interpretation is that researchers (especially medical researchers) who are engaged in philosophy of medicine increasingly realized the enlightening role of philosophy. With the confluence of philosophical perspectives with the thinking styles and research methodology of medical scientists, philosophy might play an even more important role in the future development of philosophy of medicine.

References

- Chen Li. (2006). *A Comparative Study Between Interdisciplinary Journals and Specialty Journals* (in Chinese). Master's degree thesis, Institute of Scientific and Technical Information of China

Downloads vs. Citations in Economics: Relationships, Contributing Factors and Beyond¹

Heting Chu and Thomas Krichel

hchu@liu.edu, krichel@openlib.org

Palmer School of Library & Information Science, Long Island University,
720 Northern Blvd., Brookville, NY 11548 (USA)

Abstract

Citations to 200 top downloaded papers at RePEc, a digital library in economics, were obtained from SSCI and Google Scholar respectively to address questions relating to downloads and their corresponding citations. This study finds that top downloaded documents are used in various degrees when citation is regarded as an indicator of usage. The results also show that a single downloaded paper selected for this study on average receives twice as many citations from Google Scholar as that from SSCI although the latter has been established much earlier in time. According to the coefficients computed, downloads appear having a moderate relationship with citations. However, other measures such as the download-citation ratio indicate a stronger connection between the two. While an author's reputation positively affects both download and citation frequencies, other factors (e.g., targeted readers and subject content) seem in play differently for the documents that are repeatedly downloaded or cited. The study suggests that an infrastructure which encourages downloading at digital libraries could lead to higher usage of their resources.

Keywords

digital libraries; downloads; citations; usage analysis; Google Scholar

Introduction

Along with the development of information technology, particularly the advent of the Internet, digital libraries of various kinds have been established to better serve people in all walks of life. Among all the digital libraries created so far, some are devoted to research communities in specific subject areas and meanwhile belong to the open access (OA) movement. arXiv (<http://arxiv.org>), CiteSeer (<http://citeseer.ist.psu.edu>) and RePEc (<http://repec.org>) are examples of such digital libraries.

Documents collected in digital libraries of this kind can often be freely downloaded for future reading and use. On the other hand, people may wonder whether documents downloaded from those OA digital libraries are actually used. In other words, are those documents in fact read after being downloaded? It appears difficult to find a method which allows us to directly address the aforementioned question. Most approaches have a feasibility problem. For example, it appears impractical to survey people who downloaded documents from digital libraries in order to solicit answers to the question about their usage of the downloaded documents even if we are able to trace them for survey purpose. Likewise, how can we observe others regarding their usage of downloaded documents without intrusion to their privacy even if we are permitted to do so? The quandary, however, can be circumvented by the citation approach – a time tested method although it is not controversy free.

Simply put, our approach is to gather citation counts to documents downloaded from digital libraries to find out if the documents are actually used after being downloaded. The rationale behind this methodology is that, in general, one must read a document before citing it. Based on this rationale, we intend to gather citations to the top 200 documents downloaded from RePEc (Research Papers in Economics) in the hope to answer the above question as well as to explore the following questions:

- Do any relationships exist between download frequency and citation count for the top 200 downloaded documents at RePEc?
- What factors affect downloads and citations?

¹ This research was partially funded by Long Island University, New York, USA. The authors also wish to thank Ms. Nai-yu Liu for her assistance in data collection.

- What implications will there be for RePEc given the answer to the preceding question?

Background

OA digital libraries are on the rise. The same also holds true for OA citation products and services. RePEc has been selected for this study because one of the present writers is its founder and remains actively involved in its operations and management. The reason for choosing citation tools for this study is given shortly.

The RePEc Digital Library

RePEc has been specially created to facilitate the dissemination of working papers, journal articles and software objects in the field of economics. Based on a collaborative effort of hundreds of volunteers in 54 countries, RePEc holds over 367,000 items of interest, over 266,000 of which are available online. The founders of RePEc have provided detailed descriptions about the digital library in their writings (e.g., Barrueco Cruz & Krichel, 2000). As most citation indexes traditionally cover journal articles only, documents selected for this study from RePEc are thus all journals articles even though there are actually other types of items in the collection.

Citations Products and Services

Besides the renowned citation indexes (e.g., Social Sciences Citation Index - SSCI) published by the Institute for Scientific Information (ISI, now known as Thomson Scientific), some other citation products have emerged under different auspices in recent years (Roth, 2005). Examples of such citation products include Google Scholar (scholar.google.com), Scopus (www.scopus.com), and CiteSeer. Both Google Scholar and CiteSeer fall into the OA category while Scopus is a fee-based service. Since CiteSeer focuses primarily on the literature in computer and information science, it has little relevance to economics – the subject covered by RePEc. On the other hand, Scopus has not reached the momentum that ISI enjoys with its citation indexes. ISI's citation indexes and Google Scholar are, therefore, chosen as the sources of citations for the current study. More specifically, ISI's SSCI is the target citation index for this study as economics is one discipline in the social sciences.

Literature Review

An extensive literature search for publications on downloads versus citations turned out few titles that are right on target for this study. The most frequently discussed topic along this line appears to be comparisons among various citation products or services. For example, Google Scholar and ISI citation indexes are contrasted by Charbonneau (2006) and Noruzi (2005). Jacsó (2005), on the other hand, examined Scopus as well as Google Scholar and ISI citation indexes in his comparative study. Bauer and Bakkalbasi (2005) compared citation counts from WoS (Web of Science - the Web version of ISI citation indexes), Scopus, and Google Scholar for articles published in JASIS&T (Journal of the American Society for Information Science & Technology) in 1985 and 2000 respectively.

The most related study to the present research is an editorial of International Journal of Cardiology (IJC), in which Coats (2005) compared two sets of 10 IJC papers, 10 most cited and 10 top downloaded, taken from the same one year time period in terms of their subject content and document type (e.g., review or research report). The study found no overlaps in these two sets of papers. On the other hand, Coats did not explore whether there exists any relationship between top downloaded papers and citations to them - a theme for the current study. By contrast, Bollen, Van de Sompel, Smith, and Luce (2005) did make a comparison of download and citation data. Their argument, however, was that downloading data should also be used for computing impact factor (IF) which traditionally is calculated with citation data alone.

Increasingly, scholars realize that both citations and downloads need to be combined to more completely assess the impact of journals (Kaplan & Nelson, 2002). The same notion is echoed by Darmoni, Roussel, Benichou, Faure, Thirion, and Pinhas (2000). After comparing the IF for a medical collection with a reading factor which consists of the ratio of a particular journal's download frequency to the total downloads of all journals as recorded in the system, the authors claim that the reading factor seems a credible alternative to IF.

In analyzing access data of an online archive of Physical Review and citations from Physical Review Letters to other Physical Review articles, Fosmire (2004) concluded that each article was downloaded ten times every year while both usage (measured in half-life which is again based on citations) and citation rates showed exponential decay rates with different intrinsic time scales. While Fosmire's study no doubt includes both download and citation data, it does not associate the citations with download documents. Rather, each type of data is considered independently of the other.

As one of the several major OA digital libraries, arXiv has been the subject of research on citations as well as downloads. For example, according to Brown (2001), about two-thirds of all articles submitted to arXiv since 1991 were ultimately cited somewhere between 1998 and 1999. Taking a different approach, Davis and Fromerth (2006) found that papers in the arXiv received 35% more citations on average than non-deposited articles based on an analysis of 2765 articles published in four math journals from 1997 to 2005. However, arXiv-deposited articles received 23% fewer downloads from the publisher's website (about 10 fewer downloads per article) in all but the most recent two years after publication. That is to say, an OA digital library such as arXiv enjoys a citation advantage but suffers in downloads at the publisher's website. Although the findings are interesting, the focus of both studies is placed upon whether OA has any impact on downloads and/or citations.

Relationship between downloads and citations for one single electronic journal is explored in Moed (2005) from a bibliometric perspective. A synchronous approach applied in the research revealed that downloads and citations show different patterns of obsolescence of the used materials while a diachronous approach showed that, as a cohort of documents grows older, its download distribution becomes more and more skewed, and more statistically similar to its citation distribution. In addition, Moed examined the effect of downloads and citations on each other statistically.

As shown, none of the studies reviewed above aims to investigate if top downloaded documents are also highly cited and thus frequently used by fellow researchers when citation is treated as an indicator of usage. We hence set to accomplish this research objective by analyzing citations to the 200 top downloaded journal papers from RePEc. The citation data, as described earlier, is obtained from SSCI and Google Scholar respectively.

Methodology

Our data collection was completed in about two weeks of time in early 2006. Download data for the top 200 documents, including author name, article title, journal name, publication year, volume/issue/page number, and download frequency², was automatically extracted from LogEc at <http://logec.repec.org/> using a purpose-written script. Since LogEc updates and ranks its downloading data regularly, the download data for this study was saved for future reference after being extracted.³

Citations to the selected 200 documents were gathered from SSCI and Google Scholar respectively. While citation collection from Google Scholar, again automated but followed by an extensive manual verification, was mainly based on the article title of each top download document, citation counts from SSCI were searched in Dialog with the CR (cited reference) command (e.g., CR=AKERLOF GA, 1970, V84, P488?). Citation frequencies from both SSCI and Google Scholar were gathered in about one week each to minimize the result variation due to the time difference in data collection.

The three sets of data (i.e., downloads as well as citations from SSCI and Google Scholar respectively) were then analyzed quantitatively and qualitatively to address the research questions of this study.

² Download frequency is the download request or intended download recorded at LogEc as not all the 200 journal documents can be downloaded directly from RePEc.

³ A list of the 200 top downloaded papers, including their bibliographic information and download frequency, is available upon request.

Results and Discussion

Downloads vs. Citations: A Descriptive Overview

Download counts and citation frequencies were first analyzed to present a general picture of the data collected for this study. Table 1 shows the summary measures (e.g., mean and standard deviation) for each set of the data. A visual display of the download and citation frequencies in quartile is provided in Figure 1. Although what is displayed in Table 1 and Figure 1 appears self-explanatory, several points deserve further elaboration.

Table 1. Overview of Downloads vs. Citations

Summary Measure	Download Count	Citation Frequency	
		SSCI	Google Scholar
<i>Mean</i>	1468 ⁴	344	713
<i>Maximum</i>	8739	3436	4488
<i>Minimum</i>	836	0	0
<i>Standard Deviation</i>	1023	508	762

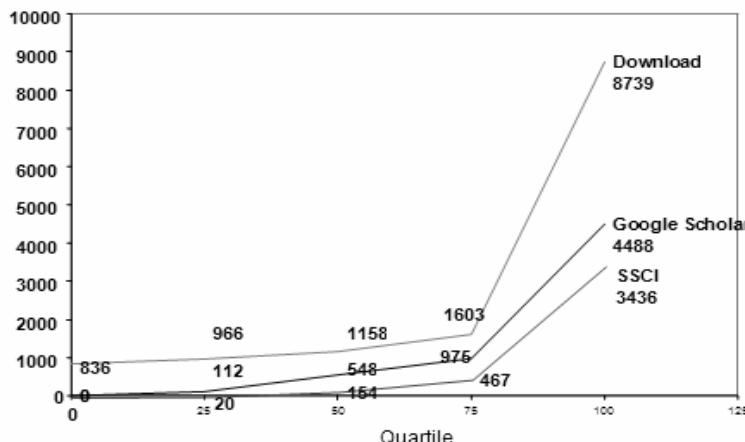


Figure 1. Download and Citation Frequencies in Quartile

First, download frequency for the 200 selected documents is on average greater than citation count, over three times more than the citation count from SSCI and twice as many as that from Google Scholar. The difference between the download and citation data can easily be explained by the fact that a document must be downloaded (or read) before cited while a document could be downloaded but never read or cited. One related note is that citations to the documents selected for this study also include those made to them from sources beyond RePEc (e.g., printed journals, e-journal sites). Otherwise, the discrepancy between downloads and citations would be even larger.

Second, a comparison of citation frequencies between SSCI and Google Scholar clearly indicates that the top 200 downloaded documents are cited more often at Google Scholar than in SSCI. Each of the 200 documents on average receives over twice as many citations from Google Scholar as that from SSCI.⁵ Since Google Scholar was established much later than SSCI, how could the former generates more citations than the latter? Possible answers to this question would be the scope and nature of both citation services. It is unarguable that Google Scholar is one of the newest “kids” on the “block of citation services”. However, Google Scholar covers a much large pool of citing sources, which differs entirely from the highly selective set of source publications (mostly journals) SSCI bases on for extracting citation data. Citations from as well as by preprints, unpublished manuscripts and other scholarly writings could all be included in Google Scholar as long as they are accessible over the

⁴ All numbers are rounded to whole digits.

⁵ The actual number is 2.07 (=713/344).

Internet. In contrast, citations would be recorded in SSCI only if they are from source publications selected by ISI.

Third, Google Scholar apparently exhibits more variation in its citation frequency distribution than that of SSCI according to their respective standard deviations. That is, some of the 200 documents were cited many times while others got fewer citations at Google Scholar. Other summary values (e.g., mean and maximum) in Table 1 for Google Scholar and SSCI also attest that different variations exist in each set of the citation data. In addition, as shown in Figure 1, most of the differences between the pair of the citation data occurred in the top 25 percent range. Likewise, the top 25 percent of the documents account for most of the variations in the download data.

The above discussion demonstrates that the top downloaded documents from RePEc are cited, many in fact highly cited, by the scholarly community. But are there any relationships between downloads and citations? The next section tries to address questions of this nature.

Downloads vs. Citations: Correlations

With the descriptive overview provided in the previous section, Pearson's r is calculated to examine correlations between downloads and each type of citations (i.e., SSCI and Google Scholar). Figure 2 presents the results obtained. All correlations are significant at the 0.01 level.

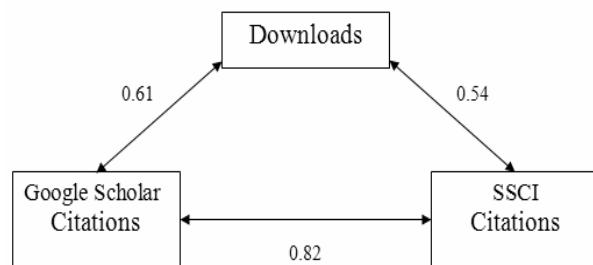


Figure 2. Correlation among Downloads and Citations

As exhibited in Figure 2, there exists little difference between the two correlation coefficients for downloads vs. Google Scholar citations (0.61) and downloads vs. SSCI counterparts (0.54). Both correlations fall within the moderate range. In other words, downloads only have a moderate correlation with citations, regardless of the source of citations being SSCI or Google Scholar, even though the coefficient value is slightly higher in the case of Google Scholar. A stronger correlation (0.82) was, however, observed between the two types of citations, perhaps reflecting the homogeneity of SSCI and Google Scholar as citation services.

A further look at the download and citation averages in Table 1 reveals that there is one citation at SSCI for around every four (4.26 to be exact) downloads from RePEc while the download number is reduced to 2.06 for each citation recorded at Google Scholar. Needless to say, both the download vs. citation ratios are much higher than what Moed (2005) reported in his study: about one citation for every 100 downloads. This huge difference in ratio could be due to the fact that this part of Moed's study covered only 25 months of time whereas the oldest downloaded document examined in the present research was published in 1897, and the average age for the top 200 downloaded documents is 15.4 years. In any case, the afore-described ratios between downloads and citations (i.e., 4.26:1 with SSCI and 2.06:1 with Google Scholar) suggest a relationship beyond the moderate range reported.

Both measures, the correlation coefficients and ratios, are computed using aggregated download and citation data. Taking a different approach, Figure 3 illustrates the connection of each downloaded document with its corresponding citations at SSCI and Google Scholar (GS). The three distributions depicted in Figure 3 appear by and large parallel. On the other hand, most of the documents positioned above the download curve are among the top 10 highly cited, which will be discussed in the following section.

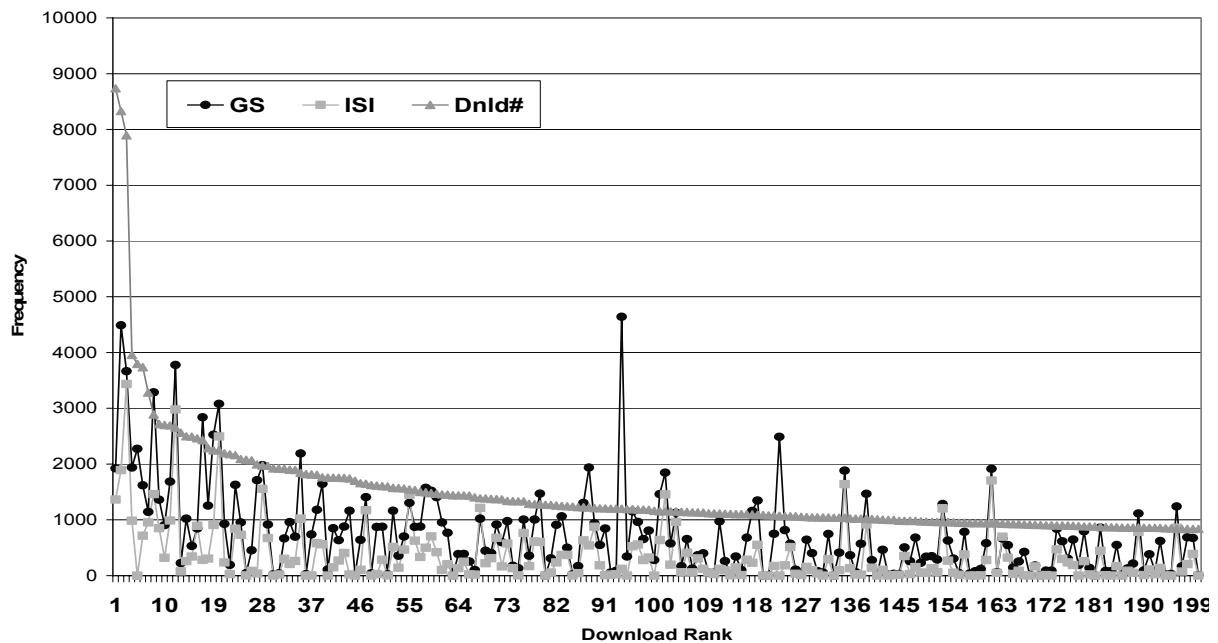


Figure 3. Download vs. Citation: A One-to-One View

Although the correlations between downloads and each type of citations seem moderate, the ratio for the two parameters at the aggregated level as well as the visualization of the three sets of data at the individual level (see Figure 3) demonstrate that downloads and citations are related in a degree stronger than moderate. A further scrutiny of that relationship is intended in a separate project with the consideration of such factors as document age to explore the question whether highly downloaded documents eventually receive more citations than those downloaded fewer times.

Downloads vs. Citations: The Top 10s

As described earlier, citations to each of the top 200 downloads were gathered using both SSCI and Google Scholar. In order to find out if papers included in this study could be both highly downloaded and heavily cited, the top 10 papers from each of the three sources (i.e., RePEc downloads, SSCI and Google Scholar citations) are to be examined and contrasted below from the perspective of their subject contents. The reason for choosing only the top 10 from each category is to keep the analysis manageable. As shown in Table 2, there are a total of 21 unique documents in the three sets of top 10s. The frequencies for all the top 10s are also listed.

Simply looking at the three columns of ranking data in Table 2, we can see that Black & Scholes 1973, Kahneman & Tversky 1979, and Romer 1986 are positioned among the top 10s in all measures while Engle & Granger 1987 and Heckman 1979 only appear in the top 10 lists of citation counts. Besides, Mankiw, Romer & Weil 1992 is ranked in both the top 10 download and Google Scholar series. The remaining documents are all placed in one of the three top 10s. In addition, there are two authors (i.e., Paul M. Romer and Robert F. Engle) who each contributed two papers to the list of 21 unique documents under discussion in this section. The download series overall shows little agreement in ranking with either the SSCI (Spearman's rho = -0.27) or Google Scholar (Spearman's rho = -0.21) citation counterparts. In other words, the 21 documents were frequently downloaded or cited for different reasons. This also indicates that the relationship between downloads and citations is hardly causal.

Among the 21 documents, nine authors are Nobel Laureates in economics. Some of them (i.e., Robert F. Engle, Clive W.J. Granger, James J. Heckman, Finn E. Kydland, and Myron S. Scholes) won the prize mainly for their papers selected in this study while others (i.e., George A. Akerlof, Milton Friedman, Daniel Kahneman, and George J. Stigler) received the award for their works in addition to what is listed in Table 2. More specifically, five out of the nine Nobel Laureates authored papers

among the top 10 downloads (i.e., #1, #2, #3, #4 & #9). In contrast, six papers from the top 10 SSCI citations (i.e., #1, #2, #3, #4, #6 & #10) were published by Nobel Prize winners whereas five papers from the Google Scholar set (i.e., #1, #2, #3, #5 & #8) belong to the same category. It becomes evident from the above amplification that the top 3 documents in each of the three measures are, as shown below, all authored by at least one Nobel Laureate.

- Top 3 downloads: #1-Akerlof 1970, #2-Black & Scholes 1973, #3-Kahneman & Tversky 1979
- Top 3 SSCI citations: #1-Kahneman & Tversky 1979, #2-Engle & Granger 1987, #3-Heckman 1979
- Top 3 Google Scholar citations: #1-Black & Scholes 1973, #2-Engle & Granger 1987, #3-Kahneman & Tversky 1979

From the viewpoint of citation frequency, what is listed above does not appear surprising at all as Garfield (1986), in one of his several studies on the same topic, pointed out that Nobel Prize winners often authored citation classics based on his analysis of ISI citations to the works of 125 Nobelists in chemistry, physics and physiology or medicine from 1965 to 1984. Citation classics at that time were defined as publications that received 300 or more citations (Garfield, 1985). Although the threshold for determining citation classics undoubtedly need to be modified as time goes by, the top 3 documents were cited at least 2,496 times and automatically fit into the category of citation classics even if the criterion is raised several times higher. Likewise, the top 3 downloads can be regarded as download masterpieces. In the age of e-publishing and open access, Nobel winners not only write citation classics but also author download masterpieces. On the other hand, three documents appeared more than once among the top 3 lists that consist of only five unique papers. This unusual concentration of top ranking downloads and citations possibly indicates

that such papers are consistently used and recognized in the scholarly community. For instance, Black & Scholes 1973, ranked #2 in downloading and #1 at Google Scholar, presented a solution to the problem of option pricing, which no one knew how before its publication. Engle & Granger 1987, positioned #2 at both SSCI and Google Scholar, reported an econometric technique that was dominating macroeconomics during the 1990s. Kahneman & Tversky 1979, the only document appeared in all top 3s (i.e., #1 at SSCI, and #3 in downloading as well as at Google Scholar), introduced the prospect theory as an alternative theory to expected value of utility maximization for economic agents. While Daniel Kahneman was awarded the Nobel Prize for this and many of his other works, the prominent ranking of Kahneman & Tversky 1979 seems to some extent unanticipated because the prospect theory is often considered as an alternative rather than a standard theory. In this case, being the first one in proposing the theory probably outweighs the alternative factor.

Akerlof 1970, although not ranked among the top 10s at either citation measure, is the most downloaded paper out of all the ones included in this study. It is a seminal research paper that uses simple mathematics, and therefore becomes an ideal teaching document. This elucidates why it received top downloads but was positioned lower than the 10th place in terms of citation ranking. The only paper remains to be discussed among the top 3s is Heckman 1979. The author described a two-stage technique that deals with the sampling bias occurring in certain studies. It was ranked #3 at SSCI perhaps because documents of this kind on econometric methodology appear more likely to get cited at SSCI than other papers included in this research.

In addition to the five individual papers that comprise the top 3s, other papers listed in Table 2 of course have particular merits that earned them a place among the top 10s. Those documents normally are the first or the most crucial in developing a theory, model or methodology. For example, Mankiw, Romer & Weil 1992, positioned #5 in downloading and #9 at Google Scholar, established a theory of economic growth which has been a popular basis for other refinements. Paul M. Romer created the theory of endogenous growth in his 1986 paper which was ranked among the top 10 in all three categories (i.e., #4 at Google Scholar, and #8 in downloading as well as at SSCI). The author later extended the theory in his 1990 document, positioned #6 at Google Scholar. The test described in Hausman 1978, ranked #5 at SSCI, is now named as the Hausman specification test. Similar accounts

can be provided for the remaining individual documents listed in Table 2. However, we decide not to do so in consideration of presentation clarity.

Table 2. 21 Unique Documents among the Top 10s

Ranking			Document
Download (f)	SSCI (f)	GS (f)	
1 (8739)	11	14	Akerlof, George A. (1970). The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. <i>Quarterly Journal of Economics</i> , 84(3), 488-500.
2 (8331)	4 (1893)	1 (4488)	Black, Fischer, and Scholes, Myron S. (1973). The Pricing of Options and Corporate Liabilities. <i>Journal of Political Economy</i> , 81(3), 637-654.
3 (7895)	1 (3436)	3 (3666)	Kahneman, Daniel, and Tversky, Amos. (1979). Prospect Theory: An Analysis of Decision under Risk. <i>Econometrica</i> , 47(2), 263-291.
4 (3958)	17	12	Stiglitz, Joseph E., and Weiss, Andrew. (1981). Credit Rationing in Markets with Imperfect Information. <i>American Economic Review</i> , 71(3), 393-410.
5 (3796)	175	9 (2273)	Mankiw, N Gregory, Romer, David, and Weil, David. (1992). A Contribution to the Empirics of Economic Growth. <i>Quarterly Journal of Economics</i> , 107(2),
6 (3737)	30	22	Cox, John C., Ingersoll, Jonathan E., and Ross, Stephen A. (1985). A Theory of the Term Structure of Interest Rates. <i>Econometrica</i> , 53(2), 385-407.
7 (3282)	19	42	Fama, Eugene F. (1980). Agency Problems and the Theory of the Firm. <i>Journal of Political Economy</i> , 88(2), 288-307.
8 (2890)	8 (1462)	4 (3287)	Romer, Paul M. (1986). Increasing Returns and Long-run Growth. <i>Journal of Political Economy</i> , 94(5), 1002-1037.
9 (2718)	25	30	Kydland, Finn E., and Prescott , Edward C. (1977). Rules Rather Than Discretion: The Inconsistency of Optimal Plans. <i>Journal of Political Economy</i> ,
10 (2701)	65	61	Krugman, Paul. (1979). A Model of Balance-of-Payments Crises. <i>Journal of Money, Credit and Banking</i> , 11(3), 311-325.
12	2 (2979)	2 (3778)	Engle, Robert F., and Granger, Clive W.J. (1987). Co-integration and Error Correction: Representation, Estimation, and Testing. <i>Econometrica</i> , 55(2),
20	3 (2496)	5 (3081)	Heckman, James J. (1979). Sample Selection Bias as a Specification Error. <i>Econometrica</i> , 47(1), 153-161.
162	5 (1706)	15	Hausman, Jerry A. (1978). Specification Tests in Econometrics. <i>Econometrica</i> , 46(6), 1251-1271.
135	6 (1639)	16	Engle, Robert F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. <i>Econometrica</i> , 50(4),
28	7 (1557)	11	Alchian, Armen A., and Demsetz, Harold. (1972). Production, Information Costs, and Economic Organization. <i>American Economic Review</i> , 62(5), 777-
102	9 (1455)	17	Hansen, Lars Peter. (1982). Large Sample Properties of Generalized Method of Moments Estimators. <i>Econometrica</i> , 50(4), 1029-1054.
55	10 (1450)	32	Stigler, George J. (1971). The Theory of Economic Regulation. <i>Bell Journal of Economics</i> , 2(1), 3-21.
17	74	6 (2840)	Romer, Paul M. (1990). Endogenous Technological Change. <i>Journal of Political Economy</i> , 98(5), s71-s102.
19	21	7 (2527)	Barro, Robert J. (1991). Economic Growth in a Cross Section of Countries. <i>Quarterly Journal of Economics</i> , 106(2), 407-443.
123	0	8 (2489)	Friedman, Milton. (1997). John Maynard Keynes. <i>Economic Quarterly</i> , Spring, 1-24.
35	15	10 (2190)	Jensen, Michael C. (1986). Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers. <i>American Economic Review</i> , 76(2), 323-329.

The above presentation reveals the following points in light of the subject contents of the top 10s examined in this section. First, papers of pedagogical nature (e.g., Akerlof 1970) would increase downloads but not necessarily citations. Second, papers about econometric methodology (e.g., Engle & Granger 1987) are usually cited more often than others at SSCI. Third, Google Scholar appears in favor of papers that are less technical and/or published more recently (e.g., Friedman 1997). Out of all the top 10 papers published in 1990s, only Mankiw, Romer & Weil 1992 is also ranked #5 in downloading. The remaining three documents (i.e., Romer 1990, Barro 1991, and Friedman 1997) all appear just in the top 10 of Google Scholar citations (i.e., #6, #7 and #8 respectively).

Conclusions

This research is conducted to address several questions regarding downloads and citations. Using citations obtained from SSCI and Google Scholar that correspond to the top 200 downloaded papers at RePEc, we find that those documents are cited or used when citing is considered as an indicator of usage. On average, a single downloaded paper receives twice as many citations from Google Scholar as that from SSCI although SSCI has been established much earlier in time.

While the relationship between downloading and citation appears to be moderate according to correlation coefficients, other measures such as the download vs. citation ratio indicate a stronger connection between the two. Therefore, we intend to explore this association further in a separate study by taking into consideration of other factors like document age. In addition, different parameters (e.g., targeted readers and subject content) seem accounting for the documents that are repeatedly downloaded or cited while an author's reputation (e.g., Nobel prize winners) simultaneously increases both measures.

What would the findings of this study imply for RePEc as a digital library in economics? In a nutshell, an infrastructure that encourages downloading at RePEc would eventually lead to higher usage of its resources. Such an infrastructure includes greater coverage of research materials that are available via open access. As the OA movement advances, more documents from digital libraries like RePEc can be downloaded to facilitate better communication in the scholarly community.

References

- Barrueto Cruz, J.M., & Krichel, T. (2000). Cataloging economics preprints: An introduction to the RePEc project. *Journal of Internet Cataloging*, 3(3), 227 -241. Retrieved June 5, 2006 from: <http://openlib.org/home/krichel/papers/shankari.html>.
- Bauer, K., & Bakkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. D-Lib Magazine, 11(9). Retrieved June 5, 2006 from: <http://www.dlib.org/dlib/september05/bauer/09bauer.html>.
- Bollen, J., Van de Sompel, H., Smith, J.A., & Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Information Processing & Management*, 41, 1419-1440.
- Brown, C. (2001) The E-volution of preprints in the scholarly communication of physicists and astronomers. *Journal of the American Society for Information Science & Technology*, 52(3), 187-200.
- Charbonneau, L. (March 2006). Google Scholar service matches Thomson ISI citation index. *University Affairs*. Retrieved June 5, 2006 from: http://www.universityaffairs.ca/issues/2006/march/google_scholar_01.html.
- Coats, A.J.S. (2005). Top of the charts: Download versus citations in the International Journal of Cardiology. *International Journal of Cardiology*, 105(2), 123-125.
- Darmoni, S.J., Roussel, F., Benichou, J., Faure, G.C., Thirion, B., & Pinhas, N. (2000). Reading factor as a credible alternative to impact factor: A preliminary study. *Technology and Health Care*, 8(3-4), 174-175.
- Davis, P.M., & Fromerth, M.J. (2006). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? Retrieved June 5, 2006 from: <http://arxiv.org/abs/cs.DL/0603056>.
- Fosmire, M. (2004). Scan it and they will come ... But will they cite it? *Science & Technology Libraries*, 25(1/2), 55-72.
- Garfield, E. (November 4, 1985). Contemporary classics in the life sciences: An autographical feast. *Current Contents*, (44), 3-8.
- Garfield, E. (June 9, 1986). Do Nobel Prize winners write citation classics? *Current Contents*, (23), 3-8.
- Jacsó, P. (2005). As we may search – Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced database. *Current Contents*, 89(9), 1537-1547.
- Kaplan, N.R., & Nelson, M.L. (2002). Determining the publication impact of a digital library. *Journal of the American Society for Information Science & Technology*, 51(4), 324-339.
- Moed, H.F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science & Technology*, 56(10), 1088-1097.
- Noruzi, A. (2005). Google Scholar: The next generation of citation indexes. *Libri*, 55(4), 170-180.
- Roth, D.L. (2005). The emergence of competitors to the Science Citation Index and the Web of Science. *Current Science*, 89(9), 1531-1536.

Does Bureaucracy Affect Research Performance of Public Research Organizations?

Mario Coccia

m.coccia@ceris.cnr.it

National Research Council of Italy and Max Planck Institute of Economics, CERIS-CNR,
Institute for Economic Research on Firm and Growth, Collegio Carlo Alberto
via Real Collegio, n. 30 10024 Moncalieri, Torino (Italy)

Abstract

The purpose of this paper is to analyse the relationship between bureaucracy and research performance within Public Research Bodies. The research methodology is applied on a sample of 100 interviewed belonging to 11 institutes of National Research Council of Italy. The main finding is that within Italian Public Research Council there is organization bureaucracy that reduces performance and efficiency of institutes. In fact, institutes have two organizational behaviours: high bureaucracy – low performance and low bureaucracy – high performance.

Keywords

bureaucracy; efficiency; research laboratory; research performance; organization rules

Introduction

Scientific research is undoubtedly one of the most debated topics in the countries of the European Union. The core of the debate concerns a European system of innovation made up of efficient research labs (Herbst, 2004), capable of producing research and innovation, both of which are necessary to boost the economic growth of the whole European Union. In almost every country, the field of public research includes university institutions as well as other agencies and bodies of different kind and size, which are usually defined as Public Research Bodies (PRB). The efficiency of PRBs depends on their structure, which is much more difficult to organise in comparison to private businesses (Lane, 1990). Crow and Bozeman (1989) analyse scientific production in public enterprises, universities and research labs, detecting lower efficiency rate and higher bureaucratisation in US public labs. At first glance, low efficiency of public research institutes is due to their nature of public organisation (Heckman *et al.*, 1997), which causes them to be pervaded by too much bureaucratisation (Gore, 1993; 1995), making them less adaptable to turbulent environmental changes. According to Green (1997), despite being chosen more and more rarely, the bureaucratic organization is still present in a large number of universities and public research bodies. Studies concerning bureaucracy within PRBs are a poorly developed area in economics and management, despite the fact that the efficiency of these institutions plays a fundamental role in today's knowledge era in order to generate technology transfer that is more and more necessary to increase economic growth (Aghion & Howitt, 1998). The aim of the present research is to analyse the relationship between bureaucracy and scientific performance in the largest Italian PRB (Public Research Body), i.e. the National Research Council (Consiglio Nazionale delle Ricerche – CNR). This research covers a gap in European economic literature on an important topic that has several contributions in North-America scientific traditions. The results of this analysis are compared with studies carried out in the US and in Northern Europe in order to detect any similarities and differences. Before dealing with the main topic, some concepts and studies referring to bureaucracy in scientific bodies are briefly outlined.

The term “bureaucracy” comes from the French *Bureau=office* and from the Greek *Kratos= power*; the origins of bureaucratic organisations date back to the Roman Empire, when a powerful administrative system, divided into offices and based on unified procedures, was systematically introduced. The “Devoto-Oli” Italian Dictionary defines bureaucracy as the whole body of public officials, a system in which public administration has too much power. In German bureaucracy is *Bürokratie*, directly derived from the French term, while in Spanish it is called *burocracia* as well as figuratively *pedantería* (pedantry), a word that refers to a person who displays unintelligent fastidiousness in his/her profession. In English, besides *bureaucracy* (needlessly time-consuming procedure), the concept is also defined as *red tape*, an expression deriving from the fact that in public

offices documents used to be sealed with a red tape. Weber (1921; 1964) claims that bureaucracy is the most modern, rational, and efficient form of administration and it can be applied to any kind of public and private organisation. Crozier (1964) and post-Weber scholars (Merton, 1970) were the first to use the word in its negative meaning, which has become prevalent today and indicates a form of organisation characterised by slowness and inefficiency. Studies on bureaucracy within Public Research Bodies (PRBs) have been carried out above all in North America (Crow & Bozeman, 1989; Bozeman *et al.*, 1992; Bozeman & Stuart, 1994; Gumpert & Pusser, 1995; Crow & Bozeman, 1998; Bozeman & Rainey, 1998; Meier *et al.*, 2000) and in Northern Europe (Gornitzka *et al.*, 1998). Crow and Bozeman (1989) analyse the *National Comparative R&D Study Project*, using a sample of over 900 US research and development labs belonging to the *Industry, Academia and Government*. The study measures bureaucracy in terms of amount of time typically required (in weeks) for each of a variety of policy and management actions; the analysis shows that *Government labs tend to be more bureaucratic on every factor. Total levels of red tape in industrial and university labs were about one-third that government labs*. Gumpert and Pusser (1995) analyse Californian universities over a period of 25 years and their study shows that an increase in the number of universities leads to the growth of administrative structures. During the period under investigation (1967–1992), the expenditure on administration functions grew disproportionately in comparison to the expenditure on instruction: *the ratio of Instructional Expenditure to Administration Expenditure went from 6 in 1966-1967 to 3 in the 1991-1992 period*. Along with growing expenditure, there was also an increase in administrative staff, which rose by 151% in comparison to a 61% increase in academic staff during the 1967–1992 period. Bozeman and Rainey (1998) analyse the topic of bureaucratic personality at National Administrative Studies Project with questionnaires administered to managers of public and private organisations. The results of their analysis show that both personal characteristics such as alienation, and organizational characteristics such as the number of records kept, have relations to references for more rules. Contrary to expectations and to much of the literature, managers in private organizations (mostly business firms) were more likely to prefer more rules than managers in public agencies. Gornitzka *et al.* (1998) take into consideration four Norwegian universities during the 1987–1995 period, showing that there was an increase in administrative personnel in comparison to academic personnel. *In this study, the growth of personnel and administrative bodies is seen as an indicator of increasing bureaucracy in Norwegian universities.*

Coccia and Gobbino (2006) investigates trends concerning scientific and academic personnel in Italy, using data from the yearbooks of the Italian National Institute of Statistics and from reports issued by the National Research Council (CNR). The results are summarised in figures 1 and 2.

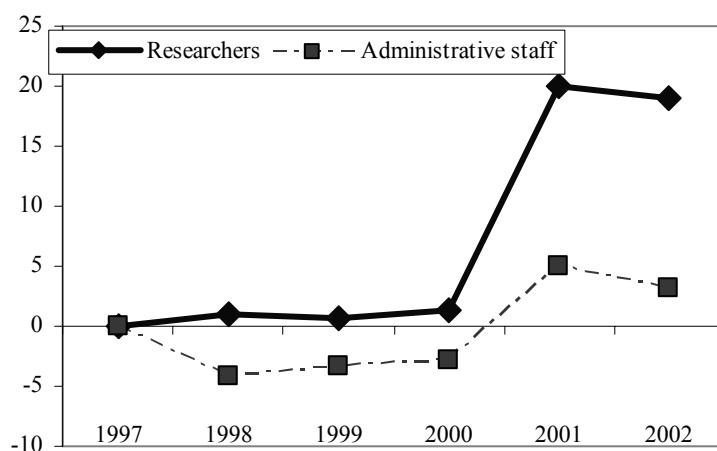


Figure 1. Temporal dynamics of researchers and administrative staff in Italian research sector (Universities and PRBs).

[Source: ISTAT 1991-2004]

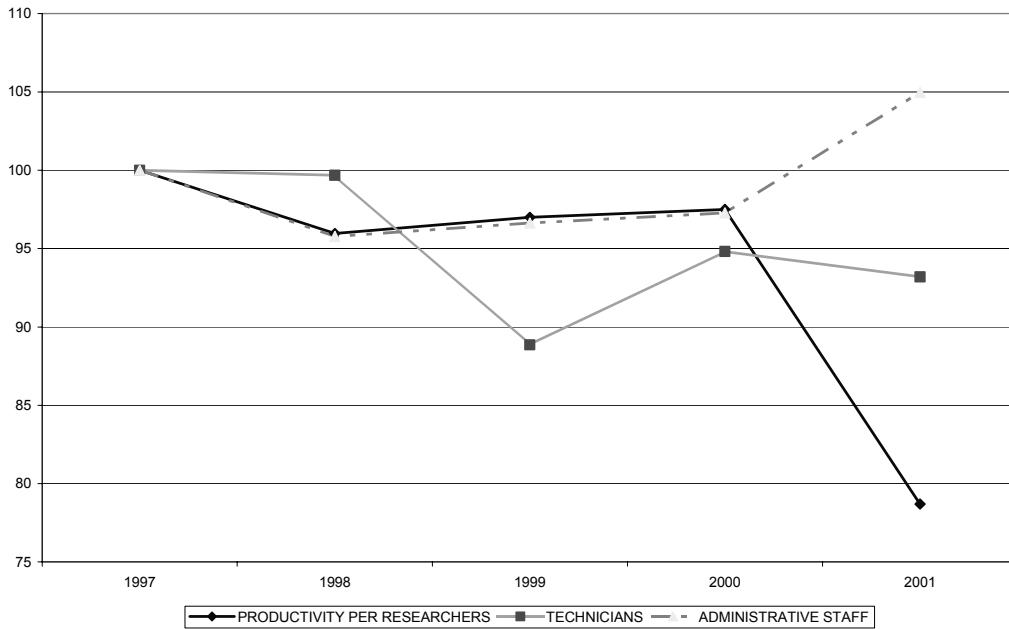


Figure 2. Temporal dynamics of productivity, technicians and administrative staff of CNR.
[Source: CNR Report, 1999-2002]

The results of Italian case-study are compared with those of Gornitzka *et al.* (1998) in Norway and Gumport and Pusser (1995) in California. Table 1 shows that the average yearly growth of administrative personnel is higher in Italian universities than in the same Californian and Norwegian institutions (respectively 15.20% in Italy versus 10.70% in California and 6.25% in Norway), while the number of researchers in Italy drops by -1.8%, versus an increase of +4.35% in California and +4.25% in Norway. The situation of Italian Public Research Bodies (PRBs) is different, since they display an increase in researchers and a decrease in administrative personnel.

Table 1. Average yearly personnel growth in percentage: comparison among states.

	California	Norway	Italy		
	University	University	Public Research Bodies	National Research Council	
Researchers (%)	4.35	4.25	-1.80	0.28	3.80
Administrative staff (%)	10.70	6.25	15.2	-1.60	0.62

In particular, the National Research Council (CNR), the largest public research body in Italy, displays low administrative personnel growth and a modest increase of researchers. These results shows that *in Italian Public Research Bodies, unlike in universities, the phenomenon called administrative bureaucracy does not occur, i.e. there is no disproportionate growth of administrative personnel in comparison to academic personnel.*

Furthermore, if we consider research productivity per researcher, the CNR in Italy has a decreasing trend (figure 2) and has the lowest productivity per researcher (Table 2), in comparison to similar

European institutions, such as the *Max-Planck Gesellschaft* - MPG in Germany¹, the *Centre National de la Recherche Scientifique* – CNRS in France, and the *Consejo Superior de Investigaciones Científicas* - CSIC in Spain. Studies carried out by Coccia (2004, 2005) on Italian CNR finds out that only 30% of researchers are high performers, whereas the remaining 70% are low performers.

Table 2. Comparison among European research bodies.

	CNR ITALY		CNRS FRANCE		CSIC SPAIN		MPG GERMANY	
	2001	2002	2001	2002	2001	2002	2001	2002
<i>Publications per researcher</i>	1.34	1.36	1.42	1.39	1.93	1.89	2.42	2.19

Source: CNR Report, 2003

In brief, CNR institutes display low efficiency, which is not ascribable to bureaucracy of the administrative type (Gornitzka *et al.*, 1998). Then, which is the cause of CNR institutes' low efficiency? The following section describes the methodology to answer such an important question, which is crucial for the correct management of public research organisations and increase their efficiency.

Research methodology

The first step of the research concerns the analysis of Reports issued by sample institutes, in order to identify the most important activities related to the functioning of such institutes. Main thematic areas and questions were included in a questionnaire. This questionnaire undergoes a pilot investigation (Bailey, 1978), in order to rectify interpretation mistakes, unnecessary or missing questions, redundant or confusing questions, etc. (Converse and Presser, 1986). The final questionnaire displays a semi-structured form (Marvulli, 1985; Manganelli Rattazzi, 1990). The questionnaire is administered by means of "face-to-face" interviews (Carli & Trentini, 1972), because when compared with other data collection methods it has several advantages in relation to the quality of the data collected, even though time and costs are higher. Semi-structured interviews using this questionnaire are carried out in a number of institutes belonging to the CNR. The sample includes 100 people (researcher, technicians and administrative staff since they represent the main subjects operating in research units) from 6 institutes and 5 sections of Piedmont and Lombardy, two large regions in Italy based on manufacturing and commercial sectors and high investments in research in comparison to other Italian regions..

Low efficiency (and research productivity) of CNR institutes is not breed by administrative staff but may be due to other causes. Bureaucracy can be also identified with the time needed to carry out administrative and scientific activities in research organisations as suggested by Crow and Bozeman (1989). Moreover according to Gornitzka *et al.*, 1998, academic bureaucracy includes the time needed to prepare meetings and to participate in them as well as all the administrative paperwork that is done inside universities. This theoretical framework is the basis to analyze the relationship between bureaucracy and research performance of Italian research institutes, given by:

$$Y = f(T_1, T_2, T_3, T_4, T_5, T_6, T_7, N)$$

where

Y = average yearly scientific production (number of domestic and international publication per researcher into institute)

T_i = time spent on the *i*-th administrative activity

N = number of documents filled in

¹ Max Planck is an association of élite research organizations that work under exceptionally rich funding conditions and therefore cannot be compared to the other organizations.

Bureaucracy is a latent variable² that is affected by a series of causes measured by the following manifest variables:

- T₁ = Time_enrol_1: average time needed to recruit term contract personnel (topic 1 in the questionnaire).
- T₂ = Time_event_organization2: time needed to organise events such as meetings, seminars, and projects (topic 2 in the questionnaire).
- T₃ = Time_event_spent3: time needed to participate in meetings and to draw up projects (topic 3 in the questionnaire).
- T₄ = Time_budget_preparation4: time needed to compile Budgets and to draw up final balances (topic 4).
- T₅ = Time_project_organization5: time elapsing from the presentation of a project application or agreement/collaboration papers to the moment when the project starts (topic 5).
- T₆ = Time_budget_approval6: time needed to approve Budgets and to make changes to the expenditure capacity of the Expenditure Centre (topic 6).
- T₇ = Time_buy_materials7: time needed to purchase scientific materials, books, journals, etc. (topic 7).
- N = Number-document-filled: number of documents required (topic 11 in the questionnaire).

The research question is: if variables T_i and N increase (indicators of the *Bureaucracy* latent variable), is there a decrease in variable Y?

The data are studied by means of a descriptive analysis, a correlation and cluster analysis using the S.P.S.S. software.

The bivariate Bravais-Pearson's correlation analysis (Girone, Salvemini, 1988) is used to find a correlation between at least two continuous variables. The value for a Pearson's can fall between 0.00 (no correlation) and 1.00 (perfect correlation). Other factors such as group size will determine if the correlation is significant (sig.). Generally, correlations above 0.80 are considered high. Moreover the correlation among the variables that is significant at the 0.01 level (2-tailed) is considered. In addition, the *Cluster Analysis* method is also applied. This procedure makes it possible to detect, within a set of items of whatever nature, a number of subsets, i.e. clusters, which are homogeneous from an internal point of view but sufficiently different from each other. Cluster Analysis techniques should display high internal (intra-cluster) homogeneity and high external (inter-cluster) heterogeneity. Therefore, if the classification is successful, items within the same cluster are close to each other, while items belonging to different clusters are further away from each other. The cluster analysis uses Ward's method and the squared measure of the Euclidean distance; results are summarised in the dendrogram.

Results

The sample is made up of 11 different institutions, of which six institute headquarters and five decentralized units. The institutes are: Institute of Ecosystem Study (ISE); Institute of Plant Virology (IVV); Institute for the hydro geological protection of the River Po basin (IRPI); Institute for economic research on firms and growth (CERIS), Milan and Turin research units; Gustavo Colonna Metrology Institute (I.M.G.C.); Institute for applied mathematics and information technologies (I.M.A.T.I.); Institute of biology and agricultural biotechnology (I.B.B.A.); Institute for Electromagnetic Sensing of the Environment (I.R.E.A.); Institute for macromolecular studies (I.S.M.A.C.), Institute of biology and agricultural biotechnology (I.B.B.A.). The questionnaire was administered to Scientific staff (Researchers and Technologists), as well as Technicians (Technical collaborators of Research Bodies and operators) and Administrative Personnel working in CNR institutes since they are the organization staff of Italian institutes. The sample used for the research is made up of 100 interviewees and is divided as follows: 27% administrative staff; 7% technical staff, and 66% scientific staff. Moreover, 51% of interviewees are females and 49% are males; 2% of the

² One of the most relevant and debated topics in the field of statistics is the so-called *latent variable*, i.e. a variable that is not directly observed, lacking both an origin and a unit of measurement. In particular, a latent variable is a variable that cannot be measured directly and that is believed to exert a causal influence on several variables that are directly observable (manifest variables).

sample belongs to the 24-30 age group, 21% belongs to the 31-40 age group, 41% belongs to the 41-50 age group, and 36% belongs to the > 50 age group. Table 3 shows the results:

Table 3. Average time and number of documents needed to carry out administrative activities within the CNR.

Topic	Item	Average value
<i>1. Contracts – staff recruitment: T_1</i>	Recruitment of staff with permanent contract	> 34.1 months
	Grant recipients	7.2 months
	Research doctorate students	6.7 months
<i>2. Organisation of events: T_2</i>	International conferences	9.1 months
	International projects	7.4 months
<i>3. Activities in one month: T_3</i>	Drawing up international projects	7.5 days
<i>4. Drawing up final balances and Budgets: T_4</i>	Drawing up Budgets	22.5 hours
	Drawing up final balances	22.8 hours
<i>5. Approval by the headquarters: T_5</i>	Approval of long-term projects	12.4 months
	Approval of one-year projects	9.7 months
<i>6. Financial activities: T_6</i>	From the allocation of funds to the approval of the Budget	59.4 days
<i>7. Purchases: T_7</i>	Materials > 7,500 €	48.5 days
	International books	12.8 days
<i>11. Documentation (number)</i>	Recruitment of staff with permanent contract	12.8
	Organisation of congresses/meetings	12.3
	Preparation of each project	12.1

Correlations Analysis

The analysis of correlations draws a distinction between scientific personnel and administrative personnel. Positive correlations are indicated with the “+” symbol, whereas negative correlations have the “–” symbol. The main results of correlation analysis for *scientific personnel* are:

“–” correlation between the SCIENTIFIC PRODUCTION variable and the Time_event_spent3 variable with r coefficient equal to -0.204 and sig. equal to 0.087.

“–” correlation between the SCIENTIFIC PRODUCTION variable and the Time_budget_preparation4 variable ($r = -0.209$ and sig. = 0.083).

The most interesting results regarding the *administrative personnel* are:

High correlation “+” between Time_event_organization2 and Number-document-filled ($r = 0.669$ and sig. = 0.003);

High correlation “+” between Time_event_spent3 and Number-document-filled ($r = 0.597$ and sig. = 0.007); correlation “+” between Time_project_organization5 and Number-document-filled ($r = 0.494$ and sig. = 0.052);

Cluster Analysis and Organisational behaviour of CNR Research Institutes considering Performance and Bureaucracy

The cluster analysis groups 11 sample institutes belonging to the CNR into two clusters made up respectively of 9 and 2 units (Fig. 3). A descriptive statistical analysis of the groups helps study differences in their organisational and strategic behaviour.

Group B (tab. 4) displays an average production value higher than that of group A. but the most remarkable result is that group B also displays the lowest average values for all the variables that are bureaucracy indicators. Therefore, *in PRBs as the time needed to carry out scientific and administrative activities and to fill in the number of necessary documents increase, there is a decrease*

in production and consequently in the production efficiency of the research institute.

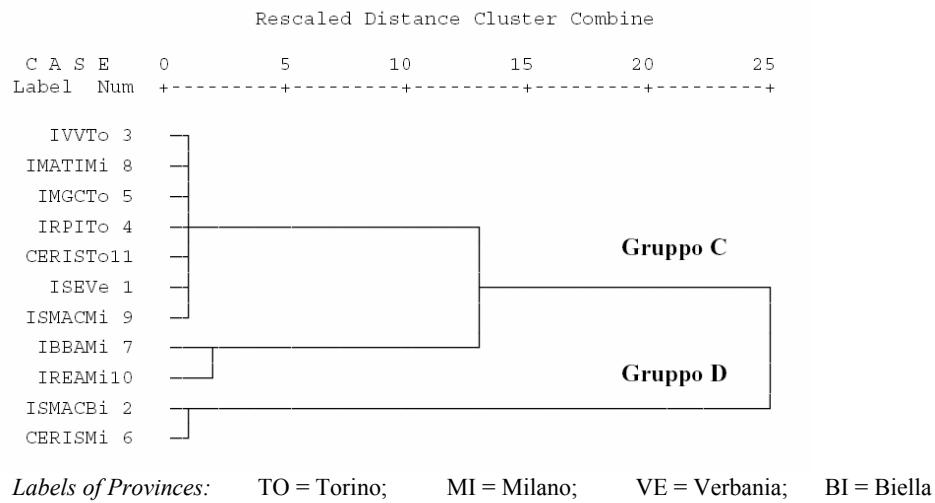


Figure 3. Dendrogram of CNR research institutes, using Ward's method.

Table 4. Descriptive statistical analysis of groups A and B following the cluster analysis of CNR institutes.

	Arithmetic mean * Group A	Arithmetic mean * Group B	Standard Deviation Group A	Standard Deviation Group B
<i>Scientific Production</i>	3.069	3.458	1.279	1.708
<i>Time_enroll_1: T₁</i>	0.402	0.338	0.076	0.012
<i>Time_event_organization2: T₂</i>	0.486	0.213	0.125	0.243
<i>Time_event_spent3: T₃</i>	0.069	0.013	0.062	0.007
<i>Time_budget_preparation4: T₄</i>	0.069	0.012	0.072	0.002
<i>Time_project_organization5: T₅</i>	0.872	0.628	0.279	0.039
<i>Time_budget_approval6: T₆</i>	0.143	0.133	0.032	0.003
<i>Time_buy_materials7: T₇</i>	0.098	0.090	0.026	0.008
<i>Number-document-filled: N</i>	6.264	4.577	1.837	0.322
<i>Age of researchers</i>	33.555	45.500	3.720	0.000
<i>Number of institutes</i>	9	2	9	2

* Some figures are low since they are standardized in annual value.

Labels:

- T₁ = Time_enrol_1: average time needed to recruit term contract personnel (topic 1 in the questionnaire).
- T₂ = Time_event_organization2: time needed to organise events such as meetings, seminars, and projects (topic 2 in the questionnaire).
- T₃ = Time_event_spent3: time needed to participate in meetings and to draw up projects (topic 3 in the questionnaire).
- T₄ = Time_budget_preparation4: time needed to compile Budgets and to draw up final balances (topic 4).
- T₅ = Time_project_organization5: time elapsing from the presentation of a project application or agreement/collaboration papers to the moment when the project starts (topic 5).
- T₆ = Time_budget_approval6: time needed to approve Budgets and to make changes to the expenditure capacity of the Expenditure Centre (topic 6).
- T₇ = Time_buy_materials7: time needed to purchase scientific materials, books, journals, etc. (topic 7).
- N = Number-document-filled: number of documents required (topic 11 in the questionnaire).

Figure 4 displays a *matrix of Bureaucracy / Performance* in which the research groups are placed. In the North – West corner there are L.B.H.P. (Low-Bureaucracy, High -Performance) units, whereas H.B.L.P. (High-Bureaucracy, Low-Performance) units are placed in the South – East corner. These areas are compatible with the study by Bozeman – Crow. The North – East (H.B.H.P.) and South – West (L.B.L.P.) corners are compatible with Weber's theories that consider Bureaucracy as the most efficient form of organisation. The latter areas do not include any of the units analysed in this study, therefore such areas are only theoretical and incompatible with the analyses of bureaucracy carried out in Italian and Foreign PRBs.

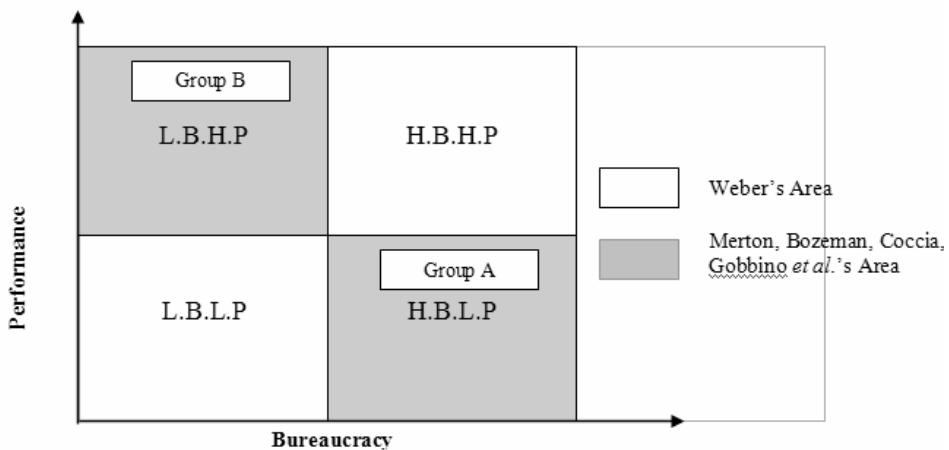


Figure 4. Strategic matrix of bureaucracy / performance and description of various related theories.

Discussion and concluding remarks

These results are particularly interesting because the correlation analysis presents a negative correlation between the variable regarding scientific production and two variables regarding governance: the one referring to the time needed to prepare projects (Time_event_spent3) and the one referring to the time needed to draw up Budgets and final balances (Time_budget_preparation4). In other words when the time needed to carry out these activities increases, the scientific production reduces. Moreover there is a high correlation between the number of documents that are filled in and the variables regarding the time needed to organise projects/congresses, the drawing up of the latter and the time needed for their approval (Number-document-filled and Time_event_organization2, Time_event_spent3 and Time_project_organization5). The above analysis shows that the time needed for certain activities is excessively long and this hinders normal and streamlined operations within the institutes. The sources of these administrative burdens are the new organization structure of CNR which is rigid in communications channels. Moreover to carry out normal scientific activities are necessary to fill several documents, for example, the procedure to recruit personnel with a permanent contract takes more than three years. The time required to allocate public funds to the institutes has increased in comparison with the past and it now takes around two months (59.4 days), slowing down the institutes' scientific activities, while the time elapsing from the presentation of a long-term project application and the beginning of that project is longer than normal due to authorisations by the headquarters in Rome (12.4 months), which often also leads to loosing external research funding. In addition to spend funds deriving from Government or self financing it is also necessary to fill several documents and to have authorizations due to national law for reducing public deficit. The CNR has organisational problems due to the high number of accounting and bureaucratic rules that origin from internal organization (CNR) and external environment (Government's law). The new accounting procedures slow down the Institutes' governance and generate uncertainties about available funds. Figure 5 summarises the results of the analysis carried out in this research. In brief, the results show that bureaucracy in Italian PRBs does not belong to the administrative type (it does not depend on a high number of administrative personnel in comparison to a low number of researchers, a situation found in Italian universities and confirmed by Gumpert and Pusser, 1995 and by Gornitzka *et al.*, 1998 in their international studies), it is instead of the academic and organisational type.

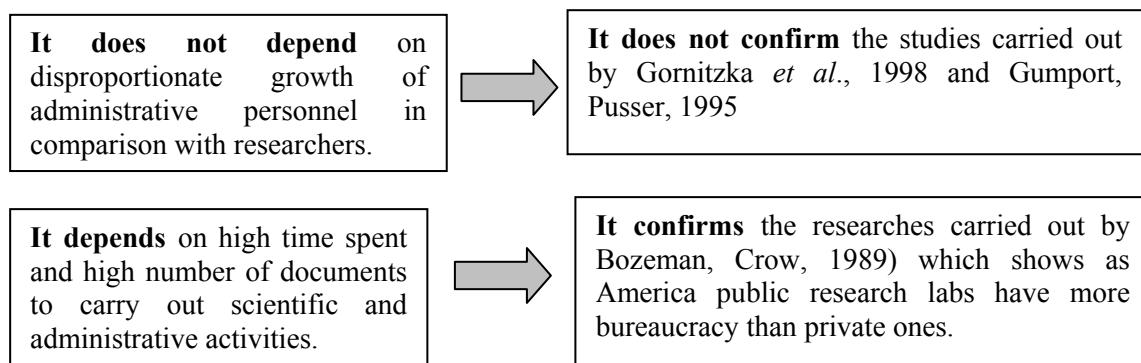


Figure5. Causes of bureaucracy in the scientific institutes of the CNR.

This organizational and academic bureaucracy can be understood if it is analyzed the reorganization of CNR. In fact, a thorough reorganisation of public research was carried out in two different phases over five years (1999 and 2003). The objective was to reduce general administrative costs and to increase technology transfer and overall efficiency of research structures. In consideration of the widely shared political objective of improving scientific research in an industrialised country such as Italy, but above all in view of the necessity for the economic system of profiting from scientific research produced in public institutions, the organisational reforms were poorly planned, creating confusion about the activities carried out by researchers, who more and more often had to deal with consultancies to external parties rather than with scientific research activities, as well as uncertainties about the future of the CNR. In particular, in order to increase the High performers' institutes the CNR is being restructured with consolidation and merger among institutes and organisation change (from line and staff to matrix). The aim of the consolidation and merger of CNR institutes is to create Italian scientific institutes of larger size, similar to the Max Planck in Germany, thinking that large labs=efficient labs. This new Italian research policy of consolidation and merger has been carried out only from an administrative and not from a scientific point of view. Although nowadays there are 107 new institutes, these often have several (2-6) decentralised units spread on the territory and far from the headquarters. This situation creates some diseconomies of scale, due to the increased costs of coordination of decentralised units. The main effects of this reform are an increase of organizational bureaucracy and confusion about the mission of this public research bodies (i.e. CNR) with a negative influence on research productivity and efficiency of the structures which have been reducing (fig. 2). This hasty and uncertain research policy reform has been transforming the Italian public research labs in hybrid organization *with many characteristics of the business firm, except for the profit motive*" (Etzkowitz, 2003) that provide a great deal of services to firms³ but they carry out little basic research. In all, the main result of this new Italian research policy is the reduction, in some cases, of certain costs (personnel, rent, and so on), but as seen has created organization bureaucracy and in terms of output increase (research performance) the effects seem very much ambiguous.

References

- Aghion, P. & Howitt, P. (1998). *Endogenous Growth Theory*. The MIT Press.
- Bailey, K.D. (1978). *Methods of social research*. New York: The Free Press.
- Bozeman, B. & Rainey, H.G. (1998). Organizational Rules and the Bureaucratic Personality. *American Journal of Political Science*, 42 (1), Jan., 163-189.
- Bozeman, B. & Stuart, B. (1994). The Publicness Puzzle' in Organization Theory: A test of Alternative

³ These activities are represented by (Coccia & Rolfo, 2002): a) analysis and technical tests (chemical and physical); b) technological services (homologation, calibration, nuclear magnetic resonance, etc.); c) quality services (accreditation, certification, quality control, etc.); d) environmental services (water monitoring, pollutant emission control, etc.); e) information technology services (data elaboration, supply of databases and data, etc.); f) health services; g) research contracts with firms and institutions.

- Explanations of Differences Between Public and Private Organizations. *Journal of Public Administration Research and Theory*, 12 (4), Jan., 197-223.
- Bozeman, B., Reed, P. & Scott, P. (1992). The presence and predictability of Red Tape in Public and Private Organizations. *Administration and Society*, 34 (24), Mar., 290-322.
- Carli, R., & Trentini, G. (1972). *L'Intervista*. Milano: Etas Kompass.
- Cnr Report (2003). *Risultati di ricerca*. Roma: D'Anselmi Editore/Hoepli.
- Coccia, M. & Gobbino, A. (2006). Analisi della burocrazia negli enti pubblici di ricerca. *Working paper Ceris-Cnr*, n. 5, Moncalieri (To).
- Coccia, M. & Rolfo, S. (2002). Technology transfer analysis in the Italian National Research Council. *Technovation*, 22, 291-299.
- Coccia, M. (2004). Models for measuring the research performance and identifying the productivity of public research institutes. *R&D Management*, Blackwell Publishers (UK), 34 (3), 267-280.
- Coccia, M. (2005). Scientometric model for the assessment of the scientific research performance within the public institutes. *Scientometrics*, Kluwer Academic publishers, 65 (30), 297-311.
- Converse, J.M. & Presser, S. (1986). *Survey question: Handcrafting the standardized questionnaire*. Beverly Hills, Ca: Sage.
- Crow, M. & Bozeman, B. (1989). "Bureaucratization in the laboratory". *Research Technology and Management*, 32 (5), 30-32.
- Crow, M. & Bozeman, B. (1998). *Limited by design. R&D Laboratories in the U.S. National Innovation System*. New York: Columbia University Press.
- Crozier, M. (1964). *Bureaucratic Phenomenon*. Canada: Tavistock Publications.
- Etzkowitz, H. (2003). Research groups as 'quasi-firm': the invention of the entrepreneurial university. *Research Policy*, 32 (1), 109-121.
- Girone, G. & Salvemini, T. (1988). *Lezioni di statistica*. Bari: Cacucci Editore.
- Gore, A. (1993). *From Red Tape to Results: Creating A Government That Works Better and Costs Less*. Washington: Government Printing Office.
- Gore, A. (1995). *Common sense government*. New York: Random House.
- Gornitzka, A., Svein, K. & Larsen, I.M. (1998). The Bureaucratization of universities. *Review of science. Learning and Policy*, Minerva, XXXVI (1), 25-47.
- Green, J. (1997). Is Bureaucracy Dead? Don't Be So Sure. *Chartered Secretary*, January, pp. 18-19.
- Gumpert, P. & Pusser, B. (1995). A Case of Bureaucratic Accretion. *Journal of Higher Education*, 66 (5), Oct., 493-520.
- Heckman, J., Heinrich, C. & Smith, J. (1997). Assessing the Performance of Performance Standards in Public Bureaucracies. *American Economic Review. Paper and Proceeding*, 87 (2), May, 389-395.
- Herbst, M. (2004). *Governance and management of research universities: funding and budgeting as instruments of change*. Center for science and technology studies, vol. 4, Bern, CH.
- ISTAT (1991-2004). *Annuari Statistici*. Roma.
- Lane, J.E. (1990). *Institutional Reform a public policy perspective*. Aldershot: Dartmouth Publishing Co.
- Manganelli Rattazzi, A. (1990). *Il Questionario*. Padova: Cleup.
- Marvulli, R. (1985). *I Questionari*. Torino Giappichelli.
- Meier, K.J., Polinard, J.L. & Wrinkle, R.D. (2000). Bureaucracy and organizational performance: causality arguments about public schools. *American Journal of Political Science*, 44 (3), 590-602.
- Merton, R.K. (1970). *Teoria e struttura sociale*. Bologna: Il Mulino.
- Weber, M. (1921). *Economy and Society*. Totowa, New Jersey: Bedminster Press.
- Weber, M. (1964). *The Theory of Social and Economic Organization*. New York: Collier Macmillan.

A Classificatory Scheme for the Analysis of Bibliometric Profiles at the Micro Level¹

Rodrigo Costas and María Bordons

*rodrigo.costas@cindoc.csic.es, mbordons@cindoc.csic.es,
Centro de Información y Documentación Científica. CINDOC-CSIC, Department of Bibliometrics,
C/Joaquín Costa, 28002 Madrid (Spain)

Abstract

Bibliometric indicators have proved to be very useful in the study of the scientific performance of countries, regions, centres, and teams as well as a support tool in scientific evaluation processes providing objective data to experts. However, different limitations and drawbacks have been described in applying bibliometric indicators at the micro level, since the low size of the unit of analysis limits the application of statistical indicators. In this study a methodology for obtaining bibliometric indicators at the individual level is presented. This methodology stresses the importance of collecting the whole production of the analysed scientists, but avoids individual rankings of scientists. Instead, it suggests the use of a classificatory scheme of researchers based on their behaviour in different dimensions. Explanations are presented in order to show how the methodology works and its advantages over other single-indicator based approaches.

Keywords

bibliometric indicators; micro-level; individual assessment; compound indicators; classification of scientists; scientific behaviours

Introduction

Bibliometric indicators are useful tools for the assessment of scientific performance of countries, regions and centres. However, they need to be properly applied, having in mind their main advantages and limitations. Their use at the micro level is usually controversial due to the low significance of statistical analyses applied to small sets of data, as well as to the fact that these units of analysis are especially sensitive to small losses of information. In spite of these limitations, analyses at the micro-level are being increasingly demanded by scientific managers, who ask for bibliometric data as a support tool for their evaluations and decision making processes.

The use of a single indicator for the assessment of a given unit has been criticized by different authors (i.e. Weingart, 2005), and the most common recommendation is to combine several indicators in order to catch a complete view of the scientific activity of the unit of analysis (Martin, 1996; van Leeuwen et al, 2001; van Leeuwen et al, 2003). This claim is especially important in the study of the scientific activity of teams and individual scientists, but up to now, there are no clear conclusions on what bibliometric indicators offer to support properly the evaluation of scientists and research teams (van Raan, 2005).

This paper presents a methodology which considers the individual bibliometric profiles of scientists, but avoids traditional ranking of individuals, and suggests the use of classification schemes to describe and analyse scientists' behaviors.

Objectives

The objective of this work is to present a useful methodology to obtain a bibliometric profile of scientists at the individual level with three complementary and consecutive aims:

- To classify researchers in different types according to their behaviour in scientific activity and impact and avoiding traditional individual rankings
- To detect a set of outstanding scientists
- To obtain a general overview of an area aggregating data from scientists.

¹ This work has been supported by an I3P grant from CSIC

Methodology

This methodology seeks to obtain an initial high number of variables which are then reduced into different compound indicators in order to simplify the analysis and interpretation by experts and peers. The different stages followed in the methodology are described below.

1. Data downloading.

The study here shown focuses on the analysis of the scientific output of 348 scientists working at the CSIC in the Natural Resources Area in 2005. These researchers are distributed in three scientific categories: Tenured Scientist (59%); Research Scientist (24%) and Research Professor (17%).

Documents published by these scientists during the period 1994-2004 were downloaded from Web of Science and gathered in a relational database. To be sure that the whole production of scientists was retrieved all different scientists' signing forms were included in the search strategy following a methodology previously described by the authors (Costas & Bordons, 2005a, 2005b). A committee of experts contribute to the study checking for the adequate assignment of documents to researchers.

2. Individual bibliometric profiles.

For every researcher a bibliometric profile comprising the following bibliometric indicators was obtained:

a) Activity indicators: total number of documents.

b) Citation based indicators: Total Number of Citations (excluding author self-citations); Citations per Document; percentage of Highly Cited Papers (HCP were those documents with more than 15 citations); h-index (following Hirsch (2005) methodology).

c) Journal Quality indicators: Median Impact Factor of documents and Normalized Journal Position (NJP) (taking into account the position of journals in descending order of impact factor within ISI subject categories) (Bordons and Barrigón, 1992).

d) Relative Impact indicators: Percentage of Documents with $RCR \geq 1$. The RCR indicator compares the citation rate of a document with that of its publication journal. A $RCR > 1$ for a document indicates that it has been cited above its journal of publication.

3. Indicator reduction.

Once the indicators were obtained for each scientist, the number of initial indicators has been grouped with Factor Analysis in order to simplify the study (Table 1).

Table 1. Factor Analysis. Rotated component matrix.

Production dimension	Observed	Expected	RCR Dimension
	Impact	Impact	
No.	.976		
<i>h</i> -index	.887		
No. Citations	.856		
HCP rate		.885	
Citations/Doc		.871	
NJP			.915
Median IF			.864
%RCR >= 1			.950

As it can be seen in Table 1, indicators were grouped in four factors or dimensions, which explain 93% of the total variance. The following dimensions were found: 1) A highly quantitative dimension related with the total production of scientists (both in documents and citations), in which the total number of documents, number of citations and h-index are contributing; 2) Observed Impact, which deals with the real impact of documents measured through the Citation per Document rate and the percentage of Highly Cited Papers; 3) Expected Impact or Journal Visibility, related with the quality of the publication journals; 4) Relative Citation Rate of papers. It is interesting to remark that the h-

index contributes mainly to the first dimension, as it correlates very well with total number of documents and total number of citations.

4. Indicator standardization.

Since the different indicators presented above have different scales, standardization is necessary in order to have all of them in the same range of values. To do that, every value of each indicator was divided by the maximum value in that indicator. As a result, all standardized indicators ranged between 0 and 1. Afterwards, the following composite indicators were built for each scientist:

Quantitative dimension= N.Doc.-ST + Total citations-ST + h-index-ST (This final value ranges between 0 and 3)

Observed Impact dim.= HCP-ST + Citations/Doc.R-ST. (Ranges between 0 and 2)

Expected Impact dim.= Median IF-ST + JNP-ST. (Ranges between 0 and 2)

Relative Citation Rate= % Documents with RCR>=1 (Ranges between 0 and 100)

("ST" means standardized indicator)

5. Classification of researchers.

Percentiles 25 and 75 were calculated for each composite indicator (see Table 2).

Table 2. Percentiles P25 and P75 values.

Dimension	N	P25	P75	Interq. Range	Outliers
Production dim.	332	0.27	0.70	0.43	1.45
Observed impact dim.	337	0.15	0.47	0.32	1.03
Expected impact dim.	327	0.87	1.17	0.30	1.70

Percentiles 25 and 75 (or quartiles 1 and 3) values are presented for each of the three main composite indicators or dimensions. Researchers were classified into 3 zones according to the following criteria(see Figure 1):

- Zone 1: values lower or equal P25;
- Zone 2: values greater than P25 and lower or equal P75;
- Zone 3: values greater than P75

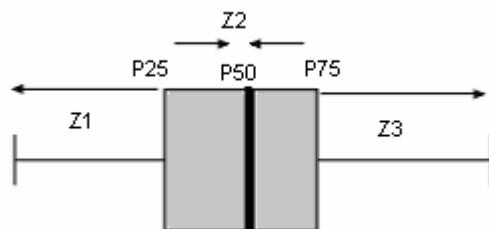


Figure 1. Zones in which are classified the researchers.

For the RCR Dimension, researchers have been classified in 3 different zones: Class “1”: researchers with less or equal 33% of their documents with RCR >=1; Class “2”: researchers with more than 33% and less than 66% of their documents with RCR>=1; Class “3”: researchers with more than 66% of their documents with RCR>=1. As a result, every scientist is located in a specific class (1, 2 or 3) in each of the four dimensions, so each researcher presents a four-dimension vector which describe different aspects of his/her behaviour. Derived from this, different scientific behaviours can be detected, and similar scientists can be grouped.

Results

The scientific production of the CSIC Natural Resources scientists amounted to 6093 documents in the Web of Science. In Table 3, different output indicators are presented.

Table 3. Average results per scientist in the CSIC Natural Resources Area

Indicator	Mean±SD	Median	Range (Min-Max)
No. Documents	25±19.50	22	1-162
No. Citations	199.22±230.17	134	0-2201
No. Citations per Document	7.25±5.08	6.44	0-40.96
Rate HCP	0.18±0.16	0.15	0-1
Median Impact Factor	1.270±0.534	0.182	0.200-3.687
NJP	0.65±0.14	0.67	0.05-0.96
h-index	7.98±4.51	8	1-29
%RCR>=1	45.14±18.89	44.44	7.14-100

Differences in the behaviour of scientists according to their scientific category were analysed by means of the methodology proposed. Concerning the Production dimension significative differences among scientific categories were found ($p<0.001$) (figure 2). Research Professors (the highest category) perform the best in this dimension. According with the box plot on the right, Research Professors present better performance also in real and expected impact, showing significative differences with the other two scientific categories ($p<0.05$).

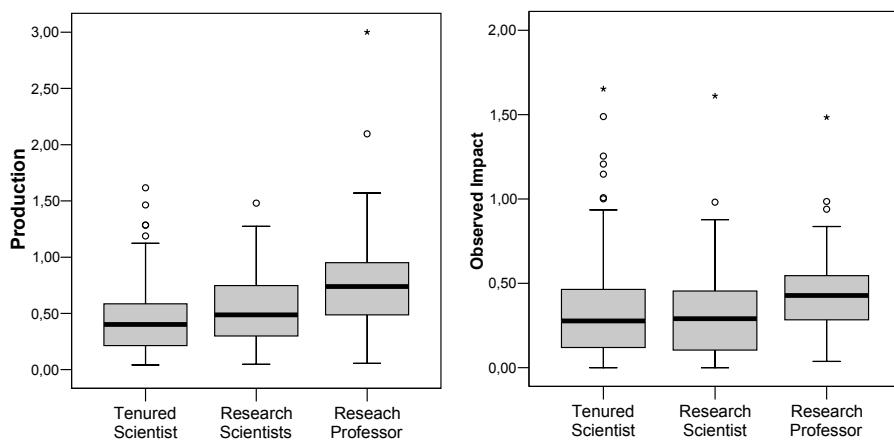


Figure 2. Production and Observed Impact dimensions by scientific category.

No significant differences between Tenured Researchers and Research Professors are observed. Perhaps this is related with the extremely good performance of new comers (Tenured Scientists), who are aware of the importance of publishing their work in the best journals of their fields to attain rewards and get a position. Concerning RCR, there are no significative differences among categories. Research Professors obtain the best performance but the differences are not significant.

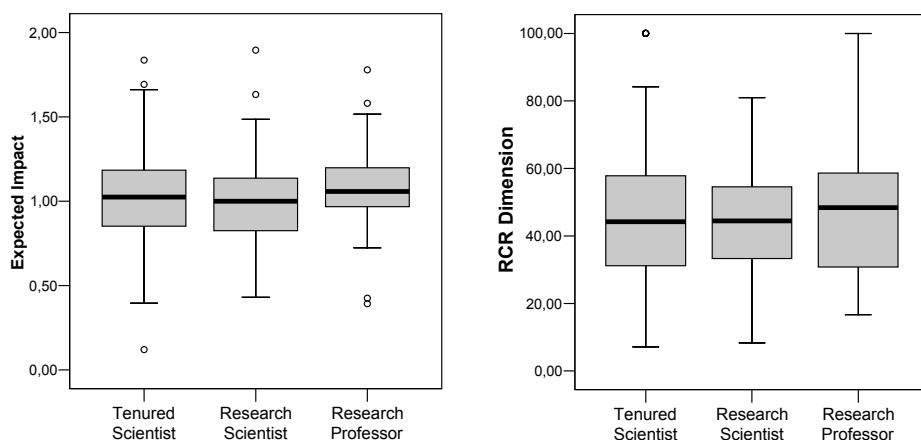


Figure 3. Dimensions of Expected Impact and Relative Citation Rate.

Different profiles of researches are observed through this methodology according to the scores they obtain in the different dimensions analysed, e.g. top scientists, which get "3" in the four dimensions or selective researchers, which show intermediate scores in quantity but very good performance in impact dimensions.

A general classificatory scheme grouping scientists with similar behaviours will be presented and discussed.

Conclusions

The usefulness of a methodology developed to obtain bibliometric data of scientists at the micro level and to describe different scientist behaviours is shown in this study. The need to overcome technical and methodological problems commonly described for the collection of data at the individual level was minimized in order to reduce the risk of incomplete retrieval of author's production.

Several advantages of this methodology can be pointed out:

- The research performance of scientists is analysed from a multidimensional point of view (involving quantitative and qualitative indicators). Scientists are classified according to their performance in each of the dimensions as compared to the whole area.
- Scientists are grouped into classes, avoiding traditional rankings which are very sensitive to small losses of information and are not very handy. The reduction of variables into 4 dimensions makes the management and interpretation of the results easier.
- Outstanding scientists can be identified and newcomers can be virtually placed in the context of the area when applying for a new position.
- As a multi-indicator approach it is difficult to be manipulated by researchers.
- This methodology can also be used for describing the behaviour of teams, departments and specialized centres.

This methodology has been considered useful by Natural Resources experts as a complementary tool to peer review and other evaluation procedures, since it provides a general overview of the activity of scientists in different facets. We are aware that the research performance of scientists involves activity in some facets that can not be measured through bibliometric indicators, so these indicators should always be considered part of the evaluative process and not the evaluation itself.

References

- Bordons, M.; Barrigon, S. (1992). Bibliometric analysis of publication of Spanish pharmacologists in the SCI (1984-89). 2. Contribution to subfields other than pharmacology and pharmacy (ISI). *Scientometrics*, 25(3): 425-446.
- Costas, R.; Bordons, M. (2005a). Bibliometric indicators at the micro-level: some results in the area of natural resources at the Spanish CSIC. *Research evaluation*, 14(2): 110-120
- Costas, R. Bordons, M (2005b). Methodological procedure to overcome the lack of normalization of author names in bibliometric analyses at the micro-level. *ISSI 2005: Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm: ISSI, p. 688
- Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102: 16569-16572
- Martin, B.R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36(3): 343-362
- Van Leeuwen, T.N.; van der Wurff, L.J.; van Raan, A.F.J. (2001). The use of combined bibliometric methods in research funding policy. *Research evaluation*, 10(3): 195-201
- Van Leeuwen, T.N.; Visser, Martijn S.; Moed, H.F.; Nederhof, T.J.; van Raan, A.F.J. (2003). The Holy Grail of science policy: Exploring and combining bibliometric tool in search of science excellence. *Scientometrics*, 57(2): 257-280
- Van Raan, A.F.J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1): 133-143
- Weingart, P. (2005). Impact of bibliometrics upon the science system: inadvertent consequences? *Scientometrics*, 62(1): 117-131.

Applying Egghe's General Theory of the Evolution of Information Production Processes to the World Wide Web

Viv Cothey

viv.cothey@wlv.ac.uk

School of Computing and IT, University of Wolverhampton, Wolverhampton, WV1 1SB (UK)

Abstract

Lotkaian informetrics in the form of the general theory of the evolution of information production processes is used to investigate, analyse and interpret the information structures found in four samples of the World Wide Web. The notion of evolution in the Web is operationalised by reference to the growth in the number of path components (or “/” separators) that are present in Web-page urls. Two different evolutionary periods are identified and their Lotkaian properties are compared and contrasted. The unexpected results lead to the discovery that the logarithmic size-frequency distribution of the incremental production process is distorted and is no longer a simple power law. It is conjectured that this is due to the prevalence of automated systematic Web-page creation which becomes significant at later stages of the Web’s evolution.

Keywords

evolution; World Wide Web; Lotka; scale factor

Introduction

A popular model when studying the World Wide Web (the Web) is a graph or complex network. Web pages become the nodes or vertices of a graph and the Web hyperlinks become arcs from a source node to a target node in a directed graph. Such graphs are sometimes called complex networks since they provide examples of scale free features. The most important of these is the frequency distribution of node indegree which is found to follow a power law (Barabási & Albert, 1999). The study of complex networks has become an active topic within the theoretical physics community (see, for example, <url:<http://arxiv.org/archive/cond-mat>>).

Lotkaian informetrics provides us with an alternative model for studying the Web (Egghe, 1997; Rousseau, 1997). This model offers a rigorous mathematical treatment of the notion of an information production process (IPP). An IPP comprises a collection of *sources* each of which has one or more *items* (Egghe & Rousseau, 1990). The Web is thus an IPP where each item is an inlink or reference from a Web page and each source is the Web page to which the inlink refers. (Note that compared with previously the notion of “source” has been inverted.) The number of items gives the size of the source. Lotkaian informetrics where the size-frequency distribution is a power law unifies many of the so-called “bibliometric laws” as well as connecting to Pareto’s econometrics, Zipf’s principle of least effort and Mandlebrot’s fractals (Egghe, 2005).

The Web IPP is Lotkaian (Barabási & Albert, 1999; Rousseau, 1997). The classic “signature” of a power law relationship is that the distribution has a linear form when plotted logarithmically. This is illustrated in Figure 1 which shows the size-frequency distribution for the Web IPP that is obtained from a sample of Canadian research and education institutions.¹

The study of Lotkaian informetrics predates that of complex networks (for example, Mitzenmacher, (2004)). Explanations such as “preferential attachment” (Barabási & Albert, 1999) for the graph indegree power law were already established as “success breeds success” (Price, 1976) within the informetrics community. The exponent or scale factor of the governing power law in respect of the Web IPP is Lotka’s α in the relationship $f(n) = C/n^\alpha$. Scale factors for the Web have been determined empirically (for example, Barabási & Albert, (1999) and Broder *et al.* (2000)) although there is

1. I am grateful to Dr J Sylvan Katz and the Rogers Communication Centre, Ryerson University, Toronto who kindly provided the Canadian data.

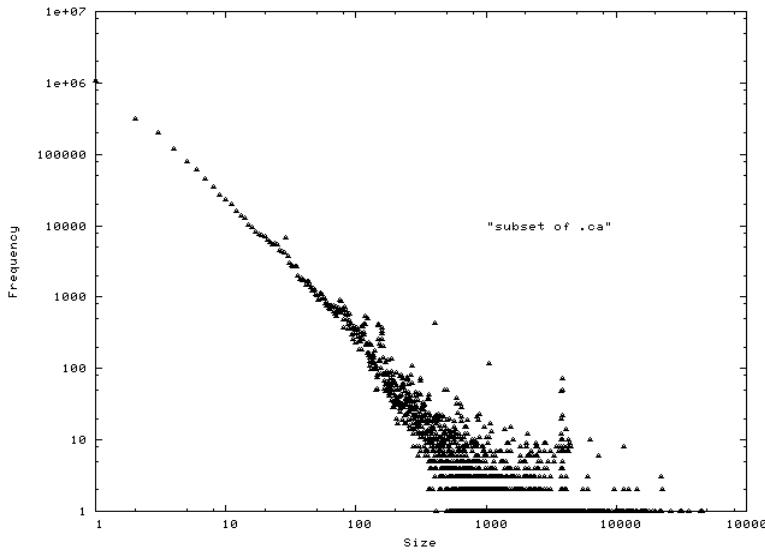


Figure 1. Logarithmic plot of Web IPP size-frequency distribution

concern regarding the validity of the computational procedures used to estimate these from empirical data (Newman, 2005; Nicholls, 1987; Rousseau, 1993).

In the study of complex networks it is tacitly assumed that the Web is essentially a *static* object. Large samples of the Web are taken to be representative and their scale factor provides an estimate for the *fixed* scale factor for the Web. The procedures adopted are essentially descriptive. This contrasts with the informetric approach that includes the possibility of mathematical proof. The study of informetrics is now considering the *evolution* of networks (Egghe, in press). An important analytic result is that the scale factor of an evolving IPP is *not* fixed. In particular Egghe's general evolutionary theory of IPPs defines the necessary and sufficient conditions under which α will increase, reduce or stay the same as the IPP grows.

Hence the alternative Lotkaian informetrics model provides a richer theoretical basis for studying the Web compared to the complex network model. Contrariwise the Web provides an accessible supply of empirical data with which to investigate Lotkaian informetrics.

In this paper we report a study aimed at applying the general evolutionary theory of IPPs to the Web. The study makes use of four substantial empirical samples of the Web in order to explore the methodology and issues involved. The study is novel in that it applies the qualitative and quantitative predictions of Egghe's general evolutionary theory to the World Wide Web. This entails a novel approach to how the notion of evolution is operationalised so that an evolving Web IPP can be constructed. The study is also novel in using theory to interpret as well as to just describe the Web IPP. That is, the study is explanatory as well as evocative.

It is concluded that the Web has two distinct evolutionary periods. The general evolutionary model is applicable to the first evolutionary period but not to the second. It is conjectured that the generative processes of the Web at work during this second period are such that the Lotkaian nature of the Web is distorted.

Method

Data collection

A Web crawler is used to collect data in respect of the Web IPP. This is an automatic technique to sample a large collection of Web-pages. A computer program (sometimes called a robot) emulates a user methodically clicking each link on a Web-page to fetch those pages and then clicking on each

link on every Web-page obtained and so on. This recursive collecting of Web-pages is controlled by some crawling policy that constrains the robot, for example to collect from only a particular Internet domain or collection of Web hosts (Cothey, 2004).

Because of the time and effort required the data obtained by the Web crawling undertaken by commercial Web search engines is proprietary and commercially confidential. Accordingly the data for this study are specially collected by the blinker (Web link crawler) robot (Cothey, 2005). Four samples of the Web are studied each being a Web IPP. Two of these relate to single Web domains (csic.es and wlv.ac.uk) the others to collections of academic institutions in each of Canada and Belgium. The blinker robot operated from several computers in order to reduce and share the computing involved.² The number of sources and items in each of the Web IPPs studied is shown below.

Constructing an evolving Web IPP

The sample data are cleaned and processed, for example to remove “dead links” and to standardise the representation of Web-page urls. Self links are removed and multiple links are considered as one. Also, Web-pages to which there are no links are removed. Each of the processed sample data thus generates a Web IPP.

It is not possible to construct a time based sequence of the creation of Web IPP sources and items other than by repetitive Web crawling. But frequent large scale Web crawling is not generally feasible.

However the format of a Web-page url provides an intrinsic sequence that can be used to operationalise the notion of evolution. This is because all the path components of the url must be created in the order in which they appear. For example the url,

<http://www.ih.csic.es/publicaciones/webasclepio/botones/ indices/532_vazquez.htm>

has four path components, publicaciones, webasclepio, botones, and indices. The sequence of creation is therefore,

1. www.ih.csic.es
2. www.ih.csic.es/publicaciones
3. www.ih.csic.es/publicaciones/webasclepio
4. www.ih.csic.es/publicaciones/webasclepio/botones
5. www.ih.csic.es/publicaciones/webasclepio/botones/indices

followed by the file 532_vazquez.htm.

Since this applies to all urls then we can construct the evolutionary sequence of Web-pages

1. pages with urls of the form http://<domain>
2. pages with urls of the form http://<domain>/<file>
3. pages with urls of the form http://<domain>/<path>/<file>
4. pages with urls of the form http://<domain>/<path>/<path>/<file>
5. pages with urls of the form http://<domain>/<path>/<path>/<path>/<file>

and so on.

Hence we can construct evolutionary steps for a Web IPP by starting with the requirement that all the urls of Web-pages in the IPP must be of the first form noted above. The next subsequent step for the evolving Web IPP the requirement is that all urls must be of either the first or second form etc. Note that new source Web pages of an earlier form may appear not until a subsequent or later IPP because their items are of the later form (IPP sources must have size > 0). Similarly the size of existing IPP sources will grow as more items evolve.

The size of each Web IPP studied is given in Table 1. This gives the number of sources and items in the evolving IPP after an evolutionary step.

² I am grateful also to Dr Paul Wouters (Virtual Knowledge Studio, Amsterdam) and Isidro Agullo (CINDOC, Madrid) for their support.

Table 1. Sources and items in an evolving Web IPP

Web IPP	Step 2	Step 5	Step 8
<i>csic.es</i>	2,723 sources 21,328 items	47,826 sources 459,202 items	59,123 sources 523,004 items
<i>wlv.ac.uk</i>	240 sources 1,637 items	58,015 sources 504,510 items	97,144 sources 884,452 items
<i>subset of .ca</i>	84,488 sources 895,855 items	1,582,130 sources 18,817,691 items	2,241,695 sources 23,533,968 items
<i>subset of .be</i>	13,158 sources 130,074 items	716,425 sources 6,538,340 items	2,106,808 sources 15,317,880 items

Computing the rate of Web IPP evolution

In general both the sources and items of an IPP will evolve. That is, an existing IPP is transformed to a new IPP as sources are created or destroyed and items are created or destroyed. Day to day experience of using the Web provides ample evidence that new sources, i.e. Web pages, become available and also that they become no longer available. The creation or destruction of items is the addition or removal of links to a (live) Web page. The operationalisation of Web evolution used for this study supports the creation of both sources and items but not their destruction. Note however that even with this restriction this does not make the Web IPP analogous to a citation IPP of papers and references. The mutability of the Web (Katz & Cothey, 2006) is preserved in that an existing Web page can evolve and refer to new Web page. The citation IPP is immutable in this sense.

In the general theory the sources and items evolve according to a power law where the exponent is the rate of evolution. Hence, given an evolutionary sequence of an IPP, the growth in sources (or items) can be plotted logarithmically and the exponent estimated as the gradient of the resulting curve. The average rate of evolution during a particular evolutionary “period” is found by a least squares best fit over the period. This is illustrated in Figure 2 which shows the growth in sources for the Web IPP obtained from the *csic.es* domain.

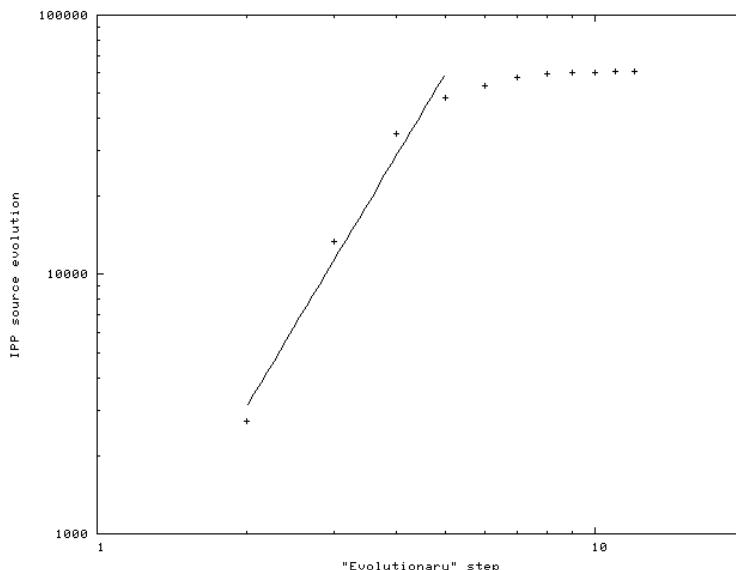


Figure 2. Evolutionary growth of Web IPP sources

The average rate of source evolution during the period step 2 to step 5 is given by the gradient of the line shown. It is evident that there are two (at least) distinct evolutionary periods. After step 5 at the end of the first period the rate of source evolution much reduces. This is the case also for the rate of

evolution of items. Both source and item evolution for each of the Web IPPs studied is illustrated in the Appendix.

Source and item average evolution rates are found for each of the two evolutionary periods step 2 to step 5 and step 5 to step 8 for each of the Web IPPs studied. The rate of source evolution is generally called b. The rates during the first and second periods are referred to as b(1) and b(2) respectively. Similarly with the rate of item evolution which is generally called c. The source and item evolution rates are given in Table 2.

Table 2. Web IPP source and item evolution rates

Web IPP	sources b(1)	items c(1)	sources b(2)	items c(2)
<i>csic.es</i>	3.22 +/- 0.36	3.53 +/- 0.49	0.46 +/- 0.05	0.28 +/- 0.05
<i>wlv.ac.uk</i>	6.13 +/- 0.37	6.29 +/- 0.45	1.09 +/- 0.19	1.20 +/- 0.10
<i>subset of .ca</i>	3.28 +/- 0.49	3.36 +/- 0.38	0.76 +/- 0.19	0.49 +/- 0.13
<i>subset of .be</i>	4.42 +/- 0.35	4.37 +/- 0.24	2.30 +/- 0.14	1.77 +/- 0.29

Computing the Web IPP scale factors

In principle the scale factor, that is Lotka's α , for an empirical IPP can be estimated statistically as the gradient of the straight line best fitting the logarithmic size-frequency distribution such as is shown in Figure 1. The difficulty lies in the detail of determining statistically the best fitting straight line to the empirical data. In consequence this approach is deprecated (Newman, 2005; Rousseau, 1993). In its place we use a maximum likelihood estimation approach (Nicholls, 1987) and a tabular fitting procedure based on Rousseau (1993) and Crovella *et al.* (1998).³

Three scale factors are computed for each of the evolving Web IPPs. These are given in Table 3. The scale factor after step 2 is called $\alpha(2)$, after step 5, $\alpha(5)$, and after step 8, $\alpha(8)$. These values correspond to the start and end of the two evolutionary periods studied, that is steps 2 to 5 and steps 5 to 8.

Table 3. Evolving Web IPP scale factors

Web IPP	$\alpha(2)$	$\alpha(5)$	$\alpha(8)$
<i>csic.es</i>	2.1954	2.1638	2.1632
<i>wlv.ac.uk</i>	2.2362	2.1583	2.1564
<i>subset of .ca</i>	2.1512	2.1228	2.1218
<i>subset of .be</i>	2.1792	2.1377	2.1314

Predictions from the general theory of the evolution of IPPs

A central result of the theory is the quantitative relationship that

$$\delta = (\alpha + c - 1 + (\alpha - 1)(b - 1))/c \quad (1)$$

where α , b and c are as already described and δ is the scale factor of the evolved IPP. The result thus allows us to predict the scale factor δ of an evolving IPP. Note that this depends only upon the evolution rates of the sources and items and the starting scale factor. In addition, as a consequence of this quantitative relationship we have the qualitative conclusion that

³ The maximum likelihood estimation approach in conjunction with reference to prepared tabulated values is used with the cumulative or complementary function (Crovella *et al.*, 1998).

1. $b < c \Leftrightarrow \alpha > \delta$
2. $b > c \Leftrightarrow \alpha < \delta$
3. $b = c \Leftrightarrow \alpha = \delta$

Hence if the rate of evolution of sources is less than the rate of evolution of items then the scale factor of the Lotkaian IPP reduces. Similarly if the rate of evolution of sources is more than the rate of evolution of items then the scale factor of the Lotkaian IPP increases. It is only when the rate of evolution of sources and items are the same that the scale factor for the evolving IPP remains constant.

Results

The four examples of Web IPP studied all share the characteristic of growing more rapidly at the outset and, after step 5 in the evolutionary process, slowing considerably. The growth data is illustrated in Figures A1 to A4 in the Appendix. Two evolutionary periods are considered for each Web IPP studied. Source and item evolution rates are shown in Table 2.

During the first period of evolution (steps 2 to 5) the source evolution rate b is less than the item evolution rate c for three out of the four Web IPPs studied. The general theory predicts that in these circumstances the scale factor in these three cases will reduce. This prediction is borne out by the data. The scale factors are shown as $\alpha(2)$ and $\alpha(5)$ in Table 3. The source evolution rate b is larger than the item evolution rate for the fourth Web IPP (subset of .be) so the scale factor should increase. This is contradicted by the data.

The numerical prediction for the evolved scale factor $\alpha(5)$ is given by Equation 1. Table 4 compares these values with the actual values estimated from the data for the three Web IPPs studied.

During the second period of evolution (steps 5 to 8) the rates of source and item evolution are both much reduced for all the Web IPPs. Also now for three out of the four Web IPPs the source evolution rate b is more than the item evolution rate c . The general theory predicts that the scale factor should now increase for these three Web IPPs. However the corresponding scale factors $\alpha(5)$ and $\alpha(8)$ shown in Table 3 show a very small contradictory decrease. For the wlv.ac.uk domain Web IPP the source evolution rate b continues to be less than the item evolution rate c . In line with the general theory the scale factor continues to reduce albeit by only a small amount.

Table 4. Predicted and actual scale factors for evolutionary period one

Web IPP	Predicted scale factor	estimated actual scale factor
<i>csic.es</i>	2.0904	2.1638
<i>wlv.ac.uk</i>	2.2048	2.1583
<i>Subset of.ca</i>	2.1238	2.1228
<i>Subset of.be</i>	2.1927	2.1377

As before the numerical prediction for the evolved scale factor $\alpha(8)$ is given by Equation 1. Table 5 compares these predicted values with the actual values estimated from the data for the four Web IPPs studied.

Discussion

The two periods of evolution for the Web IPPs studied are evident from the source and item growth illustrated in the Appendix and from the values shown in Table 2. During the first evolutionary period for the Web IPPs the rate of item growth exceeds that of source evolution for three out of the four samples. Put simply, new links are being created between Web-pages faster than Web-pages having inlinks are being created. This corresponds to a period of reduction in the scale factor of the size frequency distribution. The qualitative model provided by Egghe's general theory of the evolution of

Table 5. Predicted and actual scale factors for evolutionary period two

Web IPP	predicted scale factor	estimated actual scale factor
<i>csic.es</i>	2.9120	2.1632
<i>wlv.ac.uk</i>	2.0521	2.1564
<i>subset of.ca</i>	2.7415	2.1218
<i>subset of.be</i>	2.4784	2.1314

IPPs appears reasonable during this first evolutionary period. Lack of agreement with regard to the .be sample can perhaps be explained by the small differences found in the rates of source and item growth and their relatively wide error margin.

The theory's quantitative prediction for the Web IPP that is a subset of the .ca domain is also good. The quantitative predictions for the much smaller domain specific Web IPPs and for the subset of the .be domain are not as good. This could be a consequence of the greater statistical error associated with the smaller samples. However it may also reflect peculiarities with individual domains that are "averaged out" in the multi domain .ca Web IPP. The Web IPPs of individual domains are more likely to be subject to distortions caused by domain specific systematic behaviour discussed below. It should be noted though that the predictions are better when compared with those for period two.

During the second evolutionary period the model's predictions fail almost completely both qualitatively and quantitatively. The Web IPPs evolutions during this second period are problematic in that the absolute number of sources and items created is small (and diminishing) and in two cases the growth in items falls off and drops below that of sources. This failure of the model prompts a closer examination of the evolving distributions in order to discover a cause for the anomalous results.

Figure 3 illustrates the size-frequency distribution generated from the evolutionary increment for the *wlv.ac.uk* Web IPP during the second period from step 5 to step 8.

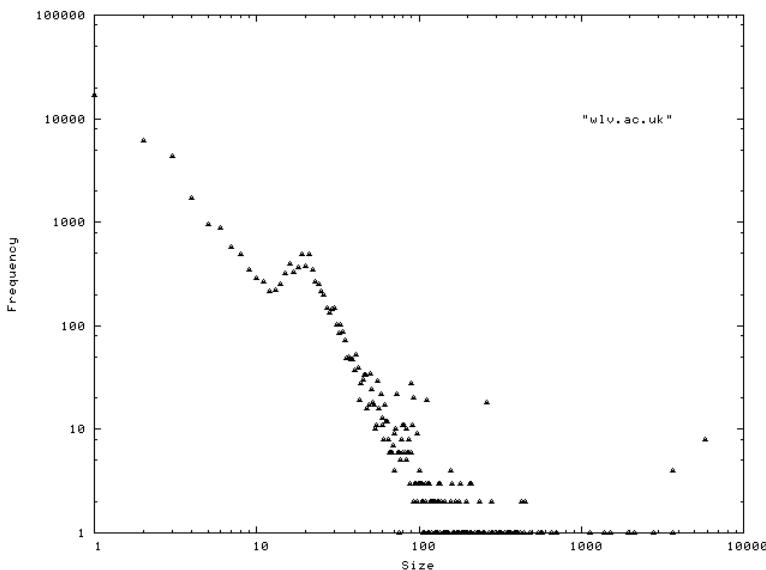


Figure 3. Logarithmic plot of incremental size-frequency distribution

The classic power law signature is absent or at least highly distorted. A possible explanation is that the evolutionary processes during the second period of evolution include a disproportionate amount of automatically created Web-pages. It is conjectured that this or a similar systematic behavior distorts

the essentially probabilistic nature that underpins the Lotkaian IPP. The general theory is thus successfully detecting non-Lotkaian production processes.

The assumption that there is single fixed scale factor for the Web entails the rate of evolution of sources being the same as the rate of evolution of items. We have seen how when the notion of evolution is operationalised as the url path length then this assumption fails. What about evolution over time? Can the rate of evolution of sources and items over time be maintained in equilibrium? This must be regarded as an open question at present. However as there is more and more systematic creation of Web content then an equilibrium is doubtful and it is possible that not only will the future Web have a variable scale factor, it is also possible that it will not be Lotkaian.

Conclusions and future research

The study has shown how an evolving Web IPP can be constructed and analysed using Egghe's general theory of evolution of information production processes.

Given how the notion of Web evolution was operationalised, the theory has been used to interpret essential differences in the production that occurs as the evolution of the Web proceeds. The Lotkaian model of the Web IPP is confirmed during the early evolutionary steps but during later evolutionary steps the theory detects distortions in the Lotkaian model. It is thought that this is because of the increasing prevalence of systematically created Web-pages and hyperlinks. It is expected that future research will focus on explaining anomalous (or non-Lotkaian) phenomena within the Web IPP.

References

- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509-512.
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000). Graph structure in the Web. *Computer networks*, 33(1-6), 309-320.
- Cothey, V. (2004). Web-crawling reliability. *Journal of the American Society for Information Science and Technology*, 55(14), 1228-1238.
- Cothey, V., (2005). Some preliminary results from a link-crawl of the European Union research area Web. In P. Ingwersen and B. Larsen (Eds.) *Proceedings of tenth international conference of the International Society for Scientometrics and Informetrics*, (pp. 2121-220). Stockholm: Karolinska UP.
- Crovella, M. E., Taqqu, M. S. and Bestavros, A. (1998). Heavy tailed probability distributions in the World Wide Web. In R. E. Feldman, R. J. Adler and M. S. Taqqu (Eds.) *A practical guide to heavy tails: statistical techniques and applications*. Boston: Birkhauser.
- Eggle, L. (1997). Fractal and informetric aspects of hypertext systems. *Scientometrics*, 40(3), 455-464.
- Eggle, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. London: Elsevier.
- Eggle, L. (in press). The general evolutionary theory of information production processes and applications to the evolution of networks. *Journal of informetrics* 1.
- Eggle, L. and Rousseau, R. (1990). Introduction to informetrics: quantitative methods in library, documentation and information science. London: Elsevier.
- Katz, J. S. and Cothey, V. (2006). Web indicators for complex innovation systems. *Research evaluation*, 15(2), 85-95.
- Price, D. J. de S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27, 292-306.
- Rousseau, R. (1993). A table for estimating the exponent in Lotka's law. *Journal of documentation*, 49(4), 409-412.
- Rousseau, R. (1997). Sitations : an exploratory study. *Cybermetrics* 1(1). Retrieved 8 February 2007 from url:<http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html>
- Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics* 1(2), 226-251. Retrieved 8 February 2007 from url:http://www.internetmathematics.org/volumes/1/2/pp226_251.pdf
- Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5), 323-351.
- Nicholls, P. T. (1987). Estimation of Zipf parameters. *Journal of the American Society for Information Science*, 38, 443-445.

Appendix

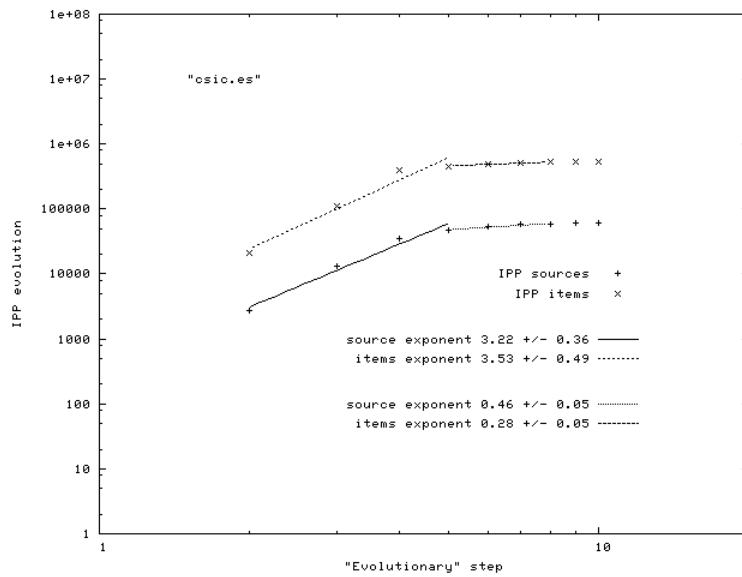


Figure A1. Evolutionary growth of sources and items for the “csic.es” Web IPP

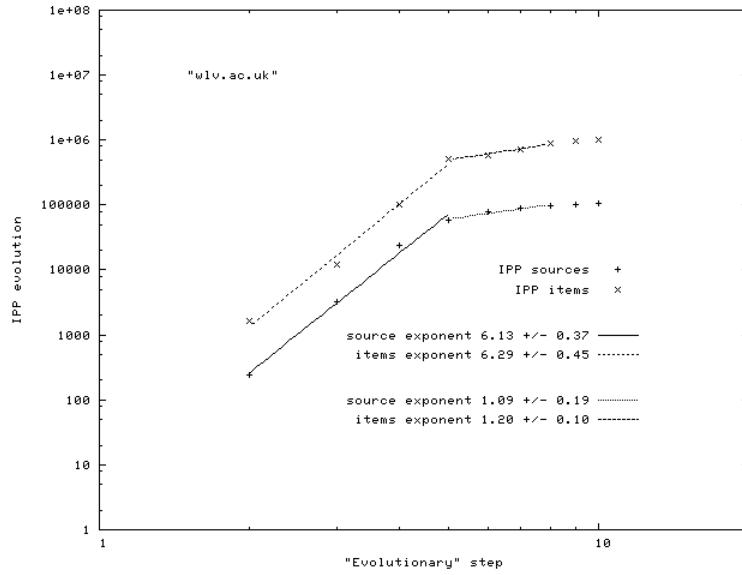


Figure A2. Evolutionary growth of sources and items for the “wlv.ac.uk” Web IPP

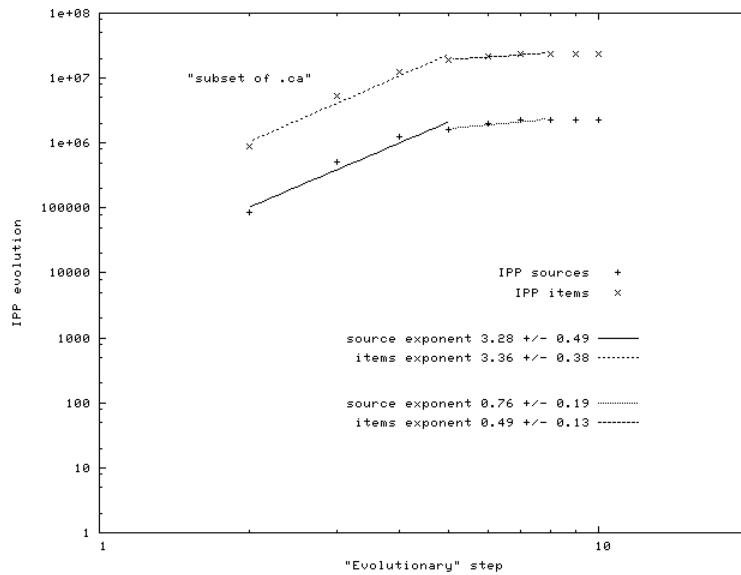


Figure A3. Evolutionary growth of sources and items for the Canadian sample Web IPP

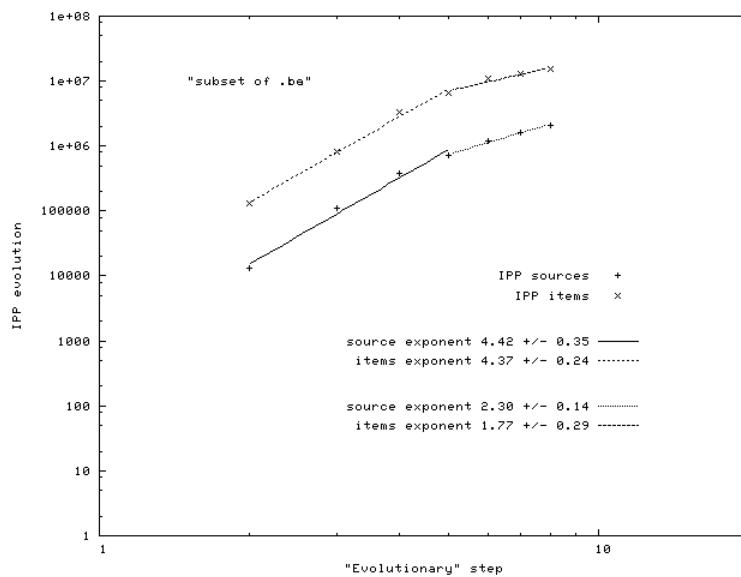


Figure A4. Evolutionary growth of sources and items for the Belgium sample Web IPP

High Productivity Physics Institutions in India: A Study of their Performance in terms of Quantitative and Qualitative Indicators¹

S.M. Dhawan * and B.M. Gupta **

* *smdhawan@yahoo.com*

Library & Information Consultant, Former Scientist F
National Physical Laboratory, New Delhi 110012 (India)

** *bmgupta1@yahoo.com, bmgupta@nistads.res.in*

National Institute of Science, Technology & Development Studies
Dr K.S.Krishnan Marg, New Delhi – 11012 (India)

Abstract

The paper analyses publications data in Science Citation Index - Expanded Version (SCIE) (Web of Science) for the period 1993 to 2001 for understanding high productivity institutions in India in physics, identifying their overall strength measured on absolute and relative indicators, and highlighting in particular their areas of specialization in different branches of physics. The paper lists suggestions for national policy formulation for growth and development of physics research in the country.

Keywords

productivity; quantitative indicators; qualitative indicators; India

Introduction

India has more than a century old tradition of creating and contributing to physics at the highest level, with a wider institutional base today, and possesses skills, knowledge for research activity. Physics research in India is an institutional activity organized under various sectors such as Universities & Colleges, Institutes of National Importance, R&D, and Industry, etc. Even though India has physics institutions spread through out the country, but its research activity is localized to select few institutions. Its high productivity institutions are relatively better placed. They possess highly qualified and skilled manpower, endowed with stronger network linkages, command superior research and technical; infrastructure and the-state-of-art research facilities. Broad characteristics of physics research and publications output have been studied by Dhawan and Gupta, etc. from time to time (Gupta, Bose, Rangarajan, and Chandrasekaran 1980), (Dhawan. and Arunachalam, S. 1998). (Dhawan 2002). (Dhawan and Gupta, 2001), (Dutta, Bidyarthi, 2000), (Dhawan and Gupta, 2004). Quality of Indian physics research has been studied by Dhawan and Gupta (2004) using journal impact factor and citations received per paper data in the Indian context. So far, no study has been undertaken on the contribution and impact of high productivity institutions.

Objectives

This paper studies the performance of high productivity institutions (HPIs) in physics by using absolute and relative publication indicators to understand their comparative strength and weakness in different areas of specialization. The study also examines how different models of institutional funding (i.e. Institutes of National Importance, Research Institutions, and Universities) influence performance in physics research.

Methodologies and Sources Used

This study is based on raw bibliographical publication data (along with their citations data) for the period 1993 to 2001, extracted and downloaded in February 2004 from the Science Citation Index - Expanded Version (SCIE) (Web of Science) published by Thomson-ISI. The delineation of the broad and narrow subject areas in research papers was based on classification of journals done by Thomson-ISI. The institutional performance was measured on absolute indicators such as: (i) average IF per publication, (ii) average citation per publication, and (iii) share of collaborative paper and international

¹ This study is derived from a project funded by the Office of the Scientific Advisor of the Government of India

collaborative papers. Besides, high productivity institutions (HPIs) which belong to three different institutional models of funding in India, differ significantly in terms of their strength, number of researchers they employed, and papers published by each. For valid inter-comparison, their performance was measured on relative indicators such as (iv) relative impact factor, (v) relative citation impact, (vi) collaboration index, (vii) international collaboration index, (viii) and relative specialization index (SI). Relative indicators measure performance relative to the group average. For example, relative impact factor per paper of 1.5 indicates that it is 0.5 times greater than the average (1.0) of all institutions in the group studied. Specialization index (SI) measures publications share of the institution in the sub-field compared to publications share of the given sub-field in the country output. Further, SI values above 0.5 indicate that the level of specialization is high, whereas between 0.2 and 0.5 and -0.2 and 0.2 represent level of specialization is above average or is just average, respectively.

Analysis & Results

As seen from the Web of Science, India had participation from 1307 institutions in physics research during 1993-01. Of these, 64 institutions were rated as high productivity institutions (HPIs) each publishing at least 100 papers during 1993-01. These HPIs accounted for 88% (23,835 papers) of the total publications output by India (27018 papers) during this period. Of the 64 HPIs, eight were Institutes of National Importance (INIs), 23 were Research Institutions (RIs), and 33 belonged to Universities & colleges (Univ).

Overall Performance of HPIs on Indicators

In this paper, the performance of high productivity institutions in physics research in the country is studied using a number of absolute and relative indicators. Publications productivity per HPI ranged between 100 and 2008 papers, with an average of 372.4 papers. One-third HPIs published research output above the HPI group average. Their average IF per paper ranged between 0.5 and 2.5 with an average of 1.49. Nearly 1/3rd HPIs showed average IF per paper above the group average. Their average citations per paper ranged between 0 and 20 with an average of 4.60 citations per paper. Nearly half of HPIs received above group average citations per paper. All HPIs participated in collaborative research, but only 1/8th stronger collaborative profile, publishing more than 50% share through collaborative research.

Institutional models of funding for research do seem to influence research performance. INIs (one of the models of institutional funding in India) topped in publications productivity (with an average of 704.1 papers per institute) and in average citations per paper (5.25) during 1993-2001. Research institutions (another model of institutional funding in India) ranked 2nd in publications productivity, but topped in terms of IF per paper (4.75). Universities (another model of institutional funding) ranked 3rd on these parameters (Table 1).

Table 1. Comparative Performance of INIs, Research Institutions and Universities

Type of Institution (Count of Insts)	TP 93-01	RPO	Av. Output 93-01	% TCP/TP	% TNCP/ TP	% TICP/ TP	Av. IF/ Paper	Av Citat/ Paper
<i>Inst of Nat. Imp (8)</i>	5433	3-98	704.13	42.41	27.52	20.22	1.34	5.25
<i>Res. Inst (23)</i>	11142	1-21	484.43	47.58	28.44	25.35	1.75	4.48
<i>Universities (33)</i>	7808	1-13	240.88	45.23	29.35	21.31	1.22	4.31

TP = Total Papers; RPO = Range of Papers Output; TCP = Total Collaborative Papers; TNCP = Total Nationally Collaborative Papers; TICP = Total Internationally Collaborative Papers

Specialization Index (SI) of High Productivity Institutions

Out of 64 HPIs, 58 achieved high SI values varying from 0.50 to 0.99. Of these, seventies achieved above average to high SI in one sub-field each. Forty institutions registered SI above average in multiple sub-fields. Twenty four HPIs achieved above average to high specialization in condensed matter physics, followed by nuclear physics (24), atomic, molecular & chemical physics (23),

crystallography (23), applied physics (21), particles & fields (20), spectroscopy (20), fluids & plasmas (18), optics (17), thermodynamics (13), astronomy & astrophysics (10), and acoustics (9). The details are shown in Table 2.

Conclusion

The present study provides significant inputs for identifying the strength of select top high productivity institutions, in terms of output, impact, and their specializations in different sub-fields of physics. The study clearly indicates that Universities & Colleges sector institutions is the weakest whereas research institutes and INIs are relatively stronger in physics research performance. The study also indicates that INI model of funding for research is more effective in giving performance in terms of quality and quantity.

It is suggested that India should constitute a Board under the Department of Science and Technology of Government of India for overall monitoring, coordination and management of physics research in the country. The Board will address issues such as setting up of research facilities common to scientists from several institutions, allocation of funds for setting up specialized facilities, the nature of specialized facilities required to be set up or improved, their location, and distribution of the funds for the purpose.

Institutional Abbreviations Used

Research Institutes - BARC = Bhabha Atomic Research Centre, Mumbai; CAT = Centre for Advanced Technology, Indore; IACS=Indian Association for cultivation of Science, Kolkata; IOP=Institute of Physics, Bhubaneshwar; JNCASR=Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore; MRIMMP=Mehta Research Institute of Mathematics & Mathematical Physics, Allahabad; NCL=National Chemical Laboratory, Pune; NPL=National Physical Laboratory, New Delhi; PRL=Physical Research Laboratory, Ahmedabad; SINP=Saha Institute of Nuclear Physics, RRL=research research Laboratory, Trivandrum; TIFR=Tata Institute of Fundamental Research, Mumbai

Universities & Colleges - DELHUD= University of Delhi; HYDEUH=University of Hyderabad; JADAUC=Jadavpur University, Kolkata; PUNEUP=University of Pune, Pune; SRIVUT = Sri Venkateswara University, Tirupati;

Institutes of National Importance - IISc-BANG=Indian Institute of Science, Bangalore; IIT-BOMB = Indian Institute of Technology, Mumbai; IIT-DELH= Indian Institute of Technology, Delhi; IIT-KANP= Indian Institute of Technology, Kanpur; IIT-KHAR= Indian Institute of Technology, Kharagpur; IIT-MADR= Indian Institute of Technology, Chennai; ISI=Indian Statistical Institute, Kolkata

References

- Gupta, B.M.; Bose, P.R; Rangarajan, K.S.; and Chandrasekaran, S. (1980) Physics research in India – A bird's eye view. Physics News 11 (2), 1-7.
- Dhawan, S.M. and Arunachalam, S. (1998) Physics Research in India, as reflected by INSPEC-Physics, 1990 & 1994. New Delhi; NISSAT, Department of Scientific and Industrial Research.
- Dhawan, S.M. (2002) Physics Research in India, based on Publications Output Indexed in INSPEC – Physics, 1998. New Delhi; NISSAT, Department of Scientific and Industrial Research.
- Dhawan, S.M. and Gupta, B.M. (2001) Physics research in India: A study of institutional performance, based on publications output. Paper presented at International Conference on Scientometrics & Informetrics, Sydney.,
- Dutta, Bidyarthi. (2000) Scientometric Study of Research Output in Physical Sciences in SAARC Countries since 1991. New Delhi, INSDOC (Associateship Thesis) (1998-2000).
- Dhawan, S.M. and Gupta, B.M. (2004) Comparative evaluation of Indian physics research: impact factor vs citations frequency. Current Science, 86 (9), 1194-1195.

Table 2. Specialization Index of 64 High Productivity Institutions

Physics Sub-Field	Institutes of National Importance	Research Institutions	Universities
<i>Acoustics</i>	IIT-MADR (0.98), IIT-KHAR (0.92), IIT-KANP (0.90), IIS-BANG (0.89), IIT-BOMB (0.88), IIT-ROOR (0.87), IIT-DELH (0.84)	NPL (0.75)	BOMB (0.83)
<i>Applied Physics</i>	IIS-BANG (0.86), IIT-KHAR (0.82), IIT-MADR (0.71), IIT-BOMB (0.70), IIT-KANP (0.68), IIT-DELHH (0.51)	NPL (0.91), IACS (0.82), RRL-TRIV (0.82), TIFR (0.78), NCL (0.77), BARC (0.72), CAT (0.71), IGCAR (0.58), SSPL (0.44),	PUNEUP (0.78), OSMAUH (0.74), COCHUC (0.58), IUCDAEF-IN (0.37), DELHUD (0.33), BANAUV (0.28), NSC (0.22)
<i>Astronomy & Astrophysics</i>	IIS-BANG (0.22)	IIA (0.99), TIFR (0.97), RMRI (0.94), PRL (0.91), SNBNCBS (0.26), IIS-BANG (0.26)	IUCAA (0.98), COLL-IT-VA (0.68), DELHUD (0.31), JADAUC (0.24)
<i>Atomic, Molecular, & Chemical Physics</i>	IIS-BANG (0.86), IIT-KANP (0.76), IIT-ROOR (0.61), IIT-BOMB (0.60), IIT-MADR (0.45)	BARC (0.92), IACS (0.88), JNSCASR (0.85), TIFR (0.72), NCL (0.57), IOP (0.49), PRL (0.44), IICT (0.44), IGCAR (0.42), BI (0.39), CAT (0.32)	NSC (0.77), HYDEUH (0.71), PUNEUP (0.48), BANAUV (0.42), VISVUS (0.55), PANJUC (0.44), JAWAUD (0.23)
<i>Condensed Matter Physics</i>	IIS-BANG (0.86), IIT-KHAR (0.70), IIT-BOMB (0.59), IIT-KANP (0.55), IIT-MADR (0.42), IIT-DELH (0.26)	TIFR (0.81), IACS (0.70), NPL (0.70), BARC (0.65), SINP (0.59), JNSCASR (0.53), SSPL (0.52), IOP (0.44), CAT (0.33), IGCAR (0.43), NCL (0.33), SNBNCBS (0.25)	HYDEUH (0.44), OSMAUH (0.44), SRIVUT (0.53), PUNEUP (0.50), IUCDAEF-IN (0.45), SHIVUK (0.43), ANNAUM (0.43), DELHUD (0.23), CALCUC (0.23)
<i>Crystallography</i>	IIS-BANG (0.72), IIT-MADR (0.53), IIT-KHAR (0.49), IIT-KANP (0.34)	IACS (0.87), IICT (0.74), RMRI (0.62), BI (0.41), BARC (0.24)	MADRUM (0.98), MYSOUM (0.93), MADUUM (0.85), DRMAUV (0.85), ANNAUM (0.84), NORTUD (0.84), JADAUC (0.82), BHARUT (0.81), HYDEUH (0.80), JAMMUJ (0.75), GURUUA (0.54), NORTUS (0.48), ALIGUA (0.47), DELHUD (0.37), BURDUB (0.30)
<i>Fluids & Plasmas</i>	IIT-DELH (0.93), IIS-BANG (0.86), IIT-KANP (0.77), ISI (0.58), IIT-MADR (0.58)	IPR (0.98), PRL (0.79), JNSCASR (0.72), BARC (0.68), IMS (0.36), IGCAR(0.27), IACS (0.25), SINP (0.25)	JAWAUD (0.91), PUNEUP (0.57), BHARUT (0.59), JADAUC (0.42), DEVIUI (0.31)
<i>Nuclear Physics</i>	IIT-DELH (0.85), IIT-BOMB (0.42), ISI (0.29)	SINP (0.95), CVEC (0.94), IOP (0.89), BARC (0.87), TIFR (0.86), IMS (0.82), MRIMMP (0.79), PRL (0.44), SNBNCBS (0.39), BI (0.23)	JAWAUD (0.91), BHARUT (0.59), PUNEUP (0.57), JADAUC (0.42), DEVIUI (0.31)
<i>Optics</i>	IIT-DELH (0.96), IIT-MADR (0.89), IIT-BOMB (0.58), IIT-KANP (0.56), IIS-BANG (0.27), IIT-KHAR (0.24))	CAT (0.87), PRL (0.57), NPL (0.53), BARC (0.50), RMRI (0.23), IMS (0.22)	PANJUC (0.92), NSC (0.81), ALIGUA (0.71), JAMMUJ (0.43), DELHUD (0.55), KALYUK (0.53), JADAUC (0.48), CALCUC (0.44), UTKAUB (0.41), BANAUV (0.39), IUCAA (0.39), MYSOUM (0.21)
<i>Particles & Fields</i>	--	MRIMMP (0.98), TIFR (0.95), PRL (0.90), IMS (0.90), SINP (0.89), IOP (0.87), CVEC (0.53), BARC (0.51), RMRI (0.47), SNBNCBS (0.38), CAT (0.31)	COLL-IT-VA (0.96), COCHUC (0.85), CALCUC (0.87), DELHUD (0.79), SRIVUT (0.66), BURDUB (0.65), HYDEUH (0.63), ANNAUM (0.56), DELHUD (0.56), JADAUC (0.49)
<i>Spectroscopy</i>	IIT-MADR (0.67), IIT-KANP (0.42), IIS-BANG (0.41)	IICT (0.87), CAT (0.84), IACS (0.78), BARC (0.73), TIFR (0.48), IGCAR (0.39), NCL (0.24)	IUCAA (0.91), DELHUD (0.82), PANJUC (0.78), JADAUC (0.75), NORTUS (0.75), UTKAUB (0.73), NSC (0.48), NORTUD (0.41), CALCUC (0.40)
<i>Thermodynamics</i>	IIT-MADR (0.98), IIT-KHAR (0.90), IIS-BANG (0.86), IIT-KANP (0.86), IIT-BOMB (0.79), IIT-ROOR (0.68)	NCL (0.54), IICT (0.44)	MADUUM (0.92), NORTUS (0.87), BANAUV (0.84), SRIVUT (0.77), BURDUB (0.76), ALIGUA (0.55), COCHUC (0.46), KALYUK (0.29), PANJUC (0.25), BOMBUB (0.22)

Distributions of the h-index and the g-index¹

Leo Egghe *

*leo.egghe@uhasselt.be

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek (Belgium)²
and Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk (Belgium)

Abstract

In every scientific research area, each scientist has a unique h-index and g-index. This paper addresses the problem of determining the distribution of these indexes over the scientists. We apply aspects of linear three-dimensional Lotkaian informetrics to determine these distributions. We show that, supposing the article – citation information production process (IPP) to be Lotkaian with exponent α and supposing the scientist-article IPP to be Lotkaian with exponent α^* , we have that the scientist-h-index IPP and the scientist-g-index IPP are Lotkaian with exponent $\alpha\alpha^*$. This model is proved for discrete as well as continuous variables. This shows that the size-frequency distributions of the h-index and the g-index are very skew in general due to the generally high value of $\alpha\alpha^*$. We also calculate the rank-frequency distributions of the h- and g-index, based on the size-frequency distributions in the continuous variables case. Examples are given.

Keywords

H-index; Hirsch index; G-index; distribution; Lotka

Introduction

In Hirsch (2005), the physicist Hirsch introduced his so-called Hirsch-index (or h-index) as follows (using our own terminology – cf. Egghe and Rousseau (2006)): if we rank an author's papers in decreasing order of the number of citations they have received (publication and citation periods are fixed but arbitrary) then this author's h-index is the largest rank $r = h$ such that all papers on ranks $1, \dots, h$ have at least h citations each.

Since its introduction, the h-index has received a lot of attention, also in the informetric community, see e.g. Ball (2005), Bornmann and Daniel (2005), Braun, Glänzel and Schubert (2005), Egghe (2007), Glänzel (2006a,b), van Raan (2005), where also advantages and disadvantages of the h-index are described. We do not go into this topic here, but we only indicate one disadvantage of the h-index, which lead Egghe (see Egghe (2006a,b,c)) to his definition of the g-index: a clear disadvantage of the h-index is that, once an article is taken into account for the calculation of the h-index (i.e. once an article has a rank in the set $\{1, 2, \dots, h\}$) its actual number of citations (above h of course), now and in the future, is not taken into account. Egghe finds this a clear disadvantage for an overall performance measure of a scientist. Egghe notes that the h-index satisfies that the first h articles have at least h^2 citations, together but h is not necessarily the largest value with this property.

Therefore, in Egghe (2006a,b,c), Egghe defines the g-index to be the largest rank $r = g$ (we use the same ranking as above) such that the papers on ranks $1, \dots, g$ have at least g^2 citations, together. Clearly, $g^2 \leq h$ and in most cases, $g > h$. Examples in Egghe (2006b) show that the g-index has more "discriminative" power amongst scientists in a field but this topic is not further addressed in this paper. Also in this paper we suppose that the total number A of citations (to all papers of a scientist) is less than the square T^2 of the total number T of papers, so that g is always defined and also $g \leq T$. This property was always encountered in the examples in Egghe (2006b) but, when $A > T^2$, in Egghe

¹Acknowledgement: The author is grateful to Prof. Dr. I.K. Ravichandra Rao (ISI, Bangalore, India) for mentioning the problem studied in this paper.

(2006b), we also give a methodology to calculate g-indexes that are superior to T, by adding fictitious articles with 0 citations.

Having any scientific field, to be considered as a group of researchers, we can calculate, for each of them, a h-index and a g-index. One can then wonder what is the distribution of these h-indexes over the researchers and we can also ask the same question for the g-indexes.

In the next section we study this problem in the discrete setting (researchers have 1, 2, 3,... cited articles) and in the third section the same problem will be studied in the continuous setting (with densities of articles and citations). In both sections we prove the same theorem (exact in the case of continuous variables and with good approximations in the case of discrete variables): suppose that articles are cited according to a Lotkaian size-frequency function (or distribution) with Lotka exponent $\alpha > 1$ and suppose that authors publish articles according to a Lotkaian size-frequency function with Lotka exponent $\alpha^* > 1$, then the distribution (size-frequency function) of the numbers of researchers with a certain h-index or g-index is Lotkaian with exponent $\alpha\alpha^*$. Concrete examples are given that make this observation clear but, of course, also exact mathematical proofs are given. So, in general, such distributions have large Lotkaian exponents showing that their size-frequency functions are very skew (take e.g. the most common Lotka-exponents $\alpha = \alpha^* = 2$, then the distributions of the h-index and of the g-index are Lotkaian with exponent $\alpha\alpha^* = 4$ which is very high and leads to very skew (concentrated – see Egghe (2005), Chapter III) size-frequency functions for the distribution of the h- and g-indexes.

Based on the size-frequency functions of the h- and g-index we also determine the rank-frequency functions of the h- and g-index.

Size-frequency functions for the h- and g-index: discrete variables case

So we have a situation where researchers in a certain field produce articles and that these articles receive citations (after their publication). We restrict ourselves to those articles that received at least one citation. Publication periods and citation periods are fixed but – for this model – are arbitrary.

We suppose the article-citation IPP to be Lotkaian, cf. Glänzel (2006b), Egghe and Rousseau (2006), Rousseau (1997), Redner (1998), but this model can also serve as a first, simple approximation of other decreasing models for the number $f(n)$ of papers with n citations, as in Burrell (2007) or Redner (2005): we suppose that the number of articles with n citations equals ($n = 1, 2, 3, \dots$)

$$f(n) = \frac{C}{n^\alpha} \quad (1)$$

where $C > 0$ is a constant and $\alpha > 1$ is the Lotkaian exponent of this IPP.

Likewise, and even more classical – cf. Lotka (1926) – we can suppose the author-publication (articles) IPP to be Lotkaian: we suppose that the number of authors with T articles equals

$$\varphi(T) = \frac{D}{T^{\alpha^*}} \quad (2)$$

where $D > 0$ is a constant and $\alpha^* > 1$ is the Lotkaian exponent of this IPP.

While the number n of citations to articles can be arbitrarily large, we assume that authors produce a number of articles between 1 and T_{\max} . We further, classically, suppose that we have only one author producing the maximum number of articles (T_{\max}). Hence, using (2)

$$1 = \frac{D}{T_{\max}^{\alpha^*}}$$

whence $D = T_{\max}^{\alpha^*}$ and hence

$$\phi(T) = \frac{T_{\max}^{\alpha^*}}{T^{\alpha^*}} \quad (3)$$

Size-frequency function for the h-index

As proved in Glänzel (2006b) in the discrete case (approximately) and exactly in the continuous case in Egghe and Rousseau (2006), we have in case (1) when there are T articles in total (for a particular author):

$$T = \sum_{n=1}^{\infty} f(n) = \sum_{n=1}^{\infty} \frac{C}{n^{\alpha}} \quad (4)$$

that the h-index equals

$$h = T^{\frac{1}{\alpha}} \quad (5)$$

Combining (3) and (5) yields that, for each $T = 1, \dots, T_{\max}$, we have

$$\frac{T_{\max}^{\alpha^*}}{T^{\alpha^*}} \text{ authors with } h = T^{\frac{1}{\alpha}} \quad (6)$$

(hereby supposing that α is fixed, independent of T). This proves that the size-frequency function for the h-index: $\phi_1(h) =$ the number of authors with h-index h equals

$$\phi_1(h) = \frac{T_{\max}^{\alpha^*}}{h^{\alpha^*}} \quad (7)$$

Indeed: for each h-index $h = T^{\frac{1}{\alpha}}$, formula (7) gives a number of authors with this h-index equal to

$$\frac{T_{\max}^{\alpha^*}}{T^{\frac{\alpha^*}{\alpha}}} = \frac{T_{\max}^{\alpha^*}}{T^{\alpha^*}}$$

which is correct according to (3), since these authors have T articles.

Size-frequency function for the g-index

When a particular author has T articles we have (proved exactly in Egghe (2006b) in the continuous case) that the g-index equals

$$g = \left(\frac{\alpha - 1}{\alpha - 2} \right)^{\frac{\alpha - 1}{\alpha}} T^{\frac{1}{\alpha}} \quad (8)$$

Hence, using (3) and (8) we now have that, for each $T = 1, \dots, T_{\max}$, we have

$$\frac{T^*}{T} \text{ authors with } g = \left(\frac{\alpha-1}{\alpha-2} \right)^{\frac{\alpha-1}{\alpha}} T^{\frac{1}{\alpha}} \quad (9)$$

This proves that the size-frequency function for the g-index: $\varphi_2(g)$ = the number of authors with g-index g equals

$$\varphi_2(g) = \frac{T^* \left(\frac{\alpha-1}{\alpha-2} \right)^{(\alpha-1)\alpha^*}}{g^{\alpha\alpha^*}} \quad (10)$$

Indeed: for each g-index $g = \left(\frac{\alpha-1}{\alpha-2} \right)^{\frac{\alpha-1}{\alpha}} T^{\frac{1}{\alpha}}$, formula (10) gives a number of authors with this g-index equal to

$$\frac{T^* \left(\frac{\alpha-1}{\alpha-2} \right)^{(\alpha-1)\alpha^*}}{\left(\frac{\alpha-1}{\alpha} \right)^{\alpha-1-\alpha\alpha^*} T^{\frac{\alpha\alpha^*}{\alpha}}} = \frac{T^*}{T^{\alpha^*}}$$

which is correct according to (3) since these authors have T articles.

Both previous subsections show that the size-frequency functions of the h- and g-indexes are Lotkaian (with different constants in the numerator) with the same Lotkaian exponent $\alpha\alpha^*$, the product of the Lotkaian exponents of the article-citation IPP and of the author-article IPP.

This simple result also shows that the size-frequency functions of the h- and g-index are very skew or concentrated – see e.g. Egghe (2005), Chapter IV, Corollary IV.3.2.1.5, p. 204-205, since, usually, $\alpha\alpha^*$ is a large Lotka exponent: take e.g. the “classical” values $\alpha \gg \alpha^* \gg 2$ then $\alpha\alpha^* \gg 4$ which is extremely large (see the Lotka exponents described in the review subsection I.4 in Egghe (2005), p. 85-98).

The above results can also be proved – essentially – in the continuous setting. This will be done in the next section. The continuous results will also enable use to calculate the rank-frequency functions for the h- and g-index, which is not possible in the discrete variable setting.

Size- and rank-frequency functions for the h-index and g-index: continuous variables case

Size-frequency function for the h-index

We again have (2) but now for continuous variables T which we do not limit: $T \in [1, +\infty]$. Also, result (5) is exact, for every T , as proved in Egghe and Rousseau (2006). Hence we have that the density of authors with h-index h is proportional to (by (2) and (5)):

$$\varphi_1(h) \sim \frac{1}{h^{\alpha\alpha^*}} \quad (11)$$

We still have to normalise formula (11): we must have that

$$\int_1^\infty \varphi_1(h) dh = \int_1^\infty \varphi(T) dT \quad (12)$$

(= total number of authors).

Note that both h and T have 1 as minimal value. For T this is so because, with (2), we consider the number of articles as items, ranging from 1 (see Egghe (2005), Chapter II). Since $T = 1$ is the minimal value of cited articles, $h = 1$ is also the minimal value for h .

Defining the proportionality factor in (11) as E :

$$\varphi_1(h) = \frac{E}{h^{\alpha\alpha^*}} \quad (13)$$

we have (since $\alpha, \alpha^* > 1$) and by (2)

$$\frac{E}{\alpha\alpha^* - 1} = \frac{D}{\alpha^* - 1}$$

hence

$$E = D \frac{\alpha\alpha^* - 1}{\alpha^* - 1}$$

so that

$$\varphi_1(h) = \frac{D \frac{\alpha\alpha^* - 1}{\alpha^* - 1}}{h^{\alpha\alpha^*}} \quad (14)$$

Hence we again find Lotka's law with the exponent $\alpha\alpha^*$.

Size-frequency function for the g-index

Now (2) is still valid and also (8) is an exact result. Hence we have that the density of authors with g-index g is proportional to (by (2) and (8)):

$$\varphi_2(g) \sim \frac{1}{g^{\alpha\alpha^*}} \quad (15)$$

(we can omit D as well as $\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}$ in (15) since both are constants and since the normalising constant in (15) still must be determined). This goes as follows: as in (12), T starts in 1 but now, since

h (or T) starts in 1 and by (8), we have that g starts in $\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}$. Hence we have the requirement:

Total number of authors

$$= \int_{\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}}^{\infty} \varphi_2(g) dg = \int_1^{\infty} \varphi(T) dT \quad (16)$$

Defining the proportionality factor in (15) as F:

$$\varphi_2(g) = \frac{F}{g^{\alpha\alpha^*}} \quad (17)$$

we have (since $\alpha, \alpha^* > 1$) and by (2):

$$\begin{aligned} \frac{F}{1-\alpha\alpha^*} \left[g^{1-\alpha\alpha^*} \right]_{\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}}^{\infty} &= \frac{D}{1-\alpha^*} \left[T^{1-\alpha^*} \right]_1^{\infty} \\ &= \frac{D}{\alpha^*-1} \end{aligned}$$

Or

$$\frac{F}{\alpha\alpha^*-1} \left(\frac{\alpha-1}{\alpha-2} \right)^{\frac{\alpha-1}{\alpha}(1-\alpha\alpha^*)} = \frac{D}{\alpha^*-1}$$

Hence

$$F = D \frac{\alpha\alpha^*-1}{\alpha^*-1} \left(\frac{\alpha-1}{\alpha-2} \right)^{\frac{\alpha-1}{\alpha}(\alpha\alpha^*-1)} \quad (18)$$

so that the size-frequency function for the g-index is

$$\varphi_2(g) = \frac{D \frac{\alpha\alpha^*-1}{\alpha^*-1} \left(\frac{\alpha-1}{\alpha-2} \right)^{\frac{\alpha-1}{\alpha}(\alpha\alpha^*-1)}}{g^{\alpha\alpha^*}} \quad (19)$$

Formulae (7), (10), (14) and (19) are evidence for the following Theorem.

Theorem:

If the article-citation IPP is Lotkaian with exponent α and if the author-article IPP is Lotkaian with exponent α^* , then both the author-h-index IPP and the author-g-index IPP are Lotkaian (i.e. their size-frequency functions are Lotkaian) with exponent $\alpha\alpha^*$.

This represents a case where one regularly finds power laws with high exponents $\alpha\alpha^*$.

Example:

Let us take the “most classical” case that $\alpha = \alpha^* = 2$ (see Egghe (2005) for a treatment of this special case of Lotka exponents equal to 2). Let us take ($\alpha^* = 2$) (cf. (2))

$$\varphi(T) = \frac{100}{T^2}$$

(author-article size-frequency function). Since also the article-citation IPP has Lotka exponent $\alpha = 2$ we have here that (Glänzel (2006b), Egghe and Rousseau (2006)) $h = \sqrt{T}$ for every production T . We have

$$\begin{aligned} \text{for } T = 1 &: 100 \text{ researchers have } h = 1 \\ T = 2 &: 25 \text{ researchers have } h = \sqrt{2} \\ &\vdots \\ T = 10 &: 1 \text{ researcher has } h = \sqrt{10} \end{aligned}$$

Hence we have (cf. (7))

$$\varphi_1(h) = \frac{100}{h^4}$$

as size-frequency distribution. Indeed: $h = 1$ occurs with 100 researchers, $h = \sqrt{2}$ occurs with $\frac{100}{h^4} = 25$ researchers, ..., $h = \sqrt{10}$ occurs with $\frac{100}{h^4} = 1$ researcher.

Redner (1998) even reports on Lotkaian exponents $\alpha \gg 3$ for the article-citation IPP making $\alpha\alpha^*$ (most likely) to be even larger than four (probably around 6 if $\alpha^* \gg 2$)!

Similar examples can be given for the g-index (for $\alpha > 2$ now) based on (10).

Based on (14) and (19) we can also determine the rank-frequency functions of the h- and g-index, i.e. the functions $h(r)$ and $g(r)$ being the h-index (g-index respectively) at rank r .

Rank-frequency function for the h-index

From every size-frequency function $\varphi(j)$ one can derive the corresponding rank-frequency function $\psi(r)$ using the following Lemma (Exercise II.2.2.6, p. 134 in Egghe (2005) or Appendix in Egghe and Rousseau (2006) where also a proof is given)

Lemma:

The following assertions are equivalent:

- (i) $\varphi(j) = \frac{C}{j^\alpha}$ with $C > 0$, $\alpha > 1$ (constants) and $j \in [1, +\infty]$ (size-frequency function)
- (ii) $\psi(r) = \frac{B}{r^\beta}$ with $B > 0$, $\beta > 0$ (constants) and $r \in [0, T]$ (rank-frequency function), where T denotes the total number of sources. Moreover, the relation between the parameters are:

$$B = \left(\frac{C}{\alpha - 1} \right)^{\frac{1}{\alpha-1}} \quad (20)$$

$$\beta = \frac{1}{\alpha - 1} \quad (21)$$

When we apply the above Lemma to (14) (as φ) we find as rank-frequency function for h

$$h(r) = \psi_1(r) = \frac{D^{\frac{1}{\alpha\alpha^*-1}} \left(\frac{\alpha\alpha^*-1}{\alpha^*-1} \right)^{\frac{1}{\alpha\alpha^*-1}}}{(r(\alpha\alpha^*-1))^{\frac{1}{\alpha\alpha^*-1}}} \quad (22)$$

Rank-frequency function for the g-index

Similarly, based on the above Lemma and (19) we have the following rank-frequency function for g

$$g(r) = \psi_2(r) = \frac{D^{\frac{1}{\alpha\alpha^*-1}} \left(\frac{\alpha\alpha^*-1}{\alpha^*-1} \right)^{\frac{1}{\alpha\alpha^*-1}}}{(r(\alpha\alpha^*-1))^{\frac{1}{\alpha\alpha^*-1}}} \left(\frac{\alpha-1}{\alpha-2} \right)^{\frac{\alpha-1}{\alpha}} \quad (23)$$

Note that it follows from (22), (23), (5) and (8) that the source-rankings in (22) are the same as the source-rankings in (23) since the source on rank r in (22) has $h(r)$ as h -index and hence has

$$\begin{aligned} h(r) \left(\frac{\alpha-1}{\alpha-2} \right)^{\frac{\alpha-1}{\alpha}} \\ = g(r) \end{aligned}$$

as g -index. This is logical since, for any two sources A and B : $h_A < h_B \hat{\cup} g_A < g_B$, hence the rankings must be the same. So formula (23) also serves as a control for the correctness of our models.

Conclusions

We showed that the size-frequency functions (or distributions) of the h -index as well as the g -index (with respect to authors) is Lotkaian if we suppose the same for the size-frequency functions of citations (with respect to articles) and of articles (with respect to authors). Moreover the Lotka exponent for the h - and g -index distribution is the product of the respective Lotka distributions of citations and articles.

This also shows that we encounter here, in most cases, large exponents making the exponent 4 the “classical” value for the h - and g -index distributions (since exponents 2 are “classical” in the underlying cases).

Conclusions

We showed that the size-frequency functions (or distributions) of the h -index as well as the g -index (with respect to authors) is Lotkaian if we suppose the same for the size-frequency functions of citations (with respect to articles) and of articles (with respect to authors). Moreover the Lotka exponent for the h - and g -index distribution is the product of the respective Lotka distributions of citations and articles.

This also shows that we encounter here, in most cases, large exponents making the exponent 4 the “classical” value for the h - and g -index distributions (since exponents 2 are “classical” in the underlying cases).

References

- Ball, P. (2005). Index aims for fair ranking of scientists. *Nature*, 436, 900.
- Bornmann, L. & Daniel, H.-D. (2005). Does the h-index for ranking of scientists really work ? *Scientometrics*, 65(3), 391-392.
- Braun, T., Glänzel, W. & Schubert, A. (2005). A Hirsch-type index for journals. *The Scientist*, 19(22), 8.
- Burrell, Q.L. (2007). Hirsch's h-index: a stochastic model. *Journal of Informetrics*, 1(1), to appear.
- Eghe, L. (2005). *Power Laws in the Informetric Production Process: Lotkaian Informetrics*. Elsevier, Oxford, UK.
- Eghe, L. (2006a). An improvement of the h-index: the g-index. *ISSI Newsletter*, 2(1), 8-9.
- Eghe, L. (2006b). Theory and practise of the g-index. *Scientometrics*, 69(1), 131-152.
- Eghe, L. (2006c). How to improve the h-index. *The Scientist*, 20(3), 14.
- Eghe, L. (2007). Dynamic h-index: The Hirsch index in function of time. *Journal of the American Society for Information Science and Technology*, 58(3), 452-454.
- Eghe, L. & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, 69(1), 121-129.
- Glänzel, W. (2006a). On the opportunities and limitations of the h-index. *Science Focus*, 1(1), 10-11 (in Chinese).
- Glänzel, W. (2006b). On the h-index – a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67(2), 315-321.
- Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569-16572.
- Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12), 317-324.
- Redner, S. (1998). How popular is your paper ? An empirical study of the citation distribution. *The European Physical Journal*, B4(2), 131-134.
- Redner, S. (2005). Citation statistics from 110 years of Physical Review. *Physics Today*, 58(6), 49-54.
- Rousseau, R. (1997). Citations: an exploratory study. *Cybermetrics*, 1(1), paper 1. <http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html>
- van Raan, A.F.J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgement for 147 chemistry research groups. *Scientometrics*, 67(1), 491-502.

Research and Application of Patent Map Analysis¹

Shu Fang ^{*,**}, Xian Zhang ^{**} and Guo-hua Xiao ^{**}

^{*}fangsh@clas.ac.cn,

School of Economics & Management, Southwest Jiaotong University, Chengdu 610031(China)

^{**}zhangx@clas.ac.cn, xiaogh@clas.ac.cn

Chengdu Library of The Chinese Academy of Sciences, Chengdu 610041 (China)

Abstract

Patent documents embed many important research results. But they often appear as legal documents covering the technology and business information wherein. Effective intelligence tools are necessary for patent information extraction. Patent Map (PM) provides visualized expression of total patent analysis results, describing the signification of all the charts, figs, graphs and so on, helping the non-specialists understand the patent analytical results easily and effectively. It is used as the starting point to investigate technology trend, and find out new emerging technology, valuable to both technologists and organizational managers or industry watchers. This paper discussed the typical representations of PM results about their functions, usual formats. A case study, patents activity analysis of the Vehicle Navigation either global or within China, was conducted by PM analyzing. Analysis results, including the development trends, technology mature degrees, technology structural distributions, major competitors, core patents, citation relationships, R&D hotspots, co-operations between major players, attributes in Chinese domestic market, etc., were presented in suitable visual maps.

Keywords

patent map; patent information analysis; patent intelligence analysis; intelligence analysis and synthesis

Introduction

As EPO disclosed, “patents reveal solutions to technical problems, and they represent an inexhaustible source of information: more than 80% of man’s technical knowledge is described in patent literature”. The investigation of existing patent literature can avoid research duplication. However, patent documents are often lengthy and rich in legal items. Data processing and comparative intelligence tools are necessary when obtaining valuable information from patent documents. Patent Map (PM) is such an intelligence tool having evolved from a simple concept to an important new discipline in intellectual property. Basing on clustering, aggregation, and other operations, it extracts the technological value from patents and provides visualized expression of total patent analysis results, describing the signification of all the charts, figs, graphs and so on. Used as the starting point to investigate technology trends, and find out new emerging technology, PM is valuable to both technologists and organizational managers or industry watchers. The functions, procedures, typical representations of PM results are discussed in this article. A case study, patents activity analysis of the Vehicle Navigation either global or within China, is conducted by PM analyzing, and many research results visualized in suitable maps are presented.

Background

By quantitative or qualitative analysis methods or both, the Administrative, technological and rights information is extracted to produce all kinds of charts, graphs, maps, etc. In general, the typical representations of patent map results are presented as follows (Jung, 2003) (JIII, 2000).

Quantitative analysis maps

They are results of quantitative analysis method such as quantity-based analysis, time-based analysis, and ranking analysis, etc., analyzing patents through numerical statistic. Most usable data come from bibliographical information including the number of patent applications, assignees, inventors, or patent classification codes, etc. The follows are the typical quantitative maps.

¹. This work was supported by West Light Foundation of The Chinese Academy of Sciences.

1. Portion rate map

Portraying composition ratios by some attributes such as applications, applicants, countries and so on, this map shows the structural difference within them. It is always represented as a pie graph, bar graph or ring graph.

2. Ranking map

It portrays the sort results in a specific field, highlighting the prominent objects such as identifying possible competitors or core patents, etc. Bar graph is more popular in this type.

3. Trend map

Depicting the development trend over time in a specific field, it is help for the future forecast. This kind of map is represented as broken line graph or bar graph usually.

Qualitative analysis maps

They are results of qualitative analysis method such as selection of core patents, citation analysis, technology development analysis, etc. They analyze the content of patents. Generally, they were performed by the inter-relationship of technology content or patent classification code, assignee, application date, etc. Typical qualitative maps are listed as following.

1. Matrix map

Matrix map shows the correlation between technical elements (such as purpose and technical item, problems and solvable technologies) obtained from patent information in the form of matrix. It helps to find important problems affecting the development of a technology field. It is even more effective to indicate the strength of the solvable technologies required for solving problems and consider the technical potential of developer. It is possible to estimate the degree of difficulty of realizing a development plan.

2. TEMPST map

TEMPST Map shows the technology analysis based on different points of analysis views. What the points of analysis views may include is listed as Table 1. In this map, the pertinent patents are classified by the analysis results of each point.

Table 1. TEMPST analysis view points

The point of analysis view		Example
<i>T</i>	<i>Treatment</i>	Temperature, Velocity, Time, frequency, Pressure, etc.
<i>E</i>	<i>Effect</i>	Purpose, Performance, Efficiency, etc.
<i>M</i>	<i>Material</i>	Material, Component, Compound, Addition, etc.
<i>P1</i>	<i>Process</i>	Manufacturing Methods, System, Procedure, etc.
<i>P2</i>	<i>Product</i>	Product, Parts, Results, Outputs, etc.
<i>S</i>	<i>Structure</i>	Structure, Form, Device, Component, Circuit, etc.

3. Citation analysis map

There are backward citation map and forward citation map (3i-Analytics, 2004). Backward map features patents being cited most frequently by patents in the field. These cited patents can be expected to cover a core technology or a valuable invention that is significant to the development of the technology. Forward map features patents citing the most number of patents among the field. These citing patents would help in mapping the trends and dependencies in the field.

4. Technology development map

Portraying the technological progress, this map discloses the expanse flow of technology from a basic patent. It is produced by extracting patents related to a specific field, layering and displaying them on time series. This map helps to grasp the source of the technology and ascertain the process by which technologies advanced. It can be used as an effective tool providing a hint for creating new ideas.

5. Claim point map

Portraying the claim points and their relationships of existing patents, claim point map is useful to identify the existing claim coverage in a specific field to investigate whether the new technology infringes (Shinmori, Okumura, Marukawa & Iwayama, 2004). It is critical in managing technical assets, planning R&D projects and avoiding the possible conflicts.

6. Component map

In component map, a product is portrayed while the key components are marked with the related patents respectively. About each component, the quantities of the relative patents and the key claims can be seen from this map. It is useful to regard the patent protection net about a specific product (JAPIO, 1998).

7. Portfolio map

This map produces multiple clusters where respective clusters contain similar technologies in terms of major attributes of interest (Ernst, 1998). It is expected that the portfolio map can detect the relationships that are not obvious in other patent classification systems. Portfolio map is used to identify technology portfolios of patentees, which provides guidance to the investment.

8. Landscape map

It portrays the concentration of patent documents on some themes (Yeap, Loo & Pang, 2003). In this map, each hill represents a concentration of patent documents of a related theme. The peak indicates a higher concentration while labels on the peak of each hill signify themes. Each black dot in the map represents a cluster of documents. The proximity between objects (patent documents and hills) in the landscape is directly related the strength of relationships between them.

9. Technology Vacuum Map

This map portrays a technology vacuum (missing area), indicating blank zones having great potential for prior occupation. It is useful in developing business strategy and making R&D projects. Technology vacuum map can be developed in two ways, dynamic and static. The static map represents a snapshot of technology vacuum at a particular point in time, while the dynamic map provides information on the changing pattern of technology vacuum over time (Yoon & Park, 2002). It is considered a starting point for discovering new emerging technologies.

Case study - Vehicle Navigation patent activities analyzing

Objectives of the study

The objectives of this study are to collect statistical data on vehicle navigation patents, estimate the development stage, discover the distributions of applications and applicants, weigh the roles that major patentees played, determine the core patents and R&D hotspots, explore the relationships between patents and corporations, and provide a better understanding of either global or domestic patent activities in this field. The analytical results are presented in suitable visual maps.

Databases and tools selection

Since the accuracy of a study largely depends on the quality of data sources and research tools, we first spent lengthy hours to locate the best suitable data source. After careful examination and repeated experimental searches, Derwent Innovation Index (DII) was decided to be used as the data collection for its worldwide content coverage and outstanding patent family registration, while the Derwent Analyst (DA) as the analytical tool for its powerful functions and compatibility with the DII data format. In study of Chinese patent activity, as the time lagging of DII's coverage, the SIPO official Web-based database was employed to insure the recall ratio.

Study process

With the selected database and tool mentioned above, this study was conducted as the following:

1. Pre-searched repeatedly and built search strategies basing the results returned.

2. Searched the DII and 13610 published records including granted patents and unexamined applications were found till June, 2006.
3. Downloaded these 13610 records and transferred them into the DA tool.
4. Cleaned all the records and analyzed them by DA.
5. Designed some self-programming for the analytical functions unavailable in DA.
6. In accordance with analytical results, performed suitable charts, diagrams and maps. As needed, some statistical data were processed by Microsoft Excel.
7. Only the records of inventions were employed in most of this study, except the developmental stage measuring in Chinese activity where analytical data covered three kinds of Chinese patent type: Invention, Utility Model and Industrial Design.

Research findings

At the end of series of database searches, data analysis, comparisons and PM drawings, some research results about vehicle navigation patent activities are summarized and presented in visual maps as follows.

Statistical information

As of Jun 2006, 13610 applications were published. The very first patent was a French application filed in 1957, followed by a gradual growth until 1993 with only 781 applications published as the 5.7% of the total. Only during the recent 10-year period time appeared a very substantial increase, indicating the rapidly growing in R&D activities. Fig. 1 shows the increasing trend of applications according to the earliest prior year. For the primary examination system, most applications filed in 2004 and 2005 have not been published yet, so the data about 2004 and 2005 is not of adequate statistical significance.

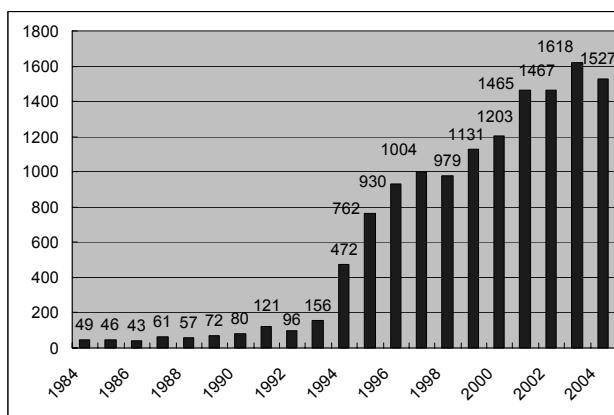


Figure 1. Vehicle Navigation Patent Applications Change Trend

Technological Life Cycle (TLC) analysis

TLC map is a plot of the number of patent applications versus the number of applicants over time (Liu, 2005). It portrays the maturity degree of a specific field. The cycle is divided into 5 periods (Fig. 2): emergence, development, mature, declining and recovery. Whether entering the recovery period is depended on whether the innovation elements appear. TLC map is useful in observing the stage of technological development, judging whether to go into the business of the technology field or not.

The TLC analysis results about global vehicle navigation technology are illustrated in Fig. 3. The mid of 1990s was the beginning of the development period, as the number of patent applications and applicants began to increase dramatically. After 1999 even more players entered. But from 2002 the applicants began to decrease while the applications kept increasing. It seemed some major players conducted this field and a beginning of gradual transmission from development to mature period appeared.

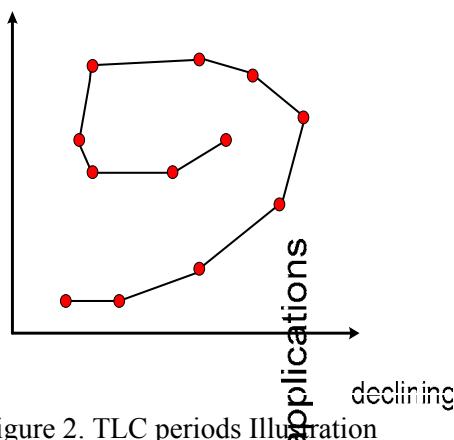


Figure 2. TLC periods Illustration

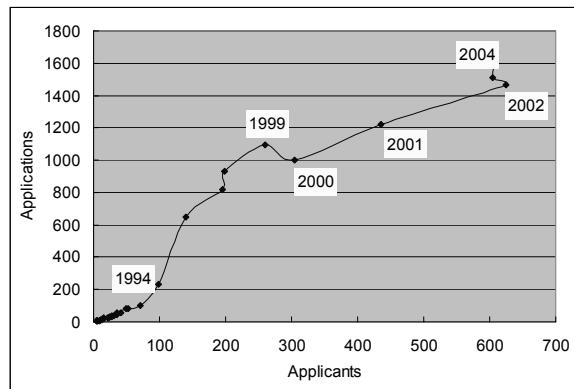


Figure 3. TLC map of Vehicle Navigation recovery

Country (regional) distributions analysis

The top 6 countries of patent applications in this field are Japan, US, Germany, Korea, France and English. 13144 applications filed by them, as 97% of the total (Fig. 4). Japanese applications advanced remotely as 75% of the total, suggesting their domination in the industry.

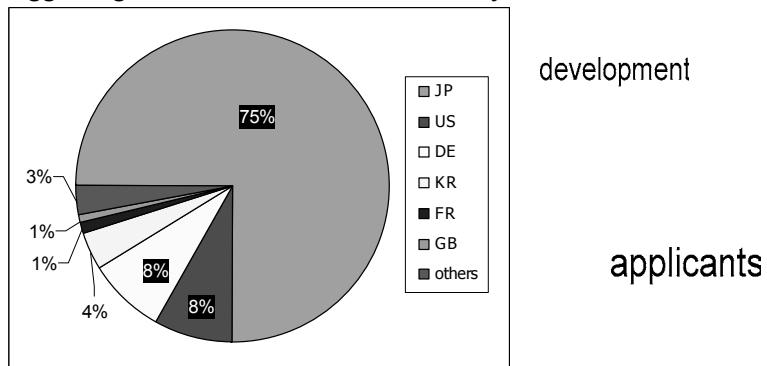


Figure 4. Country distributions of Vehicle Navigation Applications

Technology structure analysis

Employing the IPC system, the technological structure was investigated (Meyer, 2003). Top 5 IPC classes of applications are listed in Table 2. The majority applications are concentrated in 4 classes: G01C-021/00, G08G-001/0969, G09B-029/10 and G09B-029/00, indicating the main patent activities in this field focus on the technologies about navigation instruments, transmission control systems, locations, and map data storage.

Table 2. Top 5 IPC Classes about Vehicle Navigation Patents

Rank	IPC classes	No. of appl.	Class titles
1	G01C-021/00	8907	Navigation; Navigational instruments not provided for in groups G01C 1/00- G01C 19/00
2	G08G-001/0969	6711	Traffic control systems for road vehicles, having a display in the form of a map involving transmission of navigation instructions to the vehicle, giving variable traffic instructions, having an indicator mounted inside the vehicle, e.g. giving voice messages
3	G09B-029/10	4853	Map spot or co-ordinate position indicators; Map-reading aids
4	G09B-029/00	4602	Maps; Plans; Charts; Diagrams, e.g. route diagram
5	G08G-001/09	1419	Traffic control systems for road vehicles, arrangement for giving variable traffic instructions

Similar analyses could be done at both the macro and micro levels. For instance, an IPC structural comparison between major applicants helps to investigate these corporations' predominance in any unique field.

Major competitors analysis

Relative R&D ability (RRDA) was introduced to assess the levels, qualities and influences of the patent activities to determine the competitors of this field. The value of RRDA is achieved by $RRDA = NOP \times W_1 + SCI \times W_2 + OCI \times W_3$, wherein NOP means the number of patents, OCI means other-cited times, and SCI means self-cited times. In this study, W_1 , W_2 and W_3 were evaluated as 1, 1.2 and 1.4 respectively (Wu, 2003).

Regarding the maximal as 100%, the Top 10 are ranked (Table 3). 9 of them are from Japan except BOSCH GMBH ROBER from Germany, indicating Japan's absolute domination in this business. As for BOSCH, his RRDA ranks 6 although his application counts only ranks 11. This suggests his high innovation ability.

Table 3. RRDA Details of Major Competitors about Vehicle Navigation

RRDA Rank	Appl. Rank	Patentees	RRDA	Inventions	Cited Ratio	Tech. Independence
1	1	MATSUSHITA DENKI SANGYO KK (MATU)	100.0%	310	0.660	0.121
2	4	AISIN AW CO LTD (AISW)	82.8%	176	1.382	0.112
3	2	NIPPONDENSO CO LTD (NPDE)	80.3%	297	0.442	0.064
4	3	ALPINE KK (ALPN)	66.3%	166	0.386	0.100
5	9	MITSUBISHI DENKI KK (MITQ)	66.1%	236	1.893	0.062
6	11	BOSCH GMBH ROBERT (BOSC)	59.7%	348	1.892	0.164
7	8	NISSAN MOTOR CO LTD (NSMO)	58.2%	156	1.394	0.056
8	10	PIONEER ELECTRONIC CORP (PIOE)	54.7%	253	1.501	0.173
9	7	TOYOTA JIDOSHA KK (TOYT)	54.1%	170	1.136	0.059
10	14	HONDA GIKEN KOGYO KK (HOND)	46.2%	91	2.222	0.069

To investigate the R&D abilities in details, more indicators were employed in this section. For example, BOSCH had the most inventors so he held the largest R&D group of them; their cited ratios are closely to each other; the gaps of technological independences of them are not apparent.

The co-operation between the competitors was examined to investigate the penetration and interaction between major players of this field. As illustrated in Fig. 5: each node represents one assignee; the size of the node reflects the number of records associated with the assignee; these nodes are all the same because the assignees have a similar number of records (when compared to the total number of records in the dataset); the lines reflect the joint applications between the assignees; the strength of the lines is related to the number of joint applications.

From this map, Japanese corporations connected closely to each other. The co-operations between US, JP and EU are few. It seems that the European, American and Asian-Pacific R&D markets are kept under a system of its own without any connection on each other.

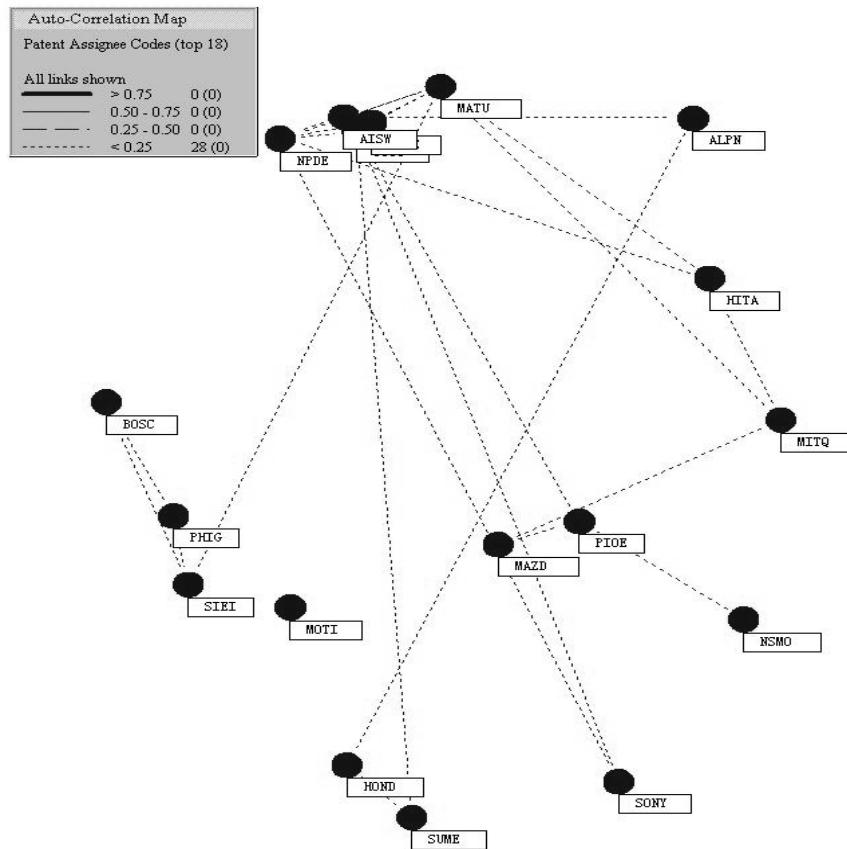


Figure 5. Co-operation Analysis of Major Competitors in Vehicle Navigation Field²

Core patents analysis

Although the earlier the patent is published the more possibly it is cited, the citation frequency is regarded as a key indicator to evaluate the patents' qualities, being used to identify the core patents of a specific field usually (Ernst, 2003).

In this study, the given patents were ranked on their citation frequencies. The top 10 are regarded as the core patents of vehicle navigation technology field (Table 4). They are all US patent documents while 4 filed by American corporations and 6 by Japanese corporations. Because of the important R&D position and high-tech market of America, Many developers regard America as the preferred country when seeking patent protection for their key innovations. On the other hand, 60% of the core patents are owned by Japanese players. It suggests Japanese corporations' core status in this field instead of the periphery in other fields usually.

Citations analysis

To investigate the technology advancement, the core patents were kept as sources and their citation process over time were explored. In this citation advancement map (Fig. 6), 5 bigger nodes are the core patents. The other smaller ones are citing patents. Arrows represent the citation relationships pointing to the cited patents. As there are too many following citing patents to show in one map completely, some main citing ones are showed in this illustration.

² Mapped by Derwent Analyst

Table 4. Times Cited Top 10 of Vehicle Navigation Patents

Rank	Pat. Num.	Title	Cited Times	Other Cited	Assignee	Appl. Date.
1	US 5177685	Automobile navigation system using real time spoken driving instructions	90	89	MASSACHUSETTS INST TECHNOLOGY (MASI), US	1990.8.9
2	US 4796191	Vehicle navigational system and method	64	61	ETAK INC (ETAK-Non-standard), US	1984.6.7
3	US 4926336	Route searching system of navigation apparatus	63	61	AISIN AW CO LTD (AISW); SHIN SANGYO KAIHATSU KK (SANG-Non-standard), JP	1988.12.27
4	US 5272638	Systems and methods for planning the scheduling travel routes	62	62	TEXAS INSTR INC (TEXI), US	1991.5.31
5	US 4992947	Vehicular navigation apparatus with help function	55	49	AISIN AW CO LTD (AISW); SHIN SANGYO KAIHATSU KK (SANG-Non-standard), JP	1988.12.27
6	US 5031104	Adaptive in-vehicle route guidance system	51	47	SUMITOMO ELECTRIC IND CO (SUME), JP	1989.11.29
7	US 4608656	Road map display system with indication of a vehicle position	50	47	NISSAN MOTOR CO LTD (NSMO), JP	1982.4.2
8	US 4782447	System and method for navigating a vehicle	49	45	NISSAN MOTOR CO LTD (NSMO); NILES PARTS CO LTD (NILE-Non-standard), JP	1986.3.28
9	US 4937753	Route end node series preparing system of navigation apparatus	49	46	AISIN AW CO LTD (AISW); SHIN SANGYO KAIHATSU KK (SANG-Non-standard), JP	1988.12.27
10	US 5243528	Land vehicle navigation apparatus with visual display	48	45	MOTOROLA INC (MOTI), US	1990.9.12

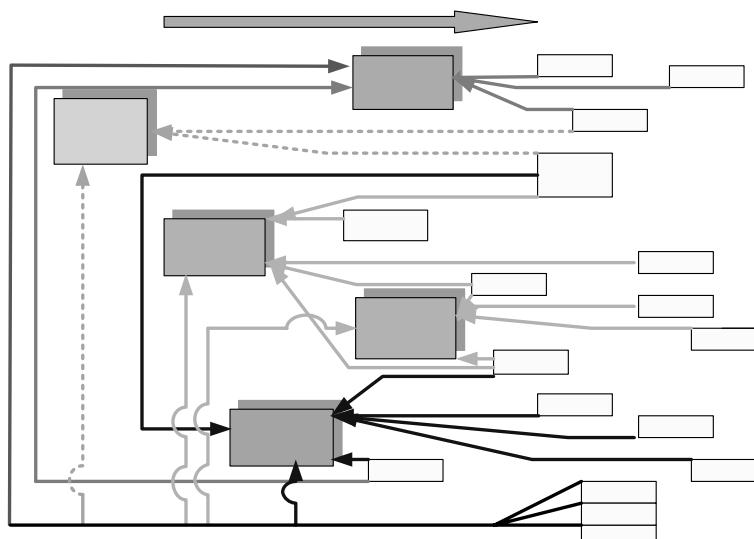


Figure 6. Citation Process Map of 5 Core Patents (Partly)

Basing the core patents, great technological families have developed by the citations. The members could be grouped on their similar subjects. The aggregations and dependences of them are plain to see in the technological family tree map (Fig. 7). Each node represents one patent. Arrows represent the citation relationship pointing to the citing patents. Some main citing ones are showed in this illustration.

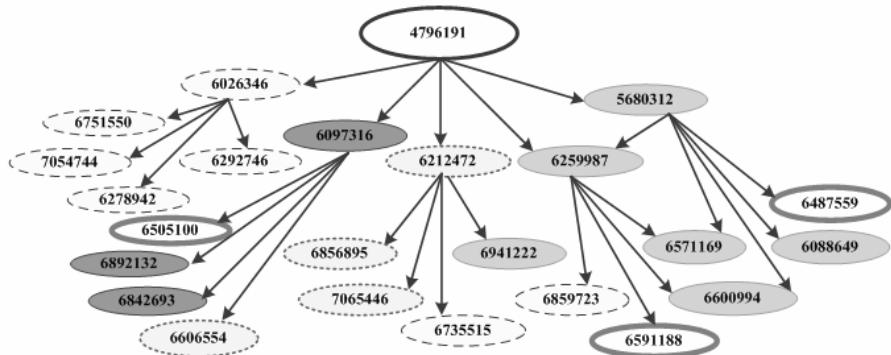


Figure 7. Technological Family Tree Map of US4796191 by Citation (Partly)

R&D Hotspots analysis

To find out the R&D hotspots, the average growth rates of IPC classes in recent 5 years were measured in this study. The top 5 are listed in Table 5, reflecting the hot technologies in this field recently.

Table 5. Average Growth Rates of IPC Classes in Vehicle Navigation Patents

Rank	IPC Classes	Aver. Growth Rate	Class Titles
1	G06T-011/60	520%	Two dimensional (2D) image generation by Editing figures and text or Combining figures or text
2	H04M-001/72	240%	Substation extension arrangements; Cordless telephones, i.e. devices for establishing wireless links to base stations without route selecting
3	G06T-003/00	220%	Geometric image transformation in the plane of the image, e.g. from bit-mapped to bit-mapped creating a different image
4	H04N-005/44	170%	Receiver circuitry of television systems
5	G06F-009/44	160%	Arrangements for executing specific programmes, using stored programme for programme control

G06T-011/60 ranks first by a growth rate of 520% per year while G06T-003/00 ranks third by 220%. Both are about image technology. It indicates that pictorial communication technology would be the hotspot in this field recently, wherein the patent activity should be further studied. Besides, major players' activities were focused to trace the new innovation directions in this study.

Claim points analysis

As the claims are the most important in patent specifications, claim points and their relationships of the key patents were examined. From the claim points map of US7126579-B2 (Fig. 8), the claims can be seen to categorize into 6 groups by their protected objects. The keywords and the relationships are illustrated.

*Patent activities in China**1. Technology developmental stage estimating*

In this section, the vehicle navigation technology developmental stage in China was estimated by employing 4 metric parameters of growth rate (V), maturity coefficient (α), aging coefficient (β) and innovation coefficient (N) (Du, 2005).

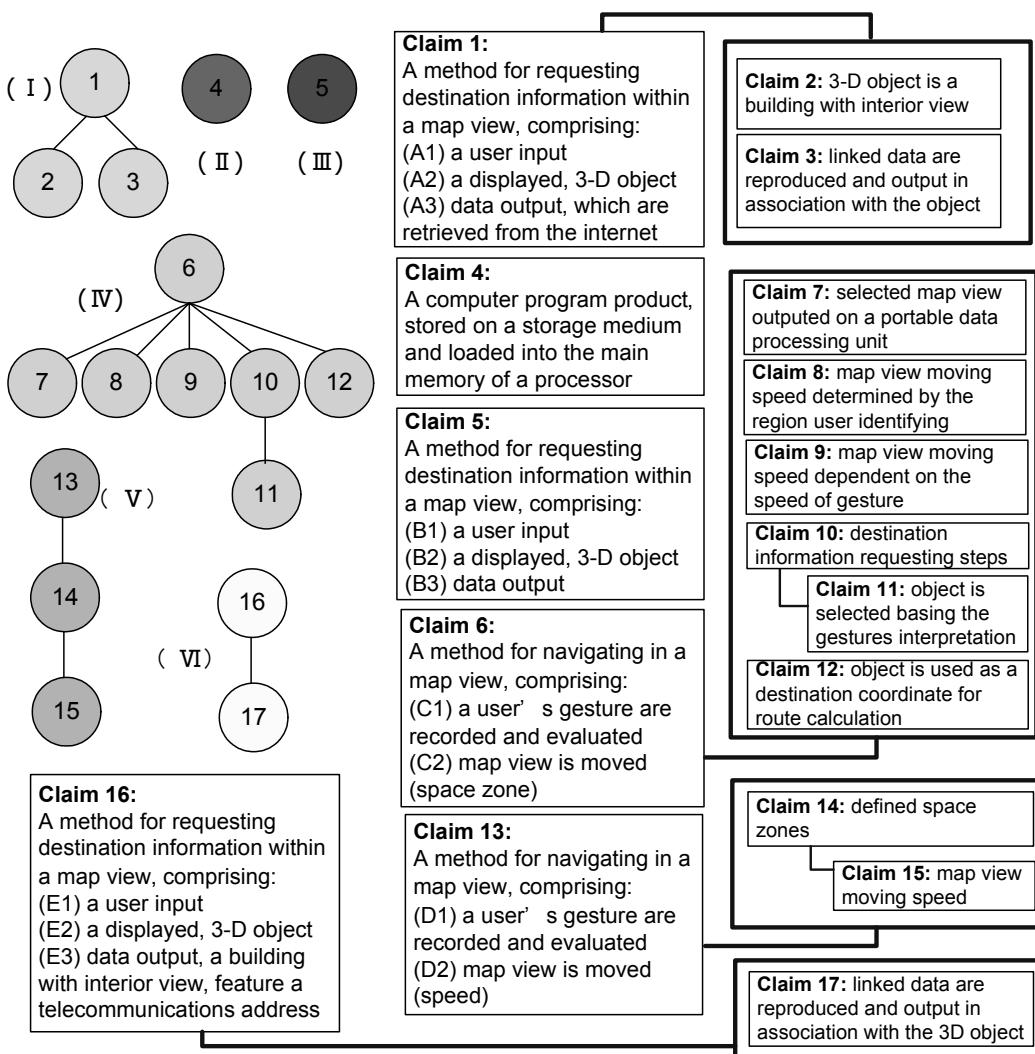
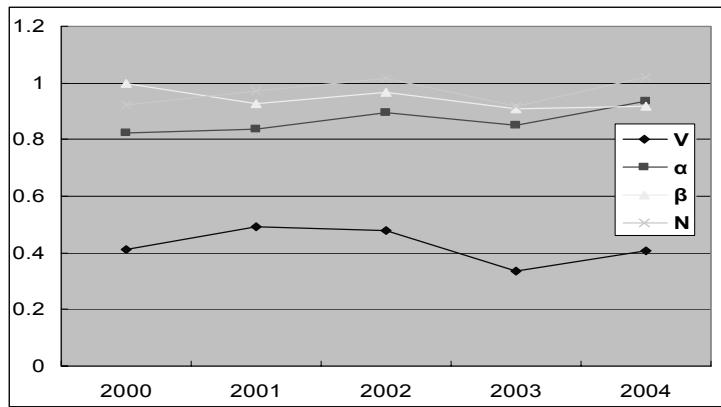


Figure 8. Claim Points Relationship Map of US7126579-B2

Table 6. Technology Developmental Stage Metric Parameters

Parameters	Formulae	Statistic Significance
<i>growth rate (V)</i>	$V=a/A$	an increasing V in several successive years indicates an emergence or growth in a specific field
<i>maturity coefficient (α)</i>	$\alpha=a/(a+b)$	a decreasing α in several successive years indicates a gradual maturity in a specific field
<i>aging coefficient (β)</i>	$\beta=(a+b)/(a+b+c)$	a decreasing β in several successive years indicates a gradual aging in a specific field
<i>innovation coefficient (N)</i>	$N=\sqrt{V^2+\alpha^2}$	The higher value of N indicates the more innovative in a specific field

Note: a: number of Invention applications (published) in a specific field in current year
b: number of Utility Model applications (published) in a specific field in current year
c: number of Industrial Design applications (published) in a specific field in current year
A: accumulation of Invention applications (published) in retrospective 5 years in a specific field

Figure 9. Value Changes of V、 α 、 β 、N over time

The measure results are illustrated in Fig. 9.

- The value of V increased from 2000 to 2001, followed by a decrease until 2003 when appeared another increase.
- The value of α kept rising except for a slight fluctuate in 2003, achieving a historic summit in 2004.
- Slight fluctuates appeared in the value of β .
- The value changes about N were similar to the V and α , keeping rising except 2003.

Above all, the vehicle navigation technology is still a new field in China, appearing a growing characteristic without any mature or aging evidence. It would continue to grow in both the short and long term. There would be more developing and marketing chances in the following years.

2. Applicants share analysis

Looking into the applicants of Chinese patents, the share of the foreign applications was around 85%. 4 Japanese corporations and 1 Korean corporation occupied the top 5 applicants, Mitsubishi Denki KK, Aisin Aw Co Ltd, Matsushita Denki Sangyo KK, Pioneer Electronic Corp, PIOE and Samsung Electronics Co Ltd. It seems that Chinese vehicle navigation technological market is dominated by the oversea companies.

To examine the original and independent innovation activities of domestic developers, the applicants were investigated (Fig. 10). The applications of mainland corporations seem to advance that of Taiwan and Hong Kong appreciably. The Universities and Academies' applications are few. A shortage of R&D in Chinese domestic developers appeared from this share. And the Industry, University and Academy should invest more in this field in future.

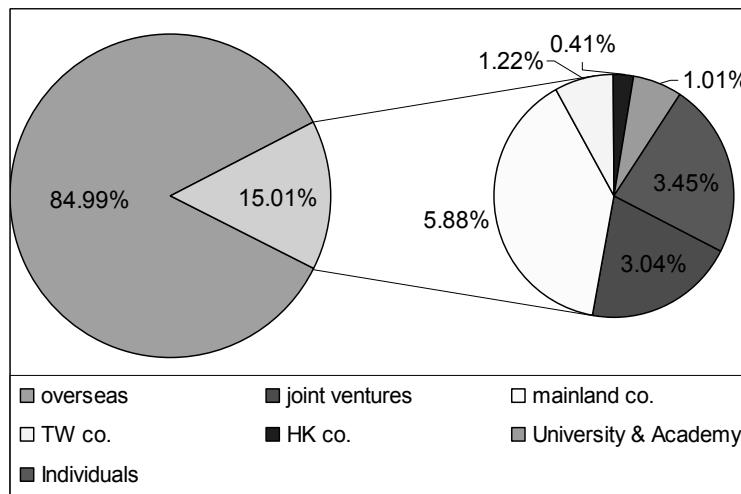


Figure 10. Share of Applications by Applicants in Chinese Patents

Conclusion and the Road Ahead

The PM and its applications were discussed in our research. A case study on Vehicle Navigation patent activities was conducted, with research results presented in suitable patent maps. Although outside the scope of our research project, it should be pointed out that this research could also form the starting point for further studies.

Existing PM, mostly relying on manual work or simple statistical tools, are limited in terms of explanatory capacity and operational efficiency since patent information encompasses numerous variables and the relationship between variables is so complex. Some improved patent analysis software should be developed to create maps that can satisfy human experts. Currently, there are some tools (Uchida, Mano & Yukawa, 2004) (Boyack, et al., 2000) (Trippe, 2003) (Eldridge, 2006) such as the DA provided by Thomson, the PM-Manager by WIPS, the PLAS by KIPO, the Patent-Lab by Delphion, and so on, making user approach to patent information easily and obtain optimized information. However, more sophisticated analysis and high-level descriptions should be incorporated into the automatic patent map generation as human-generated patent maps. A lot of research effort is required for the automatic generation system to be practically useful.

In addition, patent mapping is a true interdisciplinary skill requiring many skills, involving understanding the science, familiar to the bibliometric theory, being able to see business opportunities, a good understanding of patent law. Virtually it is difficult for one person to have all of these skills, so a “patent mapper team” appears important, which is the guarantee for the PM analysis results.

References

- 3i-Analytics. (2004). Technology patent maps. Retrieved March 9, 2006 from: <http://patentmaps.com/3i-sponsored.ipmaps/upc-196.pdf>
- Boyack, Kevin W. et al. (2000) Analysis of Patent Databases Using VxInsight. Workshop on New Paradigms in Information Visualization and Manipulation, Washington, DC (US), 12 Dec
- Du, X. (2005). Establish early warning mechanism of intellectual property. *R&D Management*, 17: 40-46
- Eldridge, J. (2006). Data visualization tools - a perspective from the pharmaceutical industry. *World Patent Information*, 28, 43-49.
- Ernst, H. (2003) Patent information for strategic technology management. *World Patent Information*, 25: 233–242
- Ernst, H. (1998). Patent portfolios for strategic R&D planning. *J Eng Technol Manage*, 14: 279-308
- Japan Patent Office Asia-Pacific Industrial Property Center, JIII. (2000). Guide Book for Practical Use of “Patent Map for Each Technology Field”. Retrieved February 12, 2006 from: http://www.okpatents.com/phosita/images/patent_map_JPO.pdf
- JAPIO. (1998). Patent map on washing technology for industry. Retrieved October 16, 2005 from: http://www.jpo.go.jp/shiryou/s_sonota/map/sennzyo/map/map10.htm
- Jung, S. (2003). WIPO/IP/BIS/GE/03/16: Patent map with exercises. Retrieved November 30, 2005 from: http://www.wipo.org/sme/en/activities/meetings/china_most_03/wipo_ip_bis_ge_03_16.1.pdf
- Liu, P., Wu, X. & Qi, C. (2005). Application of patent map in R&D management in enterprise. *R&D management*, 17: 47-52
- Meyer, M. Utecht, J. & Goloubeva, T. (2003). Free Patent Information as a resource for policy analysis. *World Patent Information* , 25: 223-231
- Shinmori, A. Okumura, M. Marukawa, Y. & Iwayama, M. (2004). Can Claim Analysis contribute toward Patent Map Generation. *Working Notes of NTCIR-4*, Tokyo, 2-4, June
- Trippe A. (2003). Patinformatics: tasks to tools. *World Patent Inform*, 25, 211-221.
- Uchida, H. Mano, A. & Yukawa, T. (2004). Patent map generation using concept-based vector space model. *Working Notes of NTCIR-4*, Tokyo, 2-4, June
- Wu, X. (2003) Packet switch architecture. Retrieved March 6, 2006 from: <http://www.gainia.com/>
- Yeap, T. Loo, G. & Pang. S. (2003). Computational Patent Mapping: Intelligent Agents for Nanotechnology. *IEEE proceedings of International Conference on MEMS, NANO and Smart Systems*, (pp.274-278)
- Yoon, B., Yoon, C., & Park, Y. (2002). On the development and application of a self-organizing feature map-based patent map. *R&D Management*, 32, 291-300

Bibliometric Study of Early Modern History in Spain Based on Bibliographic References in National Scientific Journals and Conference Proceedings¹

Francisco Fernández-Izquierdo*, Adelaida Román-Román, Cruz Rubio-Liniers**, Francisco-Javier Moreno-Díaz-del-Campo*, Carmen Martín-Moreno, Carlos García-Zorita***, María Luisa Lascurain-Sánchez***, Preiddy-Efraín García***, Elisa Povedano*** and Elías Sanz-Casado***

* fizquierdo@ih.csic.es, franciscoj.moreno@uclm.es

Instituto de Historia (Institute of History), CSIC (Spanish National Research Council), Department of Early Modern History, c/ Duque de Medinaceli 6, 28014 Madrid (Spain)

** adelaida@cindoc.csic.es, cruzrubio@cindoc.csic.es,

CINDOC (Centre for Scientific Information and Documentation), CSIC (Spanish National Research Council), c/ Pinar 25 28006 Madrid (Spain)

*** cmartin@bib.uc3m.es, czorita@bib.uc3m.es, mlascura@bib.uc3m.es, pegarcia@bib.uc3m.es,
epovedan@hum.uc3m.es, elias@bib.uc3m.es

Carlos III University, Dept of Librarianship and Information Science,
c/ Madrid 128, 28903 Getafe, Madrid (Spain)

Abstract

This study evaluates the historians' work, with a selection of 1,282 source papers published on early Modern History in Spain during 2000 and 2001 (417 articles published in 15 journals, and 865 conference papers included in 14 different proceedings, see references). They contained 44,471 bibliographic references citations (with a repetition factor of 1.59) plus 19,269 references to archive documents or manuscripts. Some conclusions are obtained in a first approach: Although conference proceedings accounted for a larger number of papers (2/3 of total) than journals (1/3), coming from a similar number of chosen proceedings and journals(14 - 15), the proceedings were not cited more frequently (5.54%) than journal articles (18.53%). Historians work usually alone, their cites are 61.5% monographs, and historical materials reached 15.59% of all citations. The vernacular languages, Spanish and Catalonian, together represented 72.50% of the citations, followed by French; other languages were more indicative of the subjects studied. The average age of the citations was fairly high, with the 50th percentile being around 16-17 years. Although a core of 111 journals was identified, dispersal was very wide, for the 7,805 articles cited appeared in 2,132 periodicals, 1,301 of which published only one of the cited articles.

Keywords

citation analysis in the humanities; early modern history; citation analysis in conference proceedings; Spain

Introduction

Humanities scholars' tendency to publish their papers primarily as monographs and in national journals, many without listings in ISI databases, constitutes an obstacle to the application of quantitative bibliometric indices (Sanz Casado and Martín Moreno, 1997; Sanz Casado et al., 1999, Moed, 2005:147-153; Coffey, 2006; Nederhof and Noyons, 2006). In late 2003 a proposal to study early modern Spanish history citations begun with journals published in 2000 and 2001 (with some exceptions of 1999, and delayed or biannual issues of periodicals) as source materials and, as a novelty, the proceedings of the most prominent international conferences published in the same two years. In addition to attempting to characterise historians' activity, the citation index obtained would afford researchers an indication of their visibility and provide an objective tool with which to evaluate scientific production.

¹ This paper contains a preview of the results of the research project titled *La Historia Moderna en España a partir de su bibliografía. Análisis y valoración de las citas en revistas y actas de congresos*, funded by the Spanish Ministry of Education and Science under references BFF 2003-09511-C02-01 and BFF 2003-09511-C02-02 and conducted jointly by Carlos III University, the Institute of History's Early Modern History Department and CINDOC (Information and Scientific Documentation Centre), of CSIC (National Spanish Research Council).

Methodology

The citations were extracted from 417 articles published in 15 journals, and 865 conference papers included in 14 different proceedings. Computer software was used to extract the citations, most in footnotes, from text files (converted from "PDF" formats or created by printed text digitisation and OCR). This software, developed entirely by Dr Fernández Izquierdo, consists in a series of highly effective PerfectScript scripts. Due to the heterogeneity of the output, standardisation, synthesis and clean-up protocols were applied to the data. The papers were classified later by two taxonomic criteria: language and document type.

In this regard, *ModernitasCitas*, the CSIC (Spanish National Research Council) Institute of History's open Internet portal for queries, comments and suggestions that runs on Windows Server with Filemaker Pro software, has been available on the Web since September 2004
<http://www.moderna1.ih.csic.es/emc/>

Results and discussion

Overview of the citing articles (source of citations)

Of the 15 conferences and 15 journals chosen, only 13, all proceedings, furnished over 40 articles each, together accounting for 71.61% of the total.

The 1,282 papers analysed were signed by 1,003 authors, indicative of a relatively low degree of multi-authorship: only 1.08 in the period analysed. Spanish authors were observed to follow patterns similar to those described by different authors in the past (Stone, 1982; WIBERLEY and JONES, 1994; Sanz et al., 2002): as Brockman et al. (2001) observed, they work jointly on research projects, but publish their papers individually.

Author sporadicity was high, with 762 authors publishing only one paper; 160 publishing two; 48 three; 20 four; 8 five; 3 six; one eight and one nine. Due to the specialisation in early modern history, the sporadicity rate in the sample analyzed was 55.26%, lower than the 75% found by Sanz et al. (2002) for the humanities in general, but higher than the 36% average reported by other authors (Schubert and Glänsel, 1991).

Bibliographic references(citations)

The 1,282 source papers contained 44,471 bibliographic references, net of the repeated references to a given article in the same citing paper initially counted by the extraction software: the mean repetition factor was 1.59 (see Table 1). An additional 19,269 citations were references to archive documents or manuscripts, i.e., 30.22% of the sum of the archive documents and published papers, a figure much higher than the 12.6% reported by Jones, Chapman and Woods (1972). No rigorous analysis or evaluation can ignore such a large volume of citations.

Monographs are the type of paper most frequently cited, followed at some distance by journal articles, for which figures similar to the findings reported by Sanz et al. (2002) in their study of Spanish history journals. These data concurred with the results of other studies with respect to the observation that humanities researchers prefer books to journals (Brockman et al., 2001, Knievel and Kellsey, 2005). Indeed, while they publish different types of papers, monographs clearly predominate, a tendency observed since the nineteen seventies (Bebout et al., 1975; Stone, 1982; Broadus, 1987). Given the focus on historians, the present study identified a specific category comprising monographs prior to 1830 and current editions of period studies, as well as other older sources (journals, press, legislation), which together accounted for 15.59% of the citations.

Citations in Spanish, the vernacular language, predominated, followed by French, a language with a long historiographic tradition in Spain (Table 2); these results are in line with other recent observations in this regard (Fernández Izquierdo and Moreno Díaz del Campo, 2006 recorded 73% in Spanish and 15% in French; Knievel and Kellsey (2005) found 80.5% of the citations in US history journals to be in English. Citations written in other languages were associated with the subjects studied or the nature of the

authors, but very far of 91% of English citations by Turkish authors in A&HCI (Al, U.; Şahiner, M.; Tonta, Y. 2006). References to period papers in Latin confirmed a practice characteristic of the humanities (Knievel and Kellsey, 2005).

Table 1. Type of document cited and disaggregation of historical materials

Type of document	freq.	%	percentile	subtype	freq.	%	freq.hcs.	%
monographs	27.349	61,50	61,50	contemp. monogs	20.717	46,59		
articles	8.286	18,63	80,13	hist. monographs			4.078	9,17
collective papers	5.476	12,31	92,44	hist. books			2.554	5,74
proceedings	2.464	5,54	97,98	editions				
theses	321	0,72	98,70	articles	8.255	18,56		
websites	168	0,38	99,08	historic articles			31	0,07
press	217	0,49	99,57	press	121	0,27		
legislation	177	0,40	99,97	historic press			96	0,22
cartography	7	0,02	99,98	legislation	1			0,00
magnetic media	6	0,01	100,00	historic legislation			49	0,11
Total (a)	44.471	100,00		modern./edited vers.			127	0,29
<i>Repeated citations(b)</i>	70.816				29.094	65,00	6.935	15,59
<i>Repetition factor (b/a)</i>	1,59							
<i>References to archive material</i>	19.269							

Table 2. Distribution of works cited by language

Language	freq.	%	percentile	Language	freq.	%	percentile
Spanish	30.139	67.77	67.77	Galician	62	0.14	99.83
French	3.396	7.64	75.41	Hungarian	45	0.10	99.93
English	2.926	6.58	81.99	Polish	8	0.02	99.95
Italian	2.486	5.59	87.58	Scandinavian langs	10	0.02	99.97
Catalonian	2.102	4.73	92.31	Turkish	5	0.01	99.98
Latin	1.293	2.91	95.21	Russian	3	0.01	99.99
German	1.226	2.76	97.97	Hebrew	3	0.01	100.00
Dutch	462	1.04	99.01	Asturian	1	0.00	100.00
Portuguese	303	0.68	99.69	Basque	1	0.00	100.00
				Total	44.471	100.00	

Although the publication dates of the materials cited (undated citations were disregarded) ranged across many centuries (a fact also observed by Tosete Herranz, 2002), two thirds of the references were less than 30 years old and the 50th percentile was around 1984. The average age of the papers when the citing paper was published – 16.98 years – is indicative of the intensity of research in recent decades.

A total of 15,892 different authors were cited in the 44,471 references; one had 230 citations; another 209; nine were cited at least 100 times; 2,435 were cited twice; and 10,219 only once. Moreover, 4,706 authors, or 10.76% of the total, were cited at least 50 times. Another index explored was the dispersal of the journals most consulted frequently by these researchers. The 7,805 articles cited were published in 2,132 journals, 1,301 of which published only one of the papers cited. The core consisted in 111 journals which together accounted for nearly 50% of the references (Figure3).

Table 3. Works cited by year of publication

Year pub.	freq.	%	percentile
500-1199	6	0.01	100.00
1200-99	1	0.00	99.99
1300-99	1	0.00	99.98
1400-99	70	0.16	99.98
1500-99	1206	2.82	99.82
1600-99	1045	2.44	97.00
1700-99	1319	3.09	94.55
1800-49	631	1.48	91.47
1850-99	1852	4.33	89.99
1900-09	476	1.11	85.66
1910-19	518	1.21	84.54
1920-29	498	1.17	83.33
1930-39	564	1.32	82.17
1940-49	991	2.32	80.85
1950-59	1723	4.03	78.53
1960-69	2576	6.03	74.50
1970-79	4736	11.08	68.47
1980-89	9401	21.99	57.39
1990-99	14408	33.71	35.40
2000	655	1.53	1.69
2001	52	0.12	0.16
2002	10	0.02	0.04
2003	5	0.01	0.01
Total	42744	100.00	

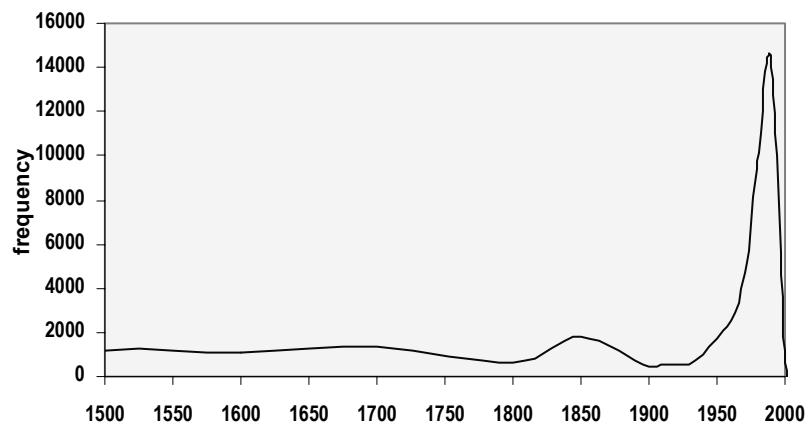


Figure 1. Distribution of citations by year of publication

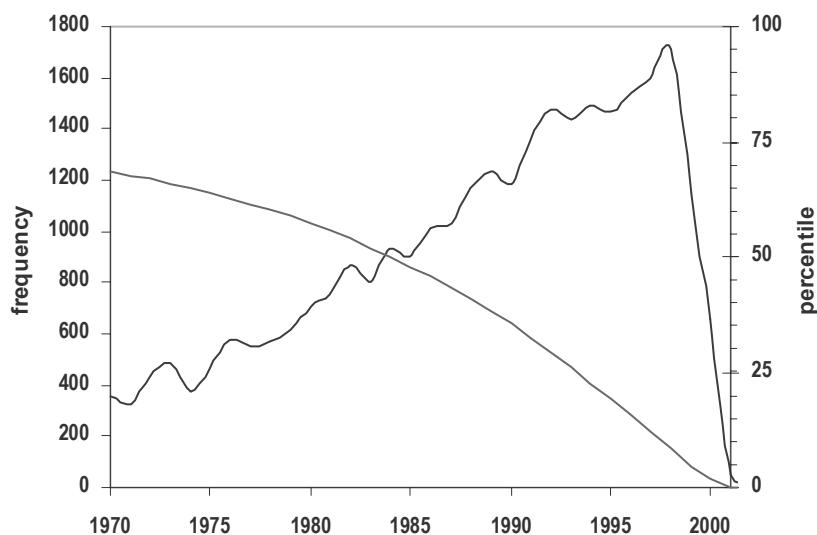


Figure 2. Citations: distribution and percentiles by year of publication (1970-2001)

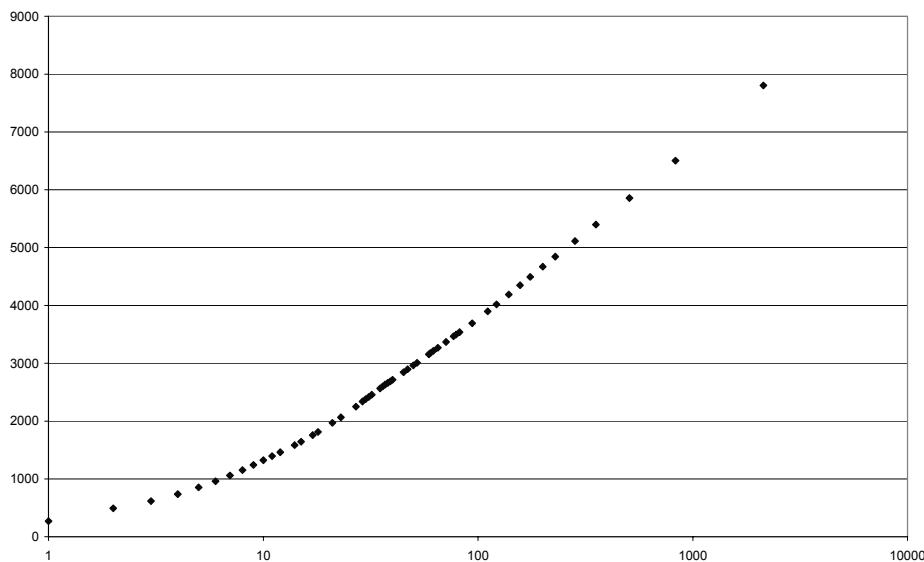


Figure 3. Dispersal of journals consulted.

Conclusions

- Although conference proceedings accounted for a larger number of papers than journals in a given period (a similar number of proceedings and journals was chosen: 14 - 15), the proceedings were not cited more frequently (5.54%) than journal articles (18.53%).
- Historians are prone to work alone, judging from the low rate of multi-authorship observed.
- Over 61.5% of the papers cited were monographs (monographs plus edited versions).
- Historical materials accounted for 15.59% of all citations, and the repetition factor found was 1.59.
- Not only printed paper: archive material and other sources would also need to be studied in citation analysis of History works.
- The vernacular languages, Spanish and Catalonian, together represented 72.50% of the citations, followed by French; other languages were more indicative of the subjects studied than of any effective internationalisation.
- The average age of the citations was fairly high, with the 50th percentile being around 16-17 years.
- Although a core of 111 journals was identified, dispersal was very wide, for the 7,805 articles cited appeared in 2,132 periodicals, 1,301 of which published only one of the articles cited. An analysis discriminating between Spanish and foreign journals should be expected to lead to results along the same lines as reported here with respect to the language of citations; another area that would need to be explored is the differences between citations from articles on the one hand and conference proceedings on the other.

References

- Al, U.; Şahiner, M.; Tonta, Y. (2006). Arts and humanities literature: Bibliometric characteristics of contributions by Turkish authors: Research Articles. *Journal of the American Society for Information Science and Technology* 57 (8, June 2006): 1011-1022.
- Bebout, L.; et al. (1975). User studies in the humanities: a survey and a proposal. *RQ*, 15 (1): 40-44.
- Broadus, R. N. (1987). Information needs of humanities scholars: A study of request made at the National Humanities Center. *Library and Information Science Research*, 9 (2): 113-29.
- Brockman, W. S; et al.. (2001). *Scholarly work in the humanities and the evolving information environment*. Washington, D. C.: Digital Library Federation; council on Library and Information Resources.
- Carr L.; Hitchcock, S.; Oppenheim, C.; et al. (2006): Extending journal-based research impact assessment to book-based disciplines (Research Proposal)
<<http://www.ecs.soton.ac.uk/~harnad/Temp/bookcite.htm>> [28-nov-2006]
- Fernández Izquierdo, F; Moreno Díaz del Campo, F.J. (2006). Análisis de citas en los artículos sobre moriscos de la revista *Sharq-Al-Andalus*. Tipologías, tendencias y reflejo historiográfico. *IX Reunión Científica de la Fundación Española de Historia Moderna, Universidad de Málaga, junio 2006*. (in print)
- Jones, C.; Chapman, M.; Woods, P.C. (1972). The Characteristics of the Literature Used by Historians. *Journal of Librarianship* 4, (3): 137-56.

- Knievel, J. E.; Kellsey, Ch. (2005). Citation analysis for collection development: a comparative study of eight humanities fields. *Library Quarterly*, 75 (2) 142–168.
- Moed, H.F. (2005). *Citation Analysis in Research Evaluation*. Springer, 350 pp.
- Nederhof, A.J.; Noyons, C.M., ed. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, 66 (1): 81-100 .
- Sanz Casado, E.; Martín Moreno, C. (1997). Técnicas bibliométricas aplicadas a los estudios de usuarios. *Revista General de Información y Documentación*, 7 (2): 41-68.
- Sanz Casado, E.; et al. (1999). La investigación española en Economía a través de las publicaciones nacionales e internacionales en el período 1990-1995. *Revista de Economía Aplicada*, 7 (20): 113-37.
- Sanz, E; et al. (2002). Creación de un índice de citas de revistas españolas de Humanidades para el estudio de la actividad investigadora de los científicos de estas disciplinas. *Revista Española de Documentación Científica*. 25 (4): 443-454.
- Schubert, A.; Glänzel, W. (1991). Publications dynamics: models and indicators. *Scientometrics*, 20 (1): 317-31.
- Stone, S. (1982). Humanities scholars: information needs and uses. *Journal of Documentation*, 38 (4): 292-313.
- Tosete Herranz, F. (2002). Midiendo la Historia Moderna: el impacto de la revista "Hispania" a través de las revistas universitarias de historia moderna españolas. *Hispania*, 210, p. 41-64.
- Wiberley, S. E.; Jones, W. G. (1994). Humanists revisited: a longitudinal look at the adoption of information technology. *College & Research Libraries*, 55: 499-509.

Appendix: Sources analysed

a) Journals:

Anuario de Historia del Derecho Español (Madrid); *Boletín de la Real Academia de la Historia* (Madrid); *Brocar* (La Rioja); *Chronica Nova* (Granada); *Contrastes. Revista de Historia Moderna* (Murcia); *Cuadernos de Historia Moderna* (Madrid); *Espacio Tiempo y Forma. Historia Moderna* (Madrid); *Estudis. Revista de Historia Moderna* (Valencia); *Hispania. Revista española de Historia* (Madrid); *Investigaciones Históricas. Época Moderna y Contemporánea* (Valladolid); *Manuscrits. Revista d'Historia Moderna* (Barcelona); *Obradoiro de Historia Moderna* (Santiago de Compostela); *Pedralbes. Revista d'Història Moderna* (Barcelona); *Recerques* (Barcelona); *Revista de Historia Moderna* (Alicante); *Studia Histórica. Historia Moderna* (Salamanca)

b) Proceedings:

- III Congrés d' Història Moderna de Catalunya* (1998). Held in Barcelona, 1998, pub. in periodical *Pedralbes* n. 18 (1998) - 2 vols, but printed in 2000.
- Rodríguez-San Pedro Bezires, Luis E., ed. (2000) *5º Congreso Internacional sobre Historia de las Universidades Hispánicas. Las Universidades Hispánicas, de la monarquía de los Austrias al centralismo liberal* (Salamanca, 1998, pub. Salamanca, 2000), 968 p.
- El emperador Carlos y su tiempo: Actas IX Jornadas Nacionales de Historia Militar* (2000) (Sevilla, 24-28 may 1999, pub. Madrid, 2000) 1.182 p.
- Bernal, A.M., ed. (2000) *Símposio Internacional "Dinero, moneda y crédito. De la Monarquía Hispánica a la integración monetaria europea"*. (Madrid, 4-7 may 1999, pub. Madrid, 2000), 898 p.
- López-Salazar Pérez, J., ed. (2000) *Las Órdenes Militares en la Península Ibérica, vol. 2 Edad Moderna* (International Conference, Ciudad Real, 1996, pub. Cuenca, 2000), 1.222 p.
- Martínez Ruiz, E. (2000) *Madrid, Felipe II y las ciudades de la Monarquía*. (Madrid, 1998, pub. Madrid, 2000), 3 vols. 1.561 p.
- Ferrer Benimelli, J.A. ,ed. (2000) *El conde de Aranda y su tiempo*. (2000) (Zaragoza, 2000), 2 vols. 1.522 p.
- Martínez Millán, J. y Carlos Reyero, coords. (2000) *El Siglo de Carlos V y Felipe II. La construcción de los mitos en el siglo XIX*. (Valladolid, 1999. Pub. Madrid, 2000.) 2 vols. 937 p.
- García Hourcade, J.L., coord. (2001) *Andrés, Laguna. Humanismo, ciencia y política en la Europa Renacentista*. (International Conference, Segovia, 1999. Pub. Valladolid, 2001), 578 p.
- Martínez Millán, J. y Ezquerra Revilla, I.J.; coords.(2001): *Carlos V y la quiebra del humanismo político en Europa (1520-1558)*. (International Conference, Madrid. 2000 – pub. 2001), 4 vol. 2032 p.
- Castellano, J.L., y Sánchez Montes González, F.; coords.(2001) *Carlos V. Europeísmo y universalidad*. (International Conference, Granada, 2000, pub. Madrid, 2001), 5 vols. 2.873 p.
- Belenguer Cebriá, E., coord. (2001) *De la unión de coronas al imperio de Carlos V*. (Congreso Internacional Barcelona, 2000, pub. Barcelona, 2001), 3 vol. 1611 p.
- Valdeón Baroque, J.; ed. (2001) *Isabel la Católica y la política. Ponencias presentadas al I Simposio sobre el reinado de Isabel la Católica*. (Valladolid, México, otoño 2000, pub. Valladolid, 2001), 421 p.
- Pereira Iglesias, J.L y Bernardo Ares, J.M. de; eds. (1999) *V Reunión Científica de la asociación Española de Historia Moderna: Felipe II y su Tiempo. La Administración Municipal en la Edad Moderna*. (Cádiz, 1998, pub. Cádiz 1999), 603+563 p.

The Effect of Patenting on the Networks and Connections of Academic Scientists¹

Enrico Forti*, Chiara Franzoni** and Maurizio Sobrero*

* *enrico.forti@unibo.it, maurizio.sobrero@unibo.it*

University of Bologna, Department of Management, via Capo di Lucca 34, 40126 Bologna (Italy)

** *chiara.franzoni@polito.it*

Polytechnic of Turin, DISPEA, Corso Duca degli Abruzzi 24b, 10129 Torino (Italy)

Abstract

Are academic inventors more or less connected in publication networks than their non-inventing peers? Are their networks sparser or denser? Are they bridges or hubs within their research communities? Does patenting alter their behaviour and role within the network along their careers? This paper tackles these specific questions, seeking to contribute to the analysis of the characteristics and the evolution of research collaborations of scientists who became academic inventors vis-à-vis those of their colleagues who never patented. Our empirical analysis compares two ego networks. The first one is based on 2899 scientific articles written between 1987 and 2006, together with 17853 co-authors by 55 Italian academic inventors, working in the field of Chemistry. The second one is based on 2406 scientific articles written between 1987 and 2006, together with 14562 co-authors by 55 Italian academics working in the field of Chemistry who never filed a patent in their career.

Keywords

academic patenting; network effect; technology transfer

Introduction

Increasing attention has been paid during the last years to academic patenting, both at a larger institutional level (Mowery et al., 2004; Murray 2002; Baldini et al. 2006), and at a more micro-individual level (Kogut and Gittelman, 2003; Murray, 2003; Baldini et al. 2007). Theoretically, the debate is centred around the possible rivalry between basic and applied research and the detrimental effects to the overall organization of scientific research within public institutions generated by the adoption of targeted output, more typical of private enterprises. This is coupled with an ever increasing attention by Universities and Public Research Organizations (PROs) all around the world for leveraging research results to foster technology transfer activities and generate new funding sources (OECD, 2003).

In particular, recent empirical evidence on the inventive activities of scientists working in academia shows that inventors, despite representing a small share of individuals in however considered scientific discipline or group (never exceeding the 10-15%), are overrepresented among the most productive individuals in science (Fabrizio and Di Minin, 2005; Stephan et al. 2005; Calderini et al., 2006). Furthermore, when the track of scientific publications are analyzed on longitudinal time-spans, the number of articles published after patenting increases (Azoulay et al., 2006), and this increase does not seem temporary (Breschi et al., 2005).

A possible explanation for this findings is that patenting, holding every other thing constant, enlarges the network of individuals whose knowledge is accessible by a scientist, which would result in newer ideas to exploit, larger cohort of peers to work with, offering additional or complementary knowledge to make research successful and, at the same time, a larger group of supporters to one's own ideas. Similarly, these findings could suggest that the presumed rivalry between patenting and publishing cannot be ascertained by the publishing behaviour of patentees. On the contrary, it points to the positive effect of patenting on knowledge diffusion through subsequent publication.

¹ This work is part of project iRis: Italian research on Innovation Systems. www.iris.unibo.it

On open-ended problem when counting publications is how to account for co-authored contributions. Co-authorship, in fact, does not say much about the true contribution of the single scientist to a research, nor of the complexity of the cooperative pattern that stand behind it. On the one hand, co-authored work might be needed to accomplish large-spectrum, multidisciplinary and complex tasks, requiring competences owned by no single scientist. In these cases, collaborations mean additive work and it could be argued that the contributions of each author can well equal the effort of writing alone. On the other hand, especially in those fields where research is conducted within labs, it is a common place to have honorary co-authorships, for instance of the principal investigator under whose supervision the group is working, or simply include long lists of authors to acknowledge for very diverse contributions in terms of effort, which would eventually be compensated along time (Peters and Van Raan, 1991; Lissoni and Montobbio, 2006).

Set aside the problem of how to interpret productivity effects in presence of multiple authors, a second aspect of co-authorship that co-publication analyses have largely stressed is that it makes visible (at least a part of) the set of competences and resources that are under the disposability of a single scientist: in other words, its network. In principle, the larger the network of a person, and the more diverse it is, the larger the pool of ideas and of open possibilities that he/she could fish in.

This property of networks to be channels of information and to make possible a combination and recombination of pieces of knowledge that would otherwise be distant, at a first glance, might offer good insight to understand the capacity of some scientists to become inventors of industrial applications, in addition to producing scientific advances. Academic inventors, indeed, are often described as bridging individuals whose role is to translate market needs into technically solvable problems or, conversely, envision why technologically feasible objects can offer value added to consumers (Allen, 1977; Etzkowitz, 1983). Hence, we can wonder if their role of arbitraging across the boundaries of science and market is somehow sustained by the access to wider spectra of information and competencies, including larger networks of relationships with otherwise unconnected individuals.

Yet, absent a detailed analysis of academic inventors publishing networks before and after their patenting experience and its comparison with the publishing network of their non-inventing peers, these analyses remain purely speculative. We intend to contribute to fill this gap with our study by offering an empirical analysis of the characteristics and the evolution of research collaborations of scientists who became academic inventors vis-à-vis those of their colleagues who never patented.

The paper is organized as follows. In the following section we start by summarizing the available evidence on the scientific productivity of academic inventors, we will then introduce the issue of network ties and the effects on the production of knowledge and finally set the research questions for the subsequent analysis. We then illustrate our research design, presenting the dataset, how we matched inventors with controls and the different indicators used in the analysis. The subsequent section is dedicated to the empirical analysis, with a detailed presentation of the results and their discussion. We conclude highlighting the contribution of our paper and some open questions for future research.

Scientific Productivity of Academic Inventors

Recent empirical evidence on the inventive activities of scientists working in academia shows that academic inventors, despite representing a small share of individuals in however considered scientific discipline or group (never exceeding the 10-15%) are overrepresented among the most productive individuals in science (Azoulay et al., 2006; Fabrizio and Di Minin, 2005; Stephan et al. 2005). The fact that the distribution of productive authors in science is strongly left-skewed, i.e. that a small number of authors (6%) is responsible for half of the overall publications, is well documented since the early 60s (De Solla Price, 1963; Allison and Stewart, 1974). Still, it is at first counterintuitive that an even smaller proportion of individuals in science seems to be capable at producing at the same time advances in the scientific understanding of principles and phenomena and applications of new technologies to industrial products.

This circumstance has raised the question of what kind of capabilities stands at the basis of a successful career in science and whether or not there are common drivers to explain the success of a scientist both in the academic and in the market world. The early years of Biotechnology, for instance, offered plenty of examples of eminent scientists that became famous for their technological applications and entrepreneurial spirit, while keeping a leading position in science (see, for instance, Zucker et al., 1998; Davies, 2001; Feldman et al., 2005).

Although in the traditional view of science as a speculative activity, scientific and entrepreneurial attitudes are traditionally seen as antonyms, at a closer look, there are several reasons to claim that this vision is oversimplified and obscures the true nature of the research work. First, it should be recognized that there are areas of investigation (the so-called “Pasteur’s Quadrant”) in which fundamental understanding and considerations of use can be pursued at the same time (Stokes, 1997). Hence, the potential gains and trades-off that a scientist might face in the pursuit of scientific and technological goals are not evenly distributed, across scientific disciplines, subfields, and topics (Calderini et al., 2007).

Second, the work of research itself is disseminated of technical problems to be solved (Franzoni, 2006). Scholars who study the creativity of scientists maintain that the rate-limiting factor of progress in science does not depend on the pace at which new ideas come to the researcher’s brains, but by the pace at which those ideas can be transformed into feasible operations at the bench (Holmes, 2004).

Third, successful scientists are often described as entrepreneurial by nature, as much as science is an inherently risk-taking activity and requires extensive organizational skills. As shown by the recent empirical evidence, research teams are steadily increasing in size (Adams et al., 2005) and, in the hard sciences, the budgets needed to equip university labs are growing as well (Ehrenberg et al., 2006). Successful research, of course, does not simply require attitude and preparation, but also extensive support and resources of various kinds: research funds, efficient division of labour, collaboration, influence and credit among the community of peers, to obtain the necessary validation and support to their research plans. This means that the job of scientists, once reached independence, is evermore one of managing a team of junior researchers and serve as brokers with the outside world (Murray, 2004; Franzoni and Lissoni, 2007).

Social capital and knowledge creation

Extensive studies on Social Capital and Network Theory have maintained the importance of the social dimension in knowledge creation (Coleman, 1988; Freeman, 1991; Auhja, 2000). Each scientist, of course, is engaged in a network of interpersonal relations, both working and purely social ones. Derek de Solla Price used to refer to “Invisible colleges” as the informal communities of scientists that interact closely by means of relationships, which are locus of cognitive formations advancing the state-of-art in research (1968). Relations bring information, for instance on the foreseeable evolution of technology and science, on the needs of firms and end-users markets, on the perceived importance of themes and topics, and so on. They overall bring competencies, in terms of potential solutions to problems.

Among the notable features of networks is that they allow reaching competencies that would otherwise be distant. If we imagine people as repository of (often idiosyncratic and tacit) knowledge and the interpersonal relationships as the links through which this knowledge can be exchanged, the degree of knowledge capital of each node increases with the size of the network and the intensity of the underlying relationships. Several studies have confirmed this idea, by finding a positive overall correlation among the collaborations and various indicators of quantity and quality of publications (see for instance De Beaver and Rosen, 1979; Kretschmer, 2004; Defazio et al., 2006).

A second argument to expect a positive correlation of productivity and a scientist’s social capital is linked to the internal approach to the world of science, extensively explored by Robert K. Merton and colleagues (see for instance Merton, 1957; Hagstrom, 1965), and refers to the idea that scientific theories, especially new and disruptive ones, can only be affirmed by raising a wide consensus among the community of peers. Hence, wide and frequent relationships within the community of reference

(and/or with contiguous communities, in the case of “boundary-spanners”) favor acceptance and, consequently, scientific performances (Allen, 1977).

On the other side, scholars assessing the impact of collaboration and networks have also stressed the dual effect of the intensity in exchanges: while cohesion generally supports productivity and trust, excess cohesion may in fact reduce creativity, because it limits the openness of the field to new ideas and lowers variety (Nahapiet and Ghoshal, 1998; McFadyen and Cannella, 2004). In contrast, loose and distant ties have a stronger potential to bring in set of resources otherwise unavailable to the single member, hence, low-density networks might be conveyors of higher achievements (Granovetter, 1983). In this respect, we should in principle expect the higher benefits to be obtained from participation in those networks that bring together individuals who are repository of very diverse pieces of knowledge, i.e. those characterized by bridging ties, where the nodes are necessarily not densely knit. While fragmented networks are overall more informative, within such networks the single egos that puts in contacts loosely coupled individuals has a higher status (Langlois, 1977).

Academic patenting and network effects: some research questions

So far very few works have carefully examined the networks of patenters and non patenters, while greater attention has been given to specific scientific communities (ex. Murray, 2004, Gittelman 2006), or to the contrast between academia and industry. On this specific topic an exploratory study by Meyer (2000, 2006) on the field of nanoscience and nanotechnology showed that academic inventors have networks of co-authors of scientific publications that are larger in size than those of academics who never patented an invention. This study however, only offers a visual inspection of the networks of academic-inventors. With regard to networks of co-inventors, Balconi et al. (2004) found that academics had higher between-centrality when compared to firm’s inventors and that the community of academic inventors was more sparse than that of non-academic inventors. They use these findings to suggest the existence of structural differences between peer-communities belonging to the realm of public and private research.

The aim of this work is to analyze the structure of the networks and the relative position of academic inventors, vis-à-vis their non-inventors colleagues. First, we make the hypotheses that academic inventors are capable of performing various tasks because they benefit from being part and/or being organizers of large teams. A large size of collaboration networks might favour scientists in various ways. Extensive team works and interdisciplinary collaborations, among the other things, serve the purpose of enlarging the set of problems that can be specified into sequences of operations and thus solved. Additionally, by means of collaborations, scientists acquire information on possible areas of interest, provide themselves access to funds (this is very clearly the case in Europe, where all communitarian funds until 2006 were targeted to cooperative research only), and favour acceptance to the research lines that they undertake. Each and all of these effects are particularly desirable for scientists that are moving along lines of research at the frontiers of academic fields, hence, in principle, we expect academic inventors to be found in higher proportions among those having networks with such properties.

This first part of the analysis is also interesting because it helps separating the effect of sheer productivity from that of networking. In principle, one of the reasons why academic inventors result to be among the most prolific authors could depends on them being in special positions, as head of labs and research groups, which allows for honorary co-authorship, rather than being more productive individuals. Here, by looking at the co-authors network, we will be able to separate size from intensity of interactions.

Second, we move from the observation offered by the descriptive literature which has portrayed academic inventors as highly connected individuals, who are gatekeepers of information from and across different pieces of fragmented communities (Allen, 1977; Etzkowitz, 1983; Murray, 2004). Shifting from the analysis of networks to egos, and, following the theory of weak ties (Granovetter, 1983), we expect academic inventors to be bridging individuals, benefiting from the fragmentation of the relatively unconnected networks of firms and academia. Here the bridging role takes two ways:

Murray (2004) says that the non-substitutable role of academic scientists working jointly in biotechnology firms was not only that of bringing competencies and ideas to develop new drugs, but also that of serving as providers of social capital to keep firms linked to the scientific community, an asset of focal importance in the field of biotechnology. Finally, we focus on the so called Anti-commons effect. More precisely, we check if the event of patenting itself is antecedent to changes in the network structure and ego-networks indicators, by comparing the network-derived indicators before and after the event of patenting to see whether or not the invention event is likely to generate changes in the network.

We expect the networks of academic inventors to differ from those of their non-inventor colleagues along three different dimensions. More precisely, we confront the structure of the network with regard to scale effects (ex. the overall amount of information being transferred), scope effects (ex. the diversity of information sources), and the role of direct vs. indirect access to information. We expect academic inventors to belong to larger networks, to have higher access to information sources and to be embedded in larger and more diversified groups, thus having higher opportunities to operate not only through strong, direct ties, but also through weak indirect ones.

Data

Academic inventors were identified using the PATUNIT database, which includes all patent applications filed by Italian Universities in Italy or abroad, directly or as extensions of patents already filed elsewhere. The dataset is complete from 1965 to 2004, as different disclosure procedure of patent applications around the world generate incompleteness for more recent years. While comprehensive at the institutional level, the dataset does not include all Italian academic inventors, as many may still be involved in patents filed by firms or other private institutions (Balconi et al., 2005). It includes, however, all academic inventors involved in patents filed by the University they belonged to at the time of discovery.

To control for interdisciplinary differences in patenting and publishing behaviour, we focused on the broadly defined Chemistry field, identified by using the classification established by the Italian Ministry of University and Research (<http://www.miur.it/UserFiles/116.htm>). A total of 59 individuals working in the field of Chemistry appeared as inventors of at least one patent filed by an Italian University. 9 were involved in more than three patents and 3 in more than five.

To identify Italian academics working in the same field at the same time to be matched to the inventors we relied on data collected by Baldini (2004). Data was collected by administering a questionnaire to a sample of Italian academics working in the same field of the Italian inventors appearing in the PATUNIT database, stratified by geographical area, field of studies and role. The questionnaire was structured in different sections and aimed at analyzing the motivations, incentives and obstacles to patenting faced by academics. For the purpose of our study, we selected all respondents in the Chemistry field as defined above who declared to have never filed a patent either in the name of their institution, or in the name of firms, or as individual inventors.

We adopted a matching pairs procedure starting from 59 treated individuals and 85 untreated potential pairs. The pre-treatment observable difference was accounted for by looking at a set of demographic (age, gender, location of affiliated university) and academic variables (owns a PhD, PhD obtained in Italy vs. abroad, subfield of Chemistry). We calculated the individual propensity score (probability of being an inventor) using a probit estimate and performed a one-to-one (nearest-neighbor) matching without replacement. Under a Caliper at .70, we dropped 4 treated IDs and remained with 55 inventors. The matching results in a mean difference probability of 19.52% (Std. Dev .2403), equalling a matching under Caliper at .67.

After completing the matching pairs procedure, we were left with a final sample of 110 individuals evenly split between inventors and (non-inventor) controls. Table 1 reports the variables comparing the ex-ante characteristics of the final sample. Around the 24% of all scientists sampled is in the most productive age-group, between 35 and 45 years old, while 46% of them is rather senior (more than 55 years old). The majority (over 50%) works in the north of the country and fewer (around 20%) in the

South, coherently with the country-level geographical distribution of Universities. Only about one third has a PhD, again coherently with the fact that PhD programs in the country were introduced only in the early 80s.

Table 1. Individual Heterogeneity. Matched-pairs comparison (Obs. 110)

Variable	Mean Inventors	St. Dev. Inventors	Mean Non Inventors	St. Dev. Non Inventors
<i>Year of Birth</i>	1952	9.48	1950	9.26
<i>Age: 35-45</i>	0.24	0.43	0.20	0.40
<i>Age: 46-55</i>	0.29	0.46	0.29	0.46
<i>Age: 56-65</i>	0.33	0.47	0.36	0.49
<i>Age: Over 65</i>	0.15	0.36	0.15	0.36
<i>Gender</i>	0.78	0.42	0.89	0.32
<i>North</i>	0.51	0.51	0.58	0.50
<i>Center</i>	0.33	0.47	0.20	0.40
<i>South</i>	0.16	0.37	0.22	0.42
<i>Phd</i>	0.35	0.48	0.24	0.43
<i>Phd in Italy</i>	0.33	0.47	0.22	0.42
<i>Phd Abroad</i>	0.02	0.14	0.02	0.14

At the end of this second step we identified two sets of egos, each composed of 55 scientists, to be used to generate their respective networks and analyze their relational behaviour. Ego-network data require three elements to be determined: the egos, in our case represented by the 55 academic inventors and their non-inventing “twins”, the alters, i.e. those who have some kind of relationship with the egos, and some measurement of these relationships. To identify all the alters we collected for both sets of egos all scientific publications present within the ISI Web of Science database, for an overall time span of 11 years, centred on the year of first patenting. Co-authorship was then used to determine a relationship between the egos and the alters.

After controlling for typical naming biases occurring within publications-based studies, our final data set consists of two ego networks. The first one is based on 2899 scientific articles written between 1987 and 2006, together with 17853 co-authors by 55 Italian academic inventors, working in the field of Chemistry. The second one is based on 2406 scientific articles written between 1987 and 2006, together with 14562 co-authors by 55 Italian academics working in the field of Chemistry who never filed a patent in their career.

After having completed the collection of publications, we merged our data with citation-based indicators on the characteristics of the journal on which the articles were published. The use of journal-based indicators as a proxy of articles’ characteristics holds several advantages and disadvantages that were heavily discussed by the scientometric literature, to which we refer for full information (see for instance Garfield, 2000; Glanzel and Moed, 2002).

We included in the dataset the Journal Impact Factor, from the ISI Journal of Citation Report, and the Journal Level Classification, from IpiQ, both calculated for the year 2003 (conventionally taken). The former indicator expresses the journal article’s short-term (2-years) average citations, whereas the latter expresses a ranking from 1 to 4 of the basicness of the journal, where 1 means very-applied and 4 means very basic (see Narin et al., 1978 and IpiQ, 2005). Additionally, we reported from the ISI publication records the total citations received by each article at the end of 2006 and computed the total number of co-authors for each paper.

Methods

We use network analysis techniques and measures to derive indicators of collaboration within the scientific community of our two populations of egos, the inventors and the controls. More precisely, we start from the ego-networks to generate an author-by-author affiliation matrix, where the authors are the 55 inventors, the 55 controls, and all co-authors of either of them. The event linking any pair of authors is a joint publication.

Algebraically, the resulting affiliation matrix is valued and symmetrical. Values in the cells along the main diagonal represent the total number of publications where the corresponding actor appears as an author between 1987 and 2006. Values off the diagonal represent the total number of publications between 1987 and 2006 where any couple of actors appears jointly as co-authors.

Affiliation networks are usually defined as two-mode networks, because they involve measurements of one set of actors and one set of events. They can be used to analyze the interaction between actors and events, as well as indirectly determine the relationships among the actors via the events, or vice-versa. Co-membership technically defines the joint participation of actors to the same event. In our case, co-membership substantively mean co-authorship, and we use co-authorship patterns to determine the characteristics of scientific networks of academic inventors and non-inventors.

Ego-network indicators are used in our analysis to compare the amount of relational activity within the scientific community of inventors and non-inventors. More specifically we focus on overall network size and density to determine the dimension of these communities and the extent to which all their members are also directly related to each other without the mediating role of the relevant ego. To control for possible biases of absolute measures of network density due to different network size, we will use relative network density instead. We then consider information centrality, using Freeman Betweenness Centrality index. Information centrality measures are so called to emphasize the role of intermediaries in the circulation of information and have widely used to study the flow of knowledge and communication within different kinds of communities. Technically, information centrality indexes are based on the geodesic distance between actors, measuring the shortest path connecting a couple of actors. The more an individual lies in between others on their geodesics, the more these latter ones will have to “pass through” her to get in touch.

We then analyze the structural patterns of relationship characterizing inventors networks and non-inventors networks through cliques. From the individual, we move to the overall network level to determine whether such networks differ in significant ways with respect to how all the actors involved are connected (or not). We focus on subgroup analysis to determine whether inventors and non-inventors belong to scientific communities characterized by more (less) cohesive subgroups, characterize by more (less) strong, intense, direct (indirect), frequent (spare) ties.

Formally, a clique at level c in valued graph is defined as a sub-graph in which the ties between all pairs of actors have a value of c or greater and there is no other actor outside the clique who's ties of strength c or greater to all actors in the clique (Wasserman and Faust, 1994: 279). Substantively, in our case a clique of level 3 would imply that any actor in the clique has co-authored at least 3 papers with all other actors in the clique. Inventors and non-inventors will be compared on the basis of the number of cliques to which they belong, the size (i.e. the number of actors included in the clique) and the level of such cliques.

Empirical analysis

We start our analysis by focusing on the publication-related indicators across the two paired groups of inventors and controls along the total observation period, i.e. the 11 years including the year of first patenting. Table 2 reports the average values for all variables and the results of the two-sample paired tests. For normally or log-normally distributed paired differences, a t-test was performed under the null hypothesis of zero mean difference. For variables that did not pass the Shapiro-Wilk test of normality or log-normality, we tested the null hypothesis that the sum of paired differences took a normal distribution with equal mean and variance, by running a one-way Kolmogorov-Smirnov test.

Table 2. Publication-related indicators over the entire observation period

Variable	Inventors	Non Inventors	Differences
<i>Total Publications</i>	52.71	43.75	*
	(35.35)	(30.21)	
<i>Total n° of Co-authors</i>	324.45	264.69	NS
	(235.8)	(243.83)	
<i>Average n° of Coauthors per Paper</i>	0.17	0.18	NS
	(0.04)	(0.04)	
<i>Total Citations</i>	696.73	581.15	NS
	(646.95)	(564.41)	
<i>Average Impact Factor</i>	2.22	2.10	NS
	(0.77)	(0.88)	
<i>Average Level</i>	2.92	2.88	NS
	(0.72)	(0.85)	

***p<.001 , **p<.05 , *p<.10

The results indicate that, after matching for observable heterogeneity, inventors had a higher mean number of published articles. Each one on average published 53 papers during eleven years of career, while controls published 44 papers. While the difference is not particularly high in absolute magnitude, this result is in line with the available empirical evidence, already commented in the previous section, showing higher scientific productivity by academic patentees. The quality of the paper, either measured in terms of journal's impact factor, (2.22 vs. 2.1), or as total citations received (686 vs. 581, but with over 90% variance), is however not significantly different. Moreover, the level of basicness of the publications of both groups is similar (2.9 vs. 2.8).

The non-parametric comparison of total number of co-authors across the whole time-span showed that inventors have larger values than controls. Inventors wrote papers on average with 324 different colleagues, while controls wrote papers with 265 different colleagues. The average number of co-authors per paper, however, is equal to 6 in both groups. We then considered the potential effects before and after the patent event. Before comparing the publication-indicators across inventors and controls, we checked if the passage of time brought variations in the indicators within each group of individuals, singularly taken. Table 3 reports the descriptive statistics for the publication-related variables calculated before the patenting event for the inventors and the controls. Table 4 reports the results of between group tests pre and post the event, as well as within group tests pre and post the event. Analyzing between group differences, we notice that before the event of the invention, the two groups have similar publication and co-authorship patterns, nor do they differ in terms of average impact factor or level of basicness of their publication. Inventors, however, have a significantly higher number of citations than controls. After the event, however, they published a higher number of articles.

All in all, the productivity of scientists tends to increase with the seniority, although not at a fixed rate, a reason why we chose to collect data for the same years for each pair of individuals, after having matched, among the other things, for the age of the individuals. This is confirmed by the evidence of increase in the total number of publications for both groups along time. A notable result is that, despite the increase in publications, the total number of co-authors resulted to be lower after the event for the subgroup of the inventors. This evidence seems to indicate that the co-authors have published more, but with a smaller number of co-authors (a circumstance confirmed by the negative value of the average number of coauthors). Quite surprisingly, both the average journal impact factor and the average level resulted to have lowered along the time for both inventors and controls.

Table 3. Pre and post between and within groups differences

Variable	Pre		Post	
	Inventors	Non Inventors	Inventors	Non Inventors
<i>Total Publications</i>	20.55 (14.94)	17.29 (13.61)	54.13 (28.90)	44.22 (20.32)
<i>Total n° of Co-authors</i>	119.20 (95.81)	101.36 (107.19)	177.13 (139.64)	138.00 (120.75)
<i>Average n° of Coauthors per Paper</i>	0.18 (0.06)	0.18 (0.06)	0.41 (0.28)	0.52 (0.58)
<i>Total Citations</i>	358.82 (357.90)	298.22 (358.36)	252.71 (231.91)	230.05 (236.27)
<i>Average Impact Factor</i>	2.09 (1.06)	1.91 (1.03)	2.32 (0.76)	2.28 (0.91)
<i>Average Level</i>	2.80 (0.87)	2.81 (1.06)	3.04 (0.76)	2.94 (0.75)

Table 4. Differences and significance

Variable	Difference Between Group (Inventors VS Non Inv.)		Difference Within Group (Pre VS Post)	
	Pre	Post	Inventors	Non Inventors
<i>Total Publications</i>	+	+**	-***	-***
<i>Total n° of Co-authors</i>	+	+	-***	-
<i>Average n° of Coauthors per Paper</i>	-	-***	-*	-
<i>Total Citations</i>	+*	+	+*	+*
<i>Average Impact Factor</i>	+	+	-***	-***
<i>Average Level</i>	-	+	-***	-***

*** $p < .001$, ** $p < .05$, * $p < .10$

We then analysed the differences in the networks indicators across the paired samples of inventors and controls. Table 5 reports the data calculated as before along the whole time-span, while Tables 6 distinguishes between the pre-invention and in the post-invention periods. With respect to network size, i.e. the total number of different co-authors registered in all publications, there is no significant difference between groups. This result, coupled with the fact that we previously showed that they publish on average more papers, seems to indicate that inventors tend to publish with a relatively stable number of co-authors. The ego-network density, however does not seem to differ significantly. This means that the co-authors of both inventors and control seem to be directly connected among them in a similar patterns. Turning to the analysis of betweenness-centrality indicators, as expected we find that inventors show larger values than controls and that the difference is statistically significant. Therefore, inventors have access to a higher amount of information within their communities than control.

When we turn to the comparison of the structural patterns of relationships within the two communities using clique-analysis, we also find, as expected, that inventors, on average, participate in a larger number of cliques and those cliques are, on average, larger in terms of number of components. Considering the structure of our data, this means that inventors participate in groups of researchers

who co-publish more than those associated with the controls. Moreover, these groups tend to be on average larger, while this did not emerge before from the simple comparison of the average number of co-authors per single publication.

Table 5. Network indicators over the entire observation period

Variable	Inventors	Non Inventors	Differences
<i>Size</i>	115.93	95.42	NS
	(94.26)	(95.67)	
<i>Density</i>	12.48	12.34	NS
	(9.70)	(8.77)	
<i>Cliques</i>	45.95	34.15	**
	(38.02)	(24.37)	
<i>Average n° of Authors per Clique</i>	6.73	6.24	**
	(1.30)	(1.20)	
<i>Normalized Betweenness</i>	1.82	1.29	*
	(1.20)	(1.71)	

***p<.001 , **p<.05 , *p<.10

Table 6. Pre and post between and within groups differences

Variable	Pre		Post	
	Inventors	Non Inventors	Inventors	Non Inventors
<i>Size</i>	50.36	42.25	72.49	59.62
	(46.03)	(46.72)	(57.87)	(58.13)
<i>Density</i>	25.31	24.20	18.40	19.27
	(22.20)	(21.39)	(14.64)	(15.84)
<i>Cliques</i>	16.60	14.69	23.71	20.04
	(14.11)	(15.06)	(18.73)	(15.99)
<i>Average n° of Authors per Clique</i>	6.15	5.71	6.88	6.28
	(1.56)	(1.65)	(1.52)	(1.48)
<i>Normalized Betweenness</i>	2.16	1.31	1.82	1.29
	(3.17)	(1.20)	(2.33)	(1.71)

Table 7 reports the results of between group tests pre and post event, as well as within group tests pre and post the event. Analyzing between group differences, we notice that overall the differences hold for both betweenness centrality and cliques, while for size before the event of the invention, inventors are shown to have a larger networks, but the difference is no longer statistically significantly after the event. Examining the variation of the different indicators within the two groups we also find that the dynamics are similar. Both the inventors and the controls increase over time the size of their network, their ability to exchange information and tend to participate in more complex and interconnected network. Coupled with the between group differences highlighted above these results suggest an underlying trait of the evolution of scientific communities which is similar regardless of individual applied productivity, although significantly different in magnitude.

Table 7 - Differences and significance

Variable	Difference Between Group (Inventors VS Non Inv.)		Difference Within Group (Pre VS Post)	
	Pre	Post	Inventors	Non Inventors
Size	+***	+	-***	-***
Density	+	-***	+**	+**
Cliques	+***	+***	-***	-***
Average n° of Authors per Clique	+	+**	-***	-**
Normalized Betweenness	+	+	+	+

***p<.001 , **p<.05 , *p<.10

Conclusions

The network structure of academic scientist has long been an area of considerable interest for social scientists for several reasons. Two have been particularly debated in recent years. If and to what extent a major involvement of scientists in the exploitation of research results could also affect their collegial behaviour within the scientific community, lowering the propensity to collaborate, to share preliminary results and holding possibly important publications to first comply with IPR rules and practices. If and to what extent individual scientific creativity and productivity could be better understood analyzing the role and dynamics of relational activities and the advantages of participating in larger, more diversified and more interconnected communities of scholars.

In this paper we tried to provide a first empirical contribution on what we believe are two sides of the same coin, namely the possibility to better understand individual behaviour by looking at her relational structure, applying social network theory to the realm of academic invention. To our knowledge, there are no papers which have so far collected detailed individual level data on scientists who also succeeded in filing patent applications, and scientists who never filed a patent in their whole career, and compared the two groups and their scientific communities.

Our results show that, in line with what has been previously found by other studies focused on the possible rivalry between patenting and publishing, inventors do not publish less or in less relevant journals than their colleagues who do not patent. Although we did not find significant differences in the quality of their publications measured by the IF, we did find that they published more and were more cited.

Analyzing their scientific community on the basis of co-authorship networks, we find new evidence going against a rival effect of patenting vs. publishing. Inventors network size, scope and structural patterns are larger, more connected and more complex than those of their colleagues who never filed a patent. Although both groups strengthen their network overtime, suggesting the presence of a seniority effect in the structuring of collaborative relationships, the difference among the groups do not disappear overtime, and the controls are unable to catch on with the inventor.

The sample we used offers different advantages, as it is geographically confined, to control for possible inter-individual differences due to contextual differences, as well as disciplinary homogeneous. Nevertheless, it suffers from some limitations. Patenters include only individuals who were inventors of patents filed by their University of affiliation. We cannot exclude that academic scientists, who appear as inventors in patents owned by private firms, show different publication behaviour, possibly more similar to scientists working in the private sector. Moreover, non-patentees where identified through a survey where they declared to never have filed a patent, either alone or as inventor for some other private or public institution. While we can think of no reason for providing us with wrong information in this matter, we cannot at the same time exclude that this might have happened. More generally, as it is well known in social network analysis, an accurate sampling at the individual level might not necessarily lead to stronger external validity with respect to the network results. This is particularly true when, as in our case, one looks at longitudinal data. While we

followed all the standard procedure normally used in this case and based on the normalization of all indexes and the definition of comparable network structure over the different time intervals, some problems with the computation of the different indexes may still remain. This is also the reason why we explicitly chose to rely on a limited number of indicators which have been proved to be less sensible to this computational problems.

Despite these possible shortcomings, we believe that our paper offers a unique and particularly original contribution to the debate of the characteristics and behaviour of academic scientists. Future work should extend our comparative static analysis to more formally include multivariate modelling of the phenomena we mapped and analyzed, as well as their dynamic evolution.

References

- Adams J., Black G.C., Clemons J.R., Stephan P.E. 2005. *Scientific teams and institutional collaborations: Evidence from U.S. universities, 1981–1999*. Research Policy, 34:259-285.
- Allen T.J. 1977. *Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information within the R&D Organization*. MIT Press.
- Allison P.D., Stewart J.A. 1974. *Productivity Differences among Scientists: Evidence of Cumulative Advantages*. American Sociological Review, 39:596-606.
- Ahuja, G. 2000. *Collaboration networks, structural holes, and innovation: A longitudinal study*. Administrative Science Quarterly. 45(3); 425-457
- Azoulay P., Ding W., Stuart T. 2006. *The Impact of Academic Patenting on the Rate, Quality, and Direction of (Public) Research*. NBER Working Paper #11917.
- Balconi, M., Breschi, S., Lissoni, F. 2004. *Networks of inventors and the role of academia: an exploration of Italian patent data*. Research Policy, 33(1):127-145.
- Baldini, N. 2004. *Cambiamenti istituzionali e processi innovativi: valorizzazione della ricerca universitaria italiana attraverso i brevetti*. Unpublished doctoral dissertation, University of Bologna, Bologna , Italy
- Baldini N., Grimaldi R., Sobrero M. 2006. *Institutional changes and the commercialization of academic knowledge: A study of Italian universities' patenting activities between 1965 and 2002*. Research Policy 35:518–532.
- Baldini, N., R. Grimaldi, et al. 2007. *To patent or not to patent? A survey of Italian inventors on motivations, incentives and obstacles to university patenting*. Scientometrics 70(2): 333-354.
- Breschi S., Lissoni F., Montobbio F. 2004. *Open science and university patenting: a bibliometric analysis of the Italian case*. Paper presented at 10th International J.S. Shumpeter Society Conference on Innovation, Industrial Dynamics and Structural Transformation Shumpeterian Legacies, Milan, 9-12th June 2004.
- Calderini M., Franzoni C., Vezzulli A. 2006. *If Star Scientists do not patent: The Effect of Productivity, Basicness and Impact on The Decision to Patent in the Academic World*. Research Policy, forthcoming.
- Coleman J.S 1988. *Social capital in the creation of human capital*. American Journal of Sociology, 94:95-120.
- Defazio D., Lockett A., Wright M. 2006. *The Impact of Collaboration and funding on productivity in research networks*. Mimeo.
- de Solla Price D.J. 1963. *Little Science, Big Science*. Columbia University Press, New York.
- Davies K. 2001. *Cracking the Genome. Inside the Race to Unlock Human DNA*. The Free Press, New York, NY.
- De Beaver D.B., Rosen R. 1979. *Studies in scientific collaboration Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite. 1799–1830*. Scientometrics 1(2):133-149.
- Ehrenberg R.G., Rizzo M.J., Jakubson G.H. 2006. *Who Bears the Growing Cost of Science at Universities?* Mimeo.
- Etzkowitz H. 1983. *Entrepreneurial Scientists and Entrepreneurial Universities in American Academic Science*. Minerva: 21:198-233.
- Fabrizio K.R., Di Minin A. 2004. *Commercializing the Laboratory: The Relationship Between Faculty Patenting and Publishing*. Haas School of Business Working Paper.
- Feldman M., Colaianni A., Liu K. 2005. Commercializing Cohen-Boyer 1980 – 1997. Mimeo.
- Franzoni C. 2006. *Do scientists get fundamental research ideas by solving practical problems?* Paper presented at the Conference “Innovation, Competition and Growth: Schumpeterian Perspectives”, Nice-Sophia Antipolis, 21-24 June 2006.
- Franzoni C., Lissoni F. 2007. *Academic entrepreneurship: definitional issues, policy implications and a research agenda*. In Varga A. (ed.). Academic entrepreneurship and regional development. Forthcoming, 2007.
- Freeman, C. 1991. *Networks of innovators: a synthesis of research issues*. Research Policy, 20 (5):499-514.
- Garfield E. 2000. *The use of JCR and JPI in Measuring Short and Long Term Journal Impact*. Paper Presented at Council of Scientific Editors Annual Meeting May 9, 2000”, <http://www.garfield.library.upenn.edu/papers/cseimpactfactor05092000.html>).

- Glanzel W., Moed H.F. 2002. *Journal impact measures in bibliometric research*. *Scientometrics*, 53(2):171-193.
- Granovetter M. 1983. *The Strength of Weak Ties: A Network Theory Revised*. *Sociological Theory*, 1:201-233.
- Hagstrom W.O. 1965. *The Scientific Community*. Basic Books Inc., New York, London.
- Holmes F.L. 2004 *Investigative Pathways. Patterns and Stages in the Careers of Experimental Scientists*. Yale University Press, New Haven & London.
- IpIQ (Intellectual Property Intelligence Quotient). *Journal Level Classification Refinement*. 5 November 2005, 222 Haddon Avenue Westmont, NJ 08108. Mimeo.
- Gittelman M., Kogut B. 2003. *Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns*. *Management Science*. Linthicum: Vol. 49, Iss. 4; p. 366
- Gittelman M. 2006. *National Institutions, public-private knowledge flows, and innovation performance: A comparative study of the biotechnology industry in the US and France*. *Research Policy*, 35(7).
- Kretschmer H. 2004. Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics*, 60(3):409-420.
- Langlois S. 1977. *Les Réseaux Personnels et la Diffusion des Informations sur les Emplois*. *Recherches Sociographiques* 2:213-245.
- Lissoni F., Montobbio F. 2006. *Co-inventorship and Co-authorship in Academic Science: Quantitative Analysis of Patent-Publication Pairs*. Paper presented at the Conference of the Society for Critical Exchange, Case Western University, Cleveland, OH, April 20.23, 2006.
- Merton R.K. 1957. *Priorities in scientific discovery: A chapter in the Sociology of Science*. *American Sociological Review*, 22(6):635-659.
- McFayden M.A., Cannella A.A. 2004. *Social Capital and Knowledge Creation: Diminishing Returns of the Number and Strength of Exchange Relationships*. *Academy of Management Journal*, 47(5):735-746.
- Meyer M. 2000. *Does science push technology? Patents citing scientific literature*. *Research Policy* 29:409-434
- Meyer M. 2006. *Knowledge integrators or weak links? An exploratory comparison of patenting researchers with their non-inventing peers in nano-science and technology*. *Scientometrics*, 68(3);545-560.
- Mowery D.C., Nelson R.R., Sampat B.N., and Ziedonis A.A. 2004. *Ivory Tower and Industrial Innovation: University-Industry Technology Transfer Before and After Bayh-Dole*. Stanford: Stanford University Press.
- Murray F. 2003. *The role of academic inventors in entrepreneurial firms: sharing the laboratory life*. *Research Policy*, 33:643-659.
- Murray F. 2002. *Innovation as co-evolution of scientific and technological networks: Exploring tissue engineering*. *Research Policy*. Amsterdam: Dec 2002. Vol. 31, Iss. 8,9; p. 1389
- Nahapiet J., Ghoshal S. 1998. *Social Capital, Intellectual Capital and the Organizational Advantage*. *Academy of Management Review*, 23(2):242-266.
- Narin F., Pinski G., Gee H.H. 1976. *Structure of the Biomedical Literature*. *Journal of the American Society for Information Science*, January-February, 25:45.
- Organization for Economic Cooperation and Development. 2003. *Turning science into business. Patenting and licensing at public research organizations*. OECD publications, Paris.
- Peters H. P. F., Van Raan A. F. J. 1991. *Structuring scientific activities by co-author analysis*. *Scientometrics* 20(1):235-255.
- Stephan P.E., Gurmu S., Sumell A.J., Black G. 2004. *Who's patenting in the university? Evidence from a Survey of Doctorate Recipients*. *Economics of Innovation and New Technology*, forthcoming.
- Stokes D.E. 1997. *Pasteur's Quadrant. Basic Science and Technological Innovation*. Brookings Institution Press, Washington D.C.
- Zucker L.G., Darby M.R. 1996. *Star scientists and institutional transformation: Patterns of invention and innovation in the formation of biotechnology industry*. *Proceedings of the National Academy of Sciences of the United states of America*, Colloquium paper, 93:12709-16.
- Zucker L.G., Darby M.R., Brewer M.B. 1998 *Intellectual human capital and the birth of US biotechnology enterprises*. *American Economic Review*, 88:290-306.

Using Content Analysis to Investigate the Research Paths Chosen by Scientists Over Time¹

Chiara Franzoni*, Christopher L. Simpkins **, Baoli Li **, Ashwin Ram **

* *chiara.franzoni@polito.it*

Chiara Franzoni, DISPEA, Polytechnic of Turin, Corso Duca degli Abruzzi 24b, Torino, 10129 (Italy)

** College of Computing, Georgia Institute of Technology, Atlanta, GA (USA)

Abstract

We present an application of a clustering technique to a large original dataset of SCI publications which is capable at disentangling the different research lines followed by a scientist, their duration over time and the intensity of effort devoted to each of them. Information are obtained by means of software-assisted content analysis, based on the co-occurrence of words in the full abstract and title of a set of SCI publications authored by 650 American star-physicists across 17 years. We estimated that scientists in our dataset over the time span contributed on average to 16 different research lines lasting on average 3.5 years and published nearly 5 publication in each single line of research. The technique is potentially useful for scholars studying science and the research community, as well as for research agencies, to evaluate if the scientist is new to the topic and for librarians, to collect timely biographic information.

Keywords

content analysis; academic scientists; semantic search; research trajectories; knowledge development.

Introduction

In recent years an increasing number of studies have concentrated on analyzing the work and behavior of individual scientists within large databases of scientific publications. The approach of taking single individuals as the main unit of analysis in large field studies took off in the 1970s within the Sociology of Science and is becoming increasingly widespread for applications of Economics of Science and Innovation and for studies of Labor Market for research.

Common bibliometric applications make use of indicators derived from scientific publications to characterize the attributes of single scientists. Citations received (Cole and Cole, 1967; Narin and Hamilton, 1996; Garfield, 1979), co-authorship (Hicks and Hamilton, 1999) and co-citation analyses (Peritz, 1992), for instance, are used to add non-subjective information to the profile of single scientists and of their academic production.

Scientific publications, of course, contain many more information than their sheer author and citations, but the bulk of information is hidden into their text and understandable only to peer-readers. To overcome these problems, it is possible to make use of Software-Assisted Content Analysis, which consists in extracting and organizing non-structured information from plain text, by means of semantics, into a standardized format suitable for several different uses. Among the other things, this set of techniques allow inferring certain characteristics and meaning of full texts, and at the same time offer the advantages of being replicable for very large sets of data in relatively short times, allowing non-subjective and unskilled reading.

In the last decades, applications of Content Analysis to scientific publications (for instance co-word and semantic analysis), have quite extensively been used to map the state and evolution of single or multiple scientific disciplines (Courtial et al., 1997; Klavans and Boyack, 2005), but have not been used so far to characterize individual scientists in terms of interests and topics of inquiry, and of their evolution over time.

¹ This work was supported by the CERIS, National Research Council of Italy and was done while one of the authors (Chiara Franzoni) was kindly guested as visiting scholar at the Andrew Young School of Policy Studies (Georgia State University, Atlanta, GA). The authors wish to thank Paula Stephan and Francesco Lissoni for comments and suggestions. All usual disclaimers apply.

In this work we propose an application of clustering algorithms to analyze the topic of inquiry (and their evolution) followed by individual scientists across a time-span of a 15 years by applying Content Analysis in a convenient way. The storage of texts both in cross-sectional and longitudinal dimensions is suitable for work that aims at analyzing the state and evolution of a single scientist's research. For instance, it makes possible to identify the different lines of research followed by a scientist at a specific point in time and along the years, to spot when a scientist enters a new topic (either new to him, or new to the entire set of documents), or abandons it, and to appreciate the amount of effort devoted to different streams of one's production.

Applications of this methodology are useful in a large number of areas, including works on careers in science (Fox; 1983; Stephan and Levin, 1992), on the production and dissemination of knowledge along research trajectories (Hackett et al., 2004), on the functioning of the scientific communities (Crane, 1072), on policy of science and research (Godin, 2003) and more generally on the dynamics of science and technology (Gibbons et al., 1994; Leydesdorff, 2002). In principle, this methodology can be used for several purposes, other than research. It can be used by librarians seeking to obtain biographic information on the topics and subfields addressed by a single author and it can prove useful for granting agencies, which are generally interested in knowing the response of scientists to the choice of program funded, for instance to see whether or not receiving funds in a certain program is likely to divert the natural path of research.

The paper is organized as follows: in the next section we address the choice of research paths followed by individual scientists. We then present the dataset collection procedure and final information stored (section "Data"). In the "Methodology" section we describe the clustering algorithm adopted and we lastly describe and comment the results in the final section.

Research Paths followed by Scientists in Academia

The choice of a research line in academia is a delicate process of balancing curiosity and opportunity. On the one side, academics enjoy the freedom of choosing their topics of enquiry according to their attitude and curiosity. Novelty in science and research pays-off with the highest rewards, as of course originality is among the general objectives of all studies (Hagstrom, 1965). On the other side, once a path has been chosen, shifting towards new interests is costly and time consuming and, in many cases, non advisable over short periods.Indeed, one feature of research that looks immediately evident while observing science directly is that the majority of scientists follow quite linear research paths along their careers (Ziman, 1968), in the sense that they exploit research lines over medium-to-long horizons and move to new topics only gradually.

Of course, to a great extent, linearity is a function of specialization, which comes along with the ever-increasing complexity of research. In all disciplines, the formal training given in undergraduate and graduate programs comprises all the foundations of a topic. Yet, specialties are chosen by young scientists quite early in their career, at the latest upon completion of graduate studies, when they are required to work at their PhD dissertations. Whereas in the early years PhD students may try several research lines, later on, specialization prevails over change, which will become rather infrequent and mostly gradual (Hagstrom, 1965). The reasons behind this tendency are to be searched in the fact that switching to a new line of research is extremely costly for a scientist in several different ways. First, it requires time and effort. For instance, scientists interviewed in several sub-disciplines of Life Sciences indicated that addressing a new topic required at least one year of work before any result could be published (Hackett, 2005). This circumstance is particularly uncomfortable in highly competitive job environments, where the score of publications is extremely important for personal careers. To face the problems of discontinuity of research lines, mature scientists often keep open several different lines of research simultaneously, which helps diversifying the risk of being tied to a single unfruitful research.

Second, in almost all hard sciences, every specific research field requires investing in a set of instruments and techniques and establishing effective collaborations. In this respect, the change of a topic is not only slow because it requires learning the foundations and the state of art of a partially new subject, but also because it requires a general adjustment of techniques, team of research, and

equipment (a set of assets that scholars of science address as “research ensemble”), that claims for some enduring interest (Hacket et al., 2004).

Third, the linearity in one’s research paths are mirrored and reinforced by the fact that, in pursuing their single career strategies, scientists see great benefits from building up and maintaining a strong personal identity within their community of reference. Within hard sciences, there are at least two sorts of benefits associated to establishing a clear identity. First, within a research groups or a laboratory, the identity of an individual is based heavily on the track of tasks and duties that he or she has accomplished over the years, including the technical skills accrued at the bench (Hackett, 2005). Acquiring a relevant experience and being knowledgeable in a specific subject is an advantage in the job-market and often the mobility of younger scientists trained in leading institutions is motivated by the desire of acquiring knowledge on certain cutting-edge processes. Second, for senior scientists, a clear identity serves to gain legitimacy on the eyes of colleagues, when addressing a certain theme: working heavily on a topic, posting contributions on journals, participating at conferences and events where the research base is nurtured and results disseminated, are necessary to establish a reputation among the colleagues and peers. As well known, this mechanism was heavily discussed by Merton and colleagues, which highlighted and empirically proved the importance of cumulativity in science brought by personal recognition, which encourages repeated participation and persistence (Merton, 1968; Allison and Stewart, 1974). Thus, an additional reason for career-minded scientists to build cumulatively on their past work, stands in the fact that being regarded for few distinctive features of their preparation, interests and achievements, as validated by the scientific community, allows certain status benefits that are otherwise unavailable to people new to a subject. A clear identity serves essentially to gain legitimacy when addressing a certain theme, and enhances the probability of obtaining support and credit in research, both in terms of research grants and support (for instance all funding agencies appreciate the fitness of a curriculum with the stated objectives of a research), and in terms of visibility vis-à-vis their community of reference (being invited to speak at conferences, write on special issues, sit in journals editorial boards, etc.). Lastly, because of the well documented fads and fashion existing in science, persistence along an established domain offers a low risk strategy in research contexts characterized by strong competition (Garner, 1979).

Having addressed the rational behind novelty and specialization, we expect that the position in the spectrum going from the one end to the other is chosen by scientists quite carefully and that we can hence characterize the scientific production of researchers with respect to the rate at which they enter into new topics, the portfolio of research lines they pursue simultaneously, the duration of those research lines and the intensity of contributions made in every specific line. In the following sections we offer an empirical assessment of these characteristics by making use of a novel technique.

Data

The original database used for the study comprises a large group of scientists doing research in American universities in all the various subfields of Physics. The collection of data was started from a list of names obtained from the American Physical Society (APS) archive of Fellows. The APS has a very large membership base, and virtually all USA physicists belong to it since their doctoral studies. The status of “Fellow” is a life-long honorary title given year after year to a very small number of scientists (by internal regulation, to a maximum of 0.5% of the members), in recognition of their scholarly merits. The selection of Fellows is made annually in a peer-review fashion, starting with the nomination of meritorious scientists from the APS members worldwide and ending with the appointment of awards by the subgroup field to which the individual has mostly contributed. The use of the APS Fellows archive hence offers the advantage of having a good, unbiased indication of the subfield of Physics to which a scientist belongs, along with a synthetic description of the major contributions for which the individuals are regarded, which may prove to be of special interest for future research.

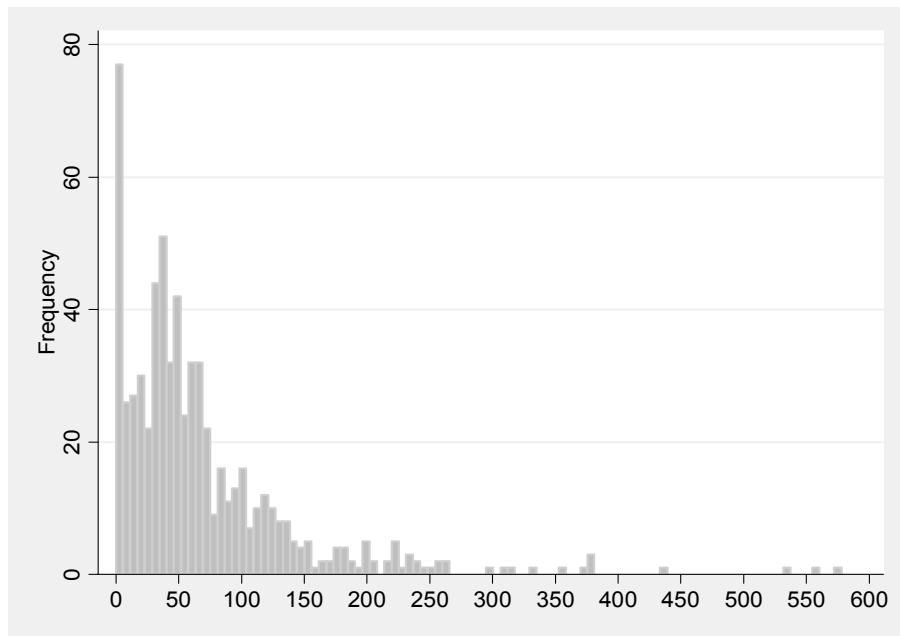


Figure 1. Total number of publication per scientist. Frequency Distribution.

We started by collecting all (1054) names of fellows awarded since 1995 to 2002, which were affiliated to US American universities at the time of the nomination. For 933 of these individuals (88%) we were able to retrieve full CV information through web search. After drops of people retired in the period of observation and of common names, we obtained a list of 650 individuals.

From the ISI Science Citation Index we extracted information on all publications made by each individual on scientific journals since 1990 to the beginning of 2006 and kept only publications for which abstracts were available. The database resulted in 45,342 unique combinations of scientist-publications and 38,178 unique SCI publications, all recorded as references and accompanied by the full abstracts content. Figure 1 shows the distribution of the total number of publications per scientist.

Clustering Methodology

The clustering algorithm is described in Figure 2. Each paper constitutes a document and is comprised by title and full abstract. Each document is represented by a vector of term weights based on the classic vector space model (VSM) (Salton 1989). Each cluster of related documents contains a centroid vector which we call the cluster's representative. To compute document vectors we first preprocess the documents to transform them into a form more amenable to vector space analysis. In the preprocessing stage we remove stopwords from documents and stem the remaining words with the efficient regular expression-based Porter stemming algorithm (Porter 1980). Stopwords are frequently occurring function words in a language which have important grammatical roles but carry no meaning, such as prepositions and articles (e.g., a, an, the, of, for, and, or). Stemming is similar in spirit to finding the root forms of words. For example, a stemmed term index might contain only the root form “walk” to represent “walk,” “walked,” and “walking.” Stopword removal and stemming measures reduce the total number of words in the corpus, which helps to reduce the dimensionality of the document vector space and improve efficiency.

Once the documents are preprocessed, term weight vectors are computed for each document. The weight w_{t_i, d_j} of a term t_i in a document d_j is calculated by equation (1). TF_{t_i, d_j} (term frequency) is the count of t_i 's occurrence in document d_j . DF_{t_i} (document frequency) is the number of documents in

1. Assign the first document D_1 as the representative for C_1 .
2. For each document D_i in D_2 to D_n
 - 2.1 $S_{max} = -1; C_{max} = -1;$
 - 2.2 For each cluster C_j
 - 2.2.1 Calculate the similarity S_{ij} between D_i and the representative (centroid) for C_j .
 - 2.2.2 If ($S_{ij} > S_{max}$)
$$S_{max} = S_{ij}; C_{max} = C_j;$$
 - 2.3 If S_{max} is greater than a threshold value S_r then
add document D_i to cluster C_{max} and recalculate the representative for C_{max} including document D_i , return to Step 2 and continue with document D_{i+1}
 - 2.4 Use D_i to initiate a new cluster C_{new} (since it was not sufficiently similar to any existing cluster).

Figure 2. Single Pass Clustering

which the term t_i occurs. N is the total number of documents in the collection. This term weight formula captures both the importance of a term in characterizing a document, and the degree to which a term discriminates between documents. Terms that occur frequently in a given document are assigned higher weights, and terms which occur in many documents receive lower weights to reflect the fact that the terms do not distinguish well between documents.

$$W_{t_i, d_j} = (\log(TF_{t_i, d_j}) + 1) \times \log\left(\frac{N}{DF_{t_i}}\right) \quad (1)$$

Representing documents as vectors in a Euclidean space allows us to use distance metrics from linear algebra to compute a similarity measure between documents (Manning and Schütze, 1999; Rasmussen, 1992). Our clustering algorithm, as most text retrieval algorithms, uses cosine similarity, which bases similarity on the angle between document vectors. The smaller the angle between two document vectors, the higher their similarity. Vectors for similar documents will be “near” each other in the vector space for some definition of “near”. Euclidean distance is the most obvious distance metric, but its sensitivity to document length requires document normalization. For example, two documents about the same topic but with different lengths would have a high Euclidean distance despite their semantic similarity. To calculate Cosine similarity measure, we hence converted them in unit vectors firstly, to avoid the need of normalization.

Once all the documents are represented as vectors in a term-weight vector space, our clustering algorithm can compute clusters for the document corpus. Our clustering algorithm, outlined in Figure 1, assigns each document to exactly one cluster. The clustering algorithm is parameterized by a similarity threshold S_r affecting the number of clusters computed for a given corpus, which was set to be 0.08. For example, a high similarity threshold requiring high similarity for documents within a cluster will result in a greater number of clusters and vice versa. The algorithm computes an appropriate number of clusters with respect to the similarity threshold, thus avoiding the need to compute or estimate the number of clusters a priori as in traditional K-means clustering. The Single Pass Clustering assigns a new document to the cluster with which it has the maximal similarity and the similarity is above a predefined threshold.

Because our clustering algorithm computes clusters automatically based only on the textual content of documents, clusters do not correspond perfectly to human-derived categories such as those determined by the American Physical Society. Our clusters thus are free to capture other sources of similarity such as research methodology, experimental instrumentation, and other ways in which research can be considered similar. As a consequence, while a scientist might stay within a particular subfield of Physics, the clusters to which his documents are assigned may reflect, for example, variations in the scientist's methodology over time. Conversely, a scientist may change subfields but retain similar methods, resulting in a low variability with respect to our automatically computed clusters. In summary, we let the documents speak for themselves in our text analysis.

Results

The clustering algorithm applied on our database of Physicists's publications resulted in 660 clusters, ideally corresponding to publications with similar content. Some statistics on the resulting clusters are reported in Table 1.

Table 1. Clusters characteristics. Summary Statistics

Variable	Obs.	Mean	Std. Dev.	Min	Max	Median
<i>size of cluster (# publications)</i>	660	57.84	117.85	1	1173	13
<i>population of cluster (# different individuals)</i>	660	16.42	25.27	1	257	6
<i>duration of cluster (years)</i>	660	7.77	5.69	1	17	7

Clusters can be more or less populated in terms of number of articles grouped by the algorithm. We call "size" the cluster dimension in terms of publications. There are nearly 58 publications on average per cluster, but the size is highly variable, as shown in Figure 3, reporting the frequency distribution for the sizes of clusters. There is a considerable number of clusters having just one publication (111, equal to 17% of clusters), which correspond to the clusters identified by the software as relatively unrelated to the rest of the publications, and a very long right-tail. The right graph in Figure 3 represents the distribution reduced to the 25th – 75th percentiles (values 2 and 58 respectively).

Clusters also differ in the number of scientists in our dataset that contributed to it. To account for this differences, we counted the number of database IDs that have at least one publication in a cluster. Because the cluster algorithm works on a unique corpus of text, we expect clusters in which many different scientists have published to be more general in content than clusters more concentrated on a single scientist, holding the number of publications constant. Figure 4 shows the distribution of clusters according to the number of single individuals that contributed to it (restricted to the 95 percentile, equal to 67 IDs per cluster).

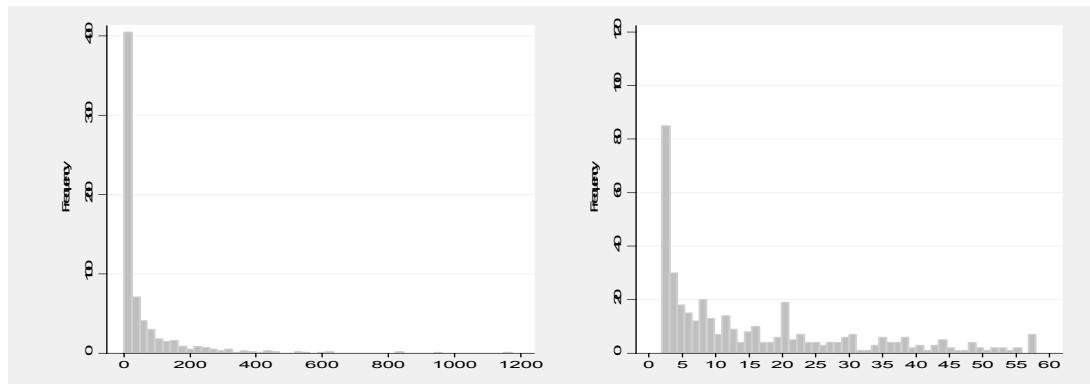


Figure 3. Left: Cluster Size (number of publications per cluster) Frequency Distribution. Right: Cluster Size (number of publications per cluster) Frequency Distribution of 25th - 75th percentile.

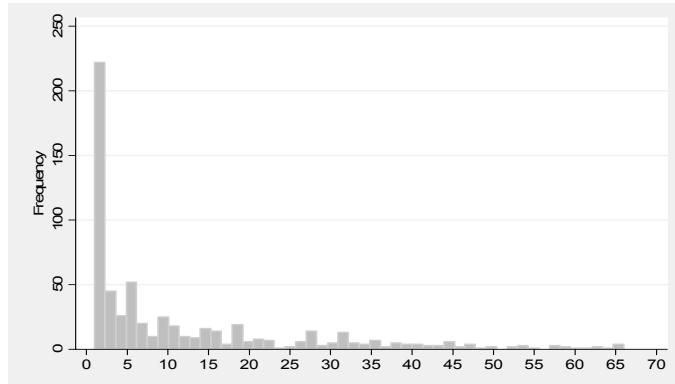


Figure 4. Population of clusters (# of different IDs in a cluster). Frequency Distribution (restricted to 95th percentile).

In terms of duration, clusters differ in the number of years elapsed since the earliest publication to the latest publications grouped in the same cluster.

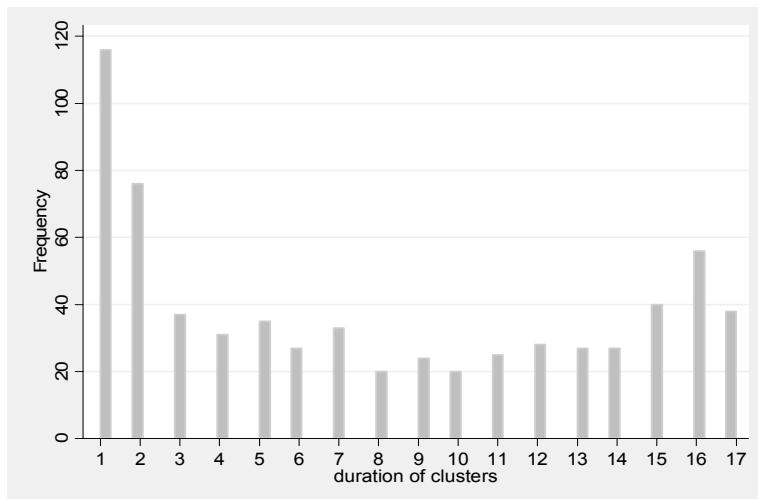


Figure 5. Duration of Clusters (in 17 years). Frequency Distribution.

Our publications were collected during 2006 and dated back to 1990 (first year for which ISI records the full article abstracts). Given the construction of the dataset, the distribution is both left and right censored, in the sense that clusters more active towards the two ends of the observation period have more chances to be not fully observed. Hence cluster duration is here given as a summary statistics to check the consistency of the clustering algorithm, but it should not be taken as an estimate of duration of single topics. The distribution of clusters duration is shown in Figure 5.

The clusters with duration equal to 1 year are heavily inflated by the 111 clusters having just one publication in it. Nonetheless, there is also a high proportion of clusters having a longer duration (16 or 17 years), indicating that there are topics which have a longer publication-cycle than our time-span.

Having presented statistics on the results generated by the clustering algorithm, we now show a possible use of the cluster data to obtain information on single individuals' scientific productivity over time. We constructed three indicators as follows:

1. Diversification of interest: number of clusters in which a scientist has at least one publication across the time-span.
2. Intensity of interest: number of papers a scientist makes on average in each single cluster

- (calculated for IDs whose diversification ≥ 1).
 3. Duration of interest: number of years a scientist stays on average in each single cluster
 (calculated for IDs whose diversification ≥ 1)..

The indicators of Intensity and Duration of interest is calculated only for the 603 (93%) scientists that had at least one publication (and one cluster).

Table 2 shows general statistics for the three indicators and on the total number of publications stored in the database. All indicators are calculated over the 17 years time-span.

Table 2. Diversification, Intensity and Duration of Interest. Summary Statistics

Variable	Obs.	Mean	Std. Dev.	Min	Max	Median
<i>Productivity (# papers)</i>	650	67.93	73.12	0	578	48
<i>Diversification of interest (# clusters)</i>	650	16.22	14.54	0	118	13
<i>Intensity of interest (# papers per cluster)</i>	603	4.73	6.38	1	82.57	3.37
<i>Duration of interest (# years)</i>	603	3.41	1.45	0	13	3.23

The number of different clusters in which a scientist has made at least one contribution gives a measure of the ranges of interests he had during the years, a reason why we call this measure “Diversification” of interests. The higher the number, the wider the spectrum of different lines of research and topics on which he/she was active. Figure 6 shows the distribution of our diversification indicator. The average scientist in our dataset had a little more than 16 different research lines during 17 years (the median scientist 13): nearly one new line per year. This is overall plausible, and in line with the fact that the scientists in our database are certainly head of big university labs and typically supervise (and co-author) the work of several post-doc and junior scientists working on different research lines simultaneously.

On average the scientists published almost 5 papers on each of their different research clusters, although this measure is highly variable per cluster and per scientist, going from 1 to nearly 83 paper per cluster. The frequency distribution of the indicator is shown in Figure 7.

Finally, the frequency distribution of the duration of the interests is shown is Figure 8. The average duration is 3.41 years per single cluster and is distributed almost normally.

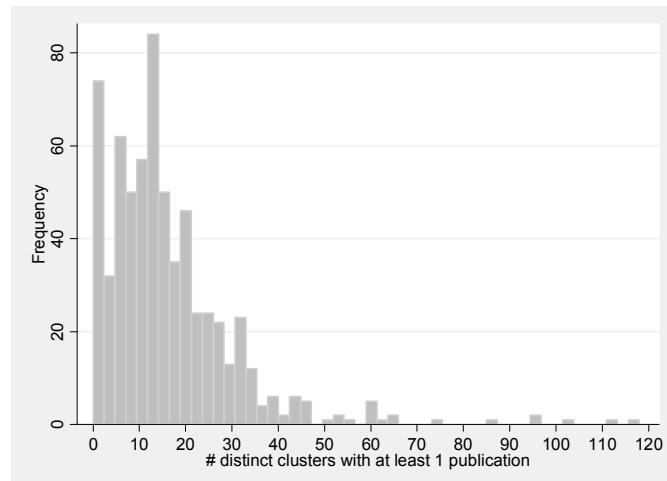


Figure 6. Diversification of interest. Frequency Distribution

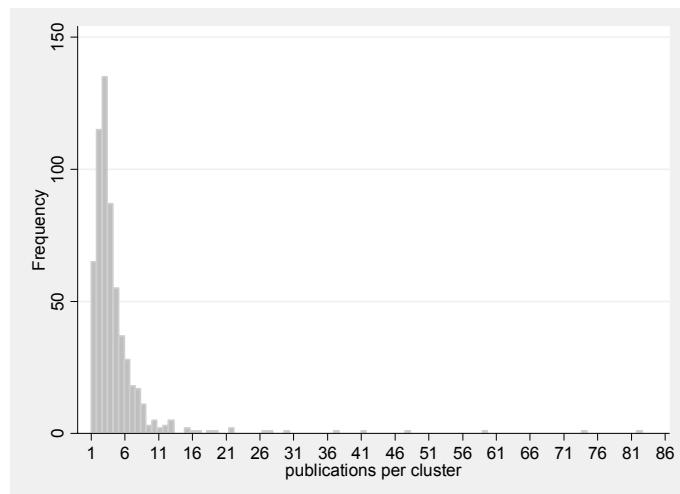


Figure 7. Intensity of Interest. Frequency Distribution

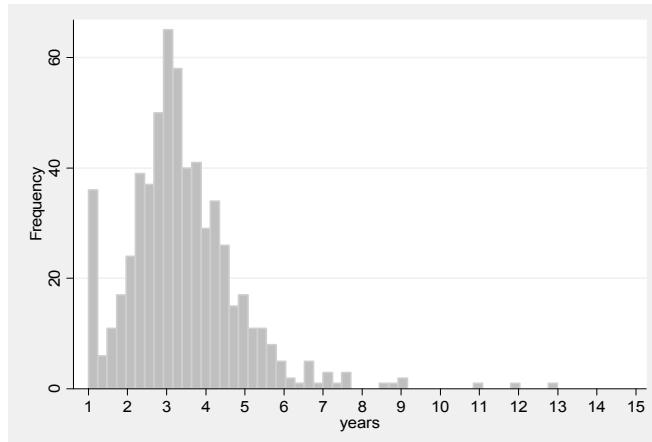


Figure 8. Duration of Interest (during 17 years). Frequency Distribution

Conclusions

We presented a methodology of Software-Assisted Content Analysis that is capable of extracting information on the characteristics of a scientist's productivity in terms of diversification of interests, intensity and duration over time, by clustering the title and abstracts of scientific publications.

Our approach is novel to the extent that clustering algorithms are here used to extract information on single scientists, rather than mapping overall scientific fields. Our results produce information for single IDs over time and are hence stored in panel format.

We expect this methodology to be useful for a number of applications ranging from the studies of the development of the knowledge needed to cultivate Science and Innovation Policy, to applications for librarians and for granting agencies that may obtain timely biographic information upon a scientist's work for large masses of data, without a subjective evaluation.

We apply the methodology on an original sample of publications, based on a dataset of 650 American star-scientists in the field of Physics for which all SCI publications were collected from 1990 to the beginning of 2006. Scientists in our sample are highly recognized by their scientific community and we expect them to be head of big university labs, supervising the work of several junior researchers. Our results estimated that scientist were working on average on 16 different research lines during the time-span, publishing nearly 5 articles per cluster, and that research lines were active on average for 3,4 years.

References

- Allison P.D.; Stewart J.A. (1974), Productivity Differences Among Scientists: Evidence for Accumulative Advantage *American Sociological Review*, 39(4), 596-606.
- Cole S., Cole J.R: (1967), Scientific Output and Recognition: A study in the Operation of the Reward System in Science, *American Sociological Review*, 32(3):377-390.
- Courtial, J. P., Sigogneau, A., and Callon, M. (1997), Identifying strategic sciences and technologies through scientometrics, in W. B. Ahston and R. A. Klavans (eds.), Keeping abreast of science and technology, technical intelligence for business, Columbus, OH: Battelle Press.
- Crane D (1972), *Invisible Colleges. Diffusion of knowledge in scientific communities*, Chicago & London: The University of Chicago Press.
- Fox M.F. (1983), Publication Productivity among Scientists: A critical Review, *Social Studies of Science*, 13(2), 285-305.
- Garfield, E. (1979), *Citation Indexing: Its theory and application in science, technology and humanities*, New York, John Wiley.
- Garner C.A. (1979), *Academic Publication, Market Signaling and Scientific Research Decisions*, Economic Inquiry, 17(4):575-584.
- Gibbons M., Limoges C., Nowotny H., Schwartzman S., Scott P., Trow M. (1994), *The new production of knowledge. The Dynamics of Science and Research in Contemporary Society*, Sage Publications.
- Godin B. (2003), *The emergence of S&T indicators: why did governments supplement statistics with indicators?*, *Research Policy*, 32, 679-691.
- Hackett E.J., Conz D., Parker J., Bashford J., DeLay S. (2004), *Tokamaks and turbulence: research ensembles, policy and technoscientific work*, *Research Policy* 33(5):747-767.
- Hackett E.J (2005), Essential Tensions: Identity, Control, and Risk in Research, *Social Studies of Science*, 35(5):787-826.
- Hagstrom W.O. (1965), *The Scientific Community*, Basic Books Inc., New York, London.
- Hicks D., Hamilton K. (1999), Does University-Industry Collaboration Adversely Affect University Research?, Issues in Science and Technology Online, Summer 1999, 74-75, http://www.nap.edu/issues/15.4/images/realmnum_big.jpg.
- Klavans R., Boyack, K.W. (2005), *Generation of large-scale maps of science and associated indicators*, SANDIA Report SAND2005-7538, 01 Dec 2005.
- Leydesdorff L. (2002), Indicators of structural change in the dynamics of science: Entropy statistics of the SCI Journal Citation Reports, *Scientometrics*, 53(1):131-159.
- Manning C., and Schütze H. (1999), *Foundations of Statistical Natural Language Processing*, Cambridge: MIT Press.
- Merton R. K. (1968), The Matthew Effect in Science, *Science, New Series*, 159(3810):56-63.
- Moed H.F., Burger W.J.M., Frankfort J.G., Van Raan A.F.J. (1985), The use of bibliometric data for the measurement of university research performance, *Research Policy*, 14, 131-149.
- Narin F., Hamilton K.S. (1996), Bibliometric performance measures, *Scientometrics*, 36(3):293-310.
- Peritz, B.C. (1992), On the Objectives of Citation Analysis: Problems of Theory and Method, *Journal of the American Society for Information Science*, 43(6):448-451.
- Porter M. F. (1980), An algorithm for suffix stripping, *Program* 14(3):130-137.
- Rasmussen E. (1992), Clustering Algorithms, in: Frakes N. and Baeza-Yates (eds.), *Information Retrieval: Data Structures & Algorithms*, New Jersey: Prentice Hall.
- Salton G. (1989), *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Reading, MA: Addison Wesley.
- Stephan P.E., Levin S.G. (1992), *Striking the Mother Lode in Science: The Importance of Age, Place, and Time*, Oxford University Press.
- Ziman J.M. (1968), *Public knowledge: an essay concerning the social dimension of science*, London, Cambridge U.P.
- Courtial, J. P., Sigogneau, A., and Callon, M. (1997), Identifying strategic sciences and technologies through scientometrics, in W. B. Ahston and R. A. Klavans (eds.), Keeping abreast of science and technology, technical intelligence for business, Columbus, OH: Battelle Press.

Using Patent Data for Monitoring the Globalisation of R&D¹

Rainer Frietsch*, Sybille Hinze* and Pari Patel**

*sybille.hinze@isi.fraunhofer.de, rainer.frietsch@isi.fraunhofer.de

Fraunhofer Institute for Systems and Innovation Research, Breslauer Str. 48; 76139 Karlsruhe (Germany)

**p.r.patel@sussex.ac.u,

SPRU (Science and Technology Policy Research), University of Sussex, Freeman Centre, Falmer, Brighton, East Sussex BN1 9QE, (UK)

Abstract

In the paper results of a study will be presented that set out to develop and test of an approach to link the location of R&D investments of companies with the geographic distribution of their patent portfolio. The rationale behind it was that patent information could be used to substitute R&D investment data at the firm level as the latter can be obtained only with comparable high efforts and often remains incomplete. However it turned out that gathering regional R&D investment data was even more challenging than anticipated and thus, the focus of the analysis shifted towards the analysis of regional patenting activities of 18 company groups for which patent portfolios were constructed based on information on the location of these activities. Based on these data patterns become visible concerning firm preferences concerning their regional engagement in selected fields of technology and the choice where to apply for patent protection. In addition methodological recommendation concerning the procedures and data used to analyse structures of R&D activities of multinational firms can be derived.

Keywords

R&D globalization; patent analysis; micro-level analysis; multinational firms; monitoring R&D

Introduction

Large multinational companies play a dominant role in the innovation activities of their home country and control a vast proportion of world's stock of advanced technologies. Their decisions in terms of the mode, location and exploitation of their R&D results greatly influence the home country's technological potential and competitiveness. Over the past two decades we observe an increasing trend towards internationalisation of R&D activities of large firms. In this context it is of interest to be able to determine the extent and the nature of the R&D activity that has been relocated abroad and to be able to evaluate the potential impact on the respective home countries. The aim of the study undertaken was to develop a methodology to address this issue. Thus, within this exploratory study we attempted to develop and test an approach that enables us to link the location of R&D investments of an industrial company with the geographic distribution of its patent portfolio. In order to do so R&D and patenting data was collected for 18 multinational companies in the areas of ICT and biotechnology in order to construct geographical distributions of their R&D investments and patent portfolios. At the same time the study should test to what extent company patent portfolios could be used as a proxy for studying the pattern of R&D investments.

Methodology

In order to test relevance of patent portfolios as a proxy for patterns of R&D investments data on R&D and patenting activities of the 18 company groups (i.e. including all their affiliates) given in table 1 was gathered.

In a first step all majority owned affiliates of the 18 companies were identified. The main data sources used are Hoovers and published annual accounts of the companies. Hoovers contains a 'family tree' for each company that lists all current majority owned subsidiaries and divisions. In a second step we tried to identify the R&D performing locations and collect the corresponding R&D expenditures for

¹ The article bases upon a feasibility study on R&D Globalisation conducted within the framework of the EU funded ERAWATCH Network, which surveys national research policies, structures, programmes and organisations (see also <http://cordis.europa.eu/erawatch/>).

those 18 company groups. The problem we encountered is that as companies are not obliged to report R&D activities beyond the company group level very few companies systematically report data on their R&D expenditures (or employment) by named subsidiaries or by geographic location. Due to the restrictions of the study we had to rely on publicly available data sources only. In particular we analysed Company accounts (including SEC filings); journal articles covering economics, business and management accessed via ABI Inform and Internet sources. Due to lacking data the aim of the study to assess the validity of using patent information for monitoring the patterns of R&D investment had to be omitted. Thus, in our analyses we rather focussed on firm strategies concerning their patterns of regional R&D and patenting activities.

Table 1. Companies under investigation

<i>ICT Companies</i>	<i>HQ Ctry</i>	<i>Biotech Companies</i>	<i>HQ Ctry</i>
<i>Alcatel</i>	FR	<i>Boehringer Ingelheim</i>	DE
<i>Ericsson</i>	SE	<i>Eli Lilly</i>	US
<i>Matsushita</i>	JP	<i>Glaxo SmithKline</i>	GB
<i>Motorola</i>	US	<i>Merck AG</i>	DE
<i>Nokia</i>	FI	<i>Novartis</i>	CH
<i>Philips Electronics</i>	NL	<i>Pfizer</i>	US
<i>Samsung</i>	KR	<i>Roche</i>	CH
<i>Siemens</i>	DE	<i>Sanofi-Aventis</i>	FR
<i>Thomson</i>	FR	<i>Biotie Therapies Pharma</i>	FI

Thus, patent data was gathered for each of the company groups. Patent data was collected using the PATSTAT database supplied by the European Patent Office (EPO). PATSTAT covers patent applications from more than 70 offices around the world, storing information on bibliographic details of the applicants, the inventors and of patents, such as the date of first filing (priority) and legal status. It only contains records of published patent filings. Applications that are withdrawn or rejected are not covered by this database. Furthermore the United States Patent and Trademark Office (USPTO) until 2001 only published granted documents and even today only a fraction of USPTO applications are published after 18 months. For example, pure national applications to the USPTO do not have to be published after 18 months. Many US-based assignees do not file an international application and therefore postpone the disclosure of their technology. This is the reason why the numbers of patent filings to the USPTO declines after the priority year 2000. However, no other publicly accessible database covers those applications either.

Patent data from the EPO and the USPTO was collected for the priority years 1998-2003². In addition direct patent filings to the EPO plus all PCT filings were retrieved. This means we created a „fictitious“ patent office that covers most of the international patent applications of the companies and thus can be used as a substitute for the patent family. This concept intends to reflect filings of higher interest or of greater value than pure national applications.

Due to the fact that PATSTAT is a database that combines data from several patent offices the quality of the data stored is heterogeneous. For instance some information might be missing e.g. for many PCT filings the inventor country is missing. This has two effects. Firstly the number of patents without an inventor country is much higher for the "fictitious" office of EP-direct plus all PCT filings, than for the EPO and the USPTO. On the other hand, the number of EP-direct plus PCT patent filings might be lower in our analyses than the actual number of patent filings for the same company in the same priority year. The reason lies in the missing inventor country information for PCT applications

² Due to procedural reasons the data is not complete for the priority year 2003. However, it was decided to include it in the analyses, as structures of the companies are analysed and thus absolute patent numbers are of secondary importance.

whereas this information is available for the Euro-PCT-RP filing that become a EPO patent application.

Another problem is that applicant names in the PATSTAT database are not unified. EUROSTAT has done some work in order to standardise the names. The respective information was used to construct the queries. At the same time the standardised names were checked against the "original names". Some discrepancies appeared if compared with the definition of affiliates as stored in Hoovers Who Owns Whom database. As far as possible these discrepancies were reduced by examining all cases with substantial numbers of patents. However, it was impossible to check all name variants, thus names appearing less than 3 times were not checked and corrected.

For the 18 company groups the following data was retrieved: priority date (1998-2003); year of application; application office; IPC classes grouped into 6 major fields (Chemistry; Mechanical Engineering; Process Engineering; Electrical Engineering; Instruments and Consumption/Construction) and 4 subfields for ICT (Telecommunications, Consumer Electronics, Computers, other ICT); name and location of inventors; name of applicants. The names of the assignees were cleaned and consolidated on the basis of the information on subsidiaries collected from Hoovers. The location of innovative activity was proxied by the country address of the inventor which is a widely used approach. Fractional counting was applied in case patents had multiple inventor country addresses.

Profiles of technological specialization by country location were constructed. The Specialization Index for a company in a particular technical field has been defined as:

$$S_{ij} = \% \text{ of all patents in technology class } j \text{ by country } i / \% \text{ of all patents by country } i$$

Thus a value of greater than unity indicates that in that particular technology a country is a relatively more favourable location (i.e. compared to the average for that company).

This index is a reflection of a company's decision to locate its technological activity in a given field in a country. In other words it is a company specific index.

Results

For each company 3 profiles have been produced, based on data from: EPO; USPTO and EPO-Direct plus PCT for 3 time periods 1998-2003; 1998-2000; 2000-2003. The profiles are based on the specialization index discussed above. They also contain data on the most important locations for each company (as proxied by inventor country addresses) and how these have changed over time. Table 2 exemplary represents the results obtained.

The main point to emerge from the analysis is that there are systematic differences across the 3 patent offices in both the indicators (specialization and volume of patenting from a location). Thus the share of patents of a particular company that have inventor addresses in a particular country at the EPO can be very different from the same share at the USPTO. For example: GlaxoSmithKline has 36% of patents invented in the US and 44% in the UK at the EPO in 1998-2003, the corresponding share at the USPTO for the two countries are: 64% in the US and 26% in the UK. At the same time the patents from EPO-Direct plus PCT filings give shares of 52% for the US inventors and 39% for those from the UK. An underlying explanation could be the relative importance of the different markets that is reflected in the relative importance of the different patent offices. In terms of specialization the results for GlaxoSmithKline show that France, Italy, Belgium, and the US have a locational advantage in Chemistry (the main technical area of patenting for the company) in terms of both EPO and USPTO patents for the period 1998-2003. However the Specialization index based on EPO-Direct plus PCT filings shows that the US no longer enjoys this advantage, although France, Italy, and Belgium do have a locational advantage.

At a more general level some of the results obtained in terms of the relative importance different locations are consistent with prior knowledge. Thus for example the US based companies (Motorola, Pfizer and Eli Lilly) consistently favour the US as a location. Thus Pfizer has more than 70% of all its patents with inventor addresses in the US, regardless of the patent office. Moreover the only Japanese company in the sample, Matsushita, has more than 90% of all its patents invented in Japan. The pattern for the European firms is more diverse but the US is consistently an important location for most of them.

Table 2. Technology profiles for GlaxoSmithkline

EPO patent Filings									
1998-2003			Specialization Index						
No. of Patents	Inventor Country	%	Chemistry	Consumption	Electrical	Instruments	Mecheng	Process	
28	DE	1.39	0.34	65.49	0.00	0.89	0.00	5.09	
55	FR	2.74	1.07	0.00	0.00	0.85	4.06	0.67	
59	IT	2.94	1.14	0.00	0.00	0.59	0.00	0.00	
192	BE	9.64	1.09	0.00	0.00	0.82	0.00	0.91	
719	US	36.06	1.03	0.00	1.20	0.77	0.16	0.67	
884	GB	44.29	0.96	0.20	1.09	1.28	1.87	1.30	
1995	N of pat	100	1713	15	39	296	9	106	

USPTO patent Filings									
1998-2003			Specialization Index						
No. of Patents	Inventor Country	%	Chemistry	Consumption	Electrical	Instruments	Mecheng	Process	
3	DE	0.35	0.25	283.50	0.00	3.15	0.00	6.44	
10	FR	1.39	1.10	0.00	0.00	1.48	0.00	0.00	
14	IT	1.98	1.13	0.00	0.00	0.00	0.00	0.00	
31	BE	4.18	1.10	0.00	0.00	0.46	0.00	0.54	
192	GB	26.27	0.93	0.00	1.50	1.55	0.00	1.23	
464	US	63.67	1.02	0.00	0.52	0.82	0.00	0.99	
729	N of pat	100	643	2	7	105	0	44	

EPO + PCT patent Filings									
1998-2003			Specialization Index						
No. of Patents	Inventor Country	%	Chemistry	Consumption	Electrical	Instruments	Mecheng	Process	
5	DE	0.42	0.70	235.45	0.00	0.24	0.00	3.86	
17	FR	1.35	1.02	0.00	0.00	0.22	0.00	1.14	
18	BE	1.42	1.06	0.00	0.00	0.70	0.00	2.68	
23	IT	1.83	1.13	0.00	0.00	1.15	0.00	0.00	
490	GB	39.14	1.02	0.00	0.59	1.06	2.04	1.13	
655	US	52.35	0.98	0.00	1.24	0.99	0.38	0.92	
1252	N of pat	100	1111	1	37	166	1	61	

There are some signs of the emergence of the East Asian countries and Israel as location of technology. Thus for example in the case of Thomson, China and Singapore have increased in importance (at least in terms of EPO filings), and the same is true for Israel in the case of Motorola.

There is a systematic bias towards US based inventors in the data from the USPTO. Thus for example for Boehringer the share of US invented patents at the EPO in 1998-2003 is 15%, but this rises to 34% in the case of the data from the USPTO. This may partly be a methodological artefact that can be explained by the fact that European firms use the PCT route to apply for a patent in the US and there is a delay in such patents entering the regional phase at the USPTO.

Discussion

The underlying rationale for the project was to make a direct link between location of R&D facilities and the patent portfolio of a company. As already discussed due to missing data this aim could not be fulfilled. However, we still think that the project shows interesting results:

Systematic differences across the three patent offices in terms of the relative importance of different locations and the specialization patterns were found.

Data from each patent office has certain advantages and disadvantages. The main advantage of the EPO being the comprehensive coverage of data on inventor addresses and thus the opportunity to more precisely assign R&D activities in geographic terms. The main disadvantage is that there may be a 'home' bias in relation to the European companies and inventors. The USPTO also has comprehensive coverage of inventor addresses but suffers from the problem of incomplete information on all applications. There is no information on patent applications before 2001 (only on those that were granted), and since then national applications that are not filed internationally are not published. Additionally there is the strong home country bias of US inventors and companies. The main advantage of the EPO-Direct plus PCT filings is that it corrects some of the home bias of the EPO and USPTO data. However the main practical problem is that so far many inventor country addresses are missing in PATSTAT. This may be overcome in future versions of this database.

Based on the findings for future patent analysis aiming at the assessment of international R&D structures of MNEs we recommend that patent families are used instead of filings at specific patent offices. Alternatively the approach introduced here based on the creation of a "fictitious" patent office could be used – at least once problem of missing inventor country information is solved. A family might be constituted even by a single application at one of the relevant offices (EPO, USPTO, WIPO). Alternatively, a family consisting of at least two applications at any – national or international – office could be used.³ And PATSTAT appears to be the perfect source of information as it covers a large number of patent offices. In fact, it is intended to include a family definition directly in the database as it is provided by the EPO so that future studies can rely on this family definition. However, some methodological work needs to be undertaken before the question of regional distributions and patterns of R&D can be simply extracted from databases.

References

SPRU, ISI (2007): Exploratory study to test the feasibility of using Patent data for monitoring the Globalization of R&D. Final Report Submitted to the IPTS

³A definition with only one patent application at any patent office might be too broad and pure national filings would predominate the picture of each firm. We would recommend comparisons of the different approaches.

Measuring the Contribution of Clinical Trials to Bibliometric Indicators: Citations and Journal Impact Factor®

Antonio García Romero*, José Navarrete Cortés**, Cristina Escudero***, Juan Antonio Fernández López****, and Juan Antonio Chaichio Moreno****

*agr33@salud.madrid.org
Agencia Laín Entralgo, c/Gran Vía 27, 28013 Madrid (Spain)

**jcortes@ujaen.es
Service of Scientific Production, Universidad de Jaén, Campus de Las Lagunillas, Jaén (Spain)

***cescuderog.hpth@salud.madrid.org
Library, Hospital Puerta de Hierro, c/ San Martín de Porres nº 4 28035 Madrid (Spain)

****jaflopez@ujaen.es, sicaalme@ual.es
Scientific Information System of Andalucía, SICA (Spain)

Abstract

Clinical trials play a relevant role in the development of new drugs. After a clinical trial has been carried out, its results are usually published in scientific journals. These papers receive a significant number of citations that can affect the scores for indicators such as the Journal Impact Factor® or the h-index. However, there is a criticism with this practice especially because around $\frac{3}{4}$ of the clinical trials are funded by industry that can use this channel to promote their products. In addition, the *Frascati Manual* (OECD) establishes that clinical trials must be considered as part of product development and not research activities. We have established two main research questions: (i) Are clinical trials cited significantly more than other papers? (ii) To what extent are Journal Impact Factors ® modified by citations to clinical trials? We use the database from Thomson-ISI Web of Science® jointly with Medline to answer these questions. Our preliminary results suggest the following remarks. Firstly, the clinical trials are significantly more cited than other papers. Secondly, the Impact Factors are significantly reduced if we do not take into account the clinical trials. We believe that this information could be useful for the Research Policy decision makers.

Keywords

clinical trials; citation; impact factor.

Introduction

Clinical trials are essential for the development of new drugs. Pharmaceutical firms¹ invest huge amounts of money in R&D brings new products to the marketplace. Clinical trials are usually published in scientific journals and these papers have a strong influence on doctors and practitioners. Two main objections to this practice have arisen from the scientific community. On the one hand, the companies themselves select the clinical trials that are going to be published and, in general, these tends to demonstrate conclusions favourable to the companies' interests (Rochon, Gurbitz & Simms *et al.*, 1994). On the other hand, clinical trials should not be considered as *research* (OECD, 2002) because in fact, the doctors who take part in them only have to follow a list of written instructions.

In this paper, we measure the contribution of clinical trials to indicators commonly used in research evaluation, such as the average number of citations per document and the Journal Impact Factors®. Our main goal is to develop a set of bibliometric indicators appropriate for the measurement of the clinical trials' citations. These indicators should provide relevant information for research evaluation.

Using data from *Thomson-ISI Web of Knowledge®* and *Medline* we have carried out two analyses in order to explore the effect that clinical trials' citations have on several bibliometric indicators such as

¹ In this paper we also consider clinical trials related to health technologies such as surgical equipment, medical imaging or nuclear medicine, etc. Nevertheless, due to space limitations, we only talk about clinical trials related to pharmaceutical industry.

the *Impact Factor®*. Our results support the hypotheses that clinical trials receive a higher number of citations than other medical papers. This effect can be observed with the Journal Impact Factors for three of the four journals analyzed. We believe that this information can be useful for decisions concerning research grants.

Motivation and Background

The pharmaceutical industry has a relevant contribution for the citizens' well being. Proof of this fact is that almost all the drugs that have improved medicine along the past century have been developed by this industry (House of Commons Health Committee, 2005). The socioeconomic benefits of pharmaceutical innovation have also been measured both in terms of mortality reduction and subsequent economic growth (Lichtenberg, 2003).

However, launching a new medicine on the market is a difficult challenge. Companies have to invest huge amounts of money in R&D. In fact, the innovation process of a new medicine can take around 15 years. It usually starts within the firms' laboratories, where the new molecules are obtained. In a second stage, the companies have to demonstrate the potential effects of the new drugs on human health through clinical trials.

A clinical trial usually comprises four steps or phases. Each phase is designed to answer a separate research question. In *Phase 1*, the researchers test a new drug or treatment in a small group of people for the first time in order to evaluate its safety, determine a safe dosage range, and identify any side effects. In *Phase 2*, the drug or treatment is given to a larger group of people in order to see if it is effective and to further evaluate its safety. In *Phase 3*, the drug or treatment is given to larger groups of people in order to confirm its effectiveness, monitor its side effects, compare it to commonly used treatments, and collect information that will allow a safe use of the drug or treatment. After this phase usually the companies receive authorization from regulatory authorities to commercialize the new drugs. Finally, *Phase 4* studies are done after the drug or treatment has been marketed in order to gather information about the drug's effect on various populations and on any side effects associated with its long-term use (see Figure 1).

But companies are not charities and are driven by profit. When a new medicine is authorized, the companies have to promote their products among doctors around the world. There are several ways to promote a new medicine. One of the most effective ways is to publish the results of a clinical trial in a medical/scientific journal. Pharmaceutical companies usually send reprints of these articles to many doctors and practitioners and this has important effects on their prescription habits.

The main problem with this practice is that there is evidence that published clinical trials are biased to results in favour of new medicines (Smith, 2006). This is a relevant fact because the results of published clinical trials are used to make many important clinical decisions which affect to the patients' health (Smith, 2005b).

In recent years, some relevant members of the scientific community and important institutions (i.e.: The House of Commons) have argued that the clinical trials publication system must be modified. To overcome the problems described above, a new system to publish clinical trials has been proposed (Smith & Roberts, 2006). One of the recommendations of this new system is that medical journals should not publish any clinical trials but rather commentaries and reports of them.

In addition to these important objections, there is another aspect of clinical trials that must be considered. This kind of studies is generally conducted by ordinary doctors, not researchers, who usually follow a protocol or a list of instructions that has been designed by the pharmaceutical companies themselves. In other words, clinical trials should not be considered as research because, in fact they are part of the development of new drugs. Indeed, the last edition of the Frascati Manual

(OECD, 2002) establishes that the phases 1, 2 and 3 of clinical trials (see above Figure 1), must be considered as development and not as research activities².

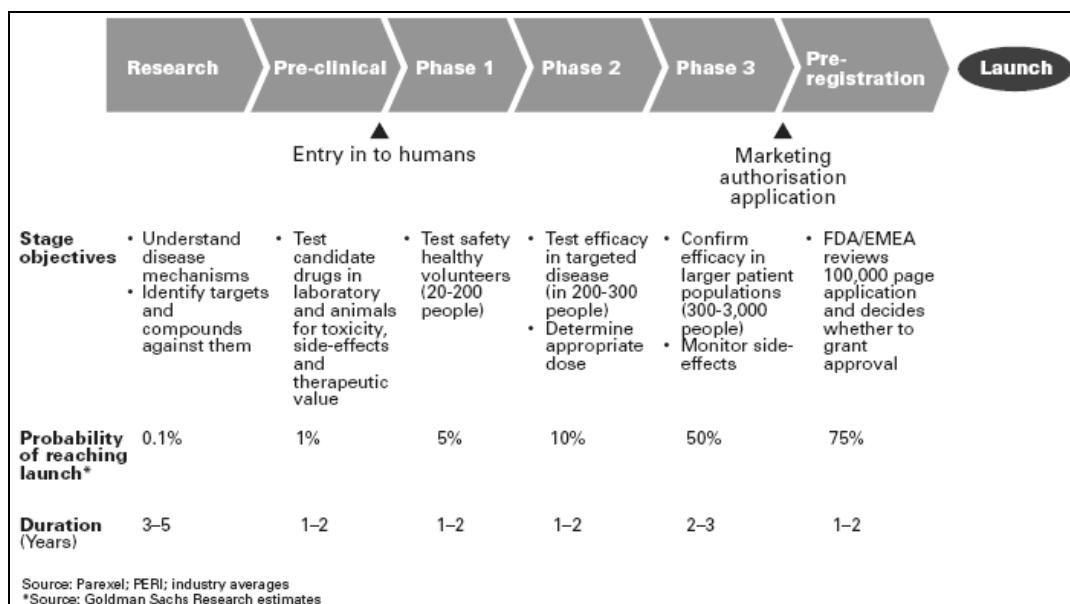


Figure 1. Drug development timeline (DTI, 2006).

Finally, clinical trials usually receive a lot of citations, but probably not because they were the best or the most important pieces of research (Smith, 2003). This high number of citations is probably due to the dissemination practices by the companies. Under these circumstances, the medical journals prefer to publish papers with a high expected number of citations. In other words, the *virtuous circle* could be transformed into a *vicious circle*.

We believe that previous arguments could have important implications on research evaluation procedures. Some of them can be formulated as research questions for this study.

1. *Are clinical trials cited significantly more than other papers?*
2. *To what extent are Journal Impact Factors affected by the citations for clinical trials?*
3. *To what extent can h-indexes be modified by the clinical trials' citations?*

There is little quantitative evidence about the previous questions; in this paper we analyze the first and the second questions. The third one, it will be considered in a specific paper that we are preparing.

Methods

To answer our research questions we have carried out two different studies with ad hoc databases. All the data have been gathered from the *Thomson-ISI Web of Science®* and *Medline*.

Firstly, in order to assess if the number of citations of clinical trials are different from the rest of the papers, we have gathered the required data as follows: (i) from WoS® database we have downloaded all the articles published in 2000 that contains the terms "*colorectal cancer*" or "*colorectal neoplasm*". For each article we have created the field "*citations*"; (ii) from Medline we downloaded the papers containing the terms "*colorectal cancer*" or "*colorectal neoplasm*". Using the thesaurus *MESH*, every article has been marked as 1 if it is clinical trial and 0 otherwise; (iii) we have merged the two datasets in order to have full information for each article (citations and clinical trial). We compare the mean values for two subsets of documents using ANOVA tests.

² As regards Pre-registration, also known as Phase 4, the Frascati Manual says that is part of the innovation process and can seldom be considered as development.

Secondly, in order to explore the effect that clinical trials have on the *Journal Impact Factors®*, we have selected for this study four of the major medical journals: JAMA, BMJ, New England Journal of Medicine and Lancet. In order to gather required data we have followed three steps for each of the four journals: (i) from the WoS database we have downloaded all citable documents corresponding to years 1998 and 1999 (dataset A₁); (ii) from Medline we downloaded only the papers indexed as clinical trials by the thesaurus MESH (dataset A₂); and (iii) we have carried out a crossed analysis of datasets A₁ and A₂ in order to identify all clinical trials of the dataset A₁. As a result, a new dataset B have been created only with papers that are not considered as clinical trials in Medline. We have used this dataset B, to compute a new Impact Factor for each journal.

Results

As regards our first analysis, we have found 1150 articles published in 2000 related to *colorectal cancer* and *colorectal neoplasm*. From those, a total of 115 clinical trials have been identified. The clinical trials have received on average 43,8 citations versus 24,4 citations received by the articles not associated with clinical trials. The ANOVA Test suggest that these distributions are significantly different ($F= 14,18$ and $p\text{-value} < 0,001$). Nevertheless, the Levenne Test of homogeneity of variance suggest ($p < 0,005$) the need to explore in depth the observed differences among the two groups. Perhaps the observed differences in citations are partially due to bias associated with the journals where articles have been published.

Secondly, as regards the *Journal Impact Factor®*, Table 1 shows the results obtained for each of the four selected journals with and without clinical trials. The columns one to three are referred to all the papers while the 4th to 6th columns contain data only regarding papers not considered as clinical trials. The columns *Articles* and *Clinical Trials* show the number of these items published in 1998 and 1999 by each of the four journals. The other two columns contain the number of citations received in 2000 for all articles (*Citations A*) or only by the articles not considered as clinical trials (*Citations B*). Finally, the columns *Impact Factor A*, and *Impact Factor B*, contains the values of the Journal Impact Factors including and excluding clinical trials respectively.

Except in the case of the *New England Journal of Medicine*, there is an important reduction of Impact Factors when the clinical trials and their citations are deleted. These figures support the hypothesis regarding the clinical trials have a significant influence in the Impact Factor values of medical journals.

From Table 1, we can also deduce that clinical trials have a significantly higher number of citations than other papers. This result is only valid for *JAMA*, *BMJ* and *Lancet*.

Conclusions

Scientific papers in biomedicine show an important dichotomy. On the one hand, there are papers based on original research, usually funded by public institutions. On the other hand, there are the papers based on clinical trials supported mainly by private firms. Recently, there has been a controversy about the appropriateness of publishing clinical trials papers in scientific journals. In order to provide additional information to research policy makers, we propose to identify and measure the contribution of the clinical trials to common bibliometric indicators, such as citation counts, *Impact Factors®* or the *h*-indexes.

Our preliminary results suggest that the effect of clinical trials can be important in terms of citation counts and Journal Impact Factors. As a consequence, we believe that this kind of information could be useful to funding research decisions.

In further research we will explore the effect of clinical trials on other bibliometric indicators such as the h-indexes. We also are going to apply new methods to identify clinical trials within ISI databases. Particularly, we are interested in differentiating between clinical trials for each of the phases and also whether or not clinical trials have been funded by private firms or not. Another interesting study can be carried out by comparing the effect clinical trials on different pathologies, regions, countries or hospitals.

Table 1. Published articles, citations, clinical trials and Impact Factor for the four selected journals.

Articles	Citations A (mean)	Impact Factor A	Clinical Trials	Citations B (mean)	Impact Factor B
<i>New England Journal of Medicine</i>					
1998	390	12878	29,512	5	197
1999	380	9846		11	1338
Total	770	22724		16	1535
<i>British Medical Journal (BMJ)</i>					
1998	916	5053	5,331	69	4647
1999	761	3887		40	1234
Total	1677	8940		298	5881
<i>Journal of American Medical Association (JAMA)</i>					
1998	494	7776	15,402	58	3172
1999	364	5439		58	3450
Total	858	13215		116	6622
<i>Lancet</i>					
1998	1009	11979	10,232	45	2789
1999	1108	9683		63	2836
Total	2117	21662		108	5625
<i>Journal of the Royal Society of Medicine</i>					
Total	770	22724		16	1535
<i>Public Library of Science Medicine</i>					
Total	1677	8940		298	5881
<i>Public Library of Science Clinical Trials</i>					
Total	858	13215		116	6622
<i>Journal of the American Medical Association (JAMA)</i>					
Total	2117	21662		108	5625
<i>British Medical Journal (BMJ)</i>					
Total	770	22724		16	1535
<i>New England Journal of Medicine</i>					
Total	1677	8940		298	5881
<i>Lancet</i>					
Total	858	13215		116	6622
<i>Journal of the Royal Society of Medicine</i>					
Total	2117	21662		108	5625
<i>Public Library of Science Medicine</i>					
Total	770	22724		16	1535
<i>Public Library of Science Clinical Trials</i>					
Total	1677	8940		298	5881
<i>Journal of the American Medical Association (JAMA)</i>					
Total	858	13215		116	6622
<i>British Medical Journal (BMJ)</i>					
Total	2117	21662		108	5625
<i>New England Journal of Medicine</i>					
Total	770	22724		16	1535
<i>Lancet</i>					
Total	858	13215		116	6622
<i>Journal of the Royal Society of Medicine</i>					
Total	2117	21662		108	5625
<i>Public Library of Science Medicine</i>					
Total	770	22724		16	1535
<i>Public Library of Science Clinical Trials</i>					
Total	1677	8940		298	5881
<i>Journal of the American Medical Association (JAMA)</i>					
Total	858	13215		116	6622
<i>British Medical Journal (BMJ)</i>					
Total	2117	21662		108	5625
<i>New England Journal of Medicine</i>					
Total	770	22724		16	1535
<i>Lancet</i>					
Total	858	13215		116	6622
<i>Journal of the Royal Society of Medicine</i>					
Total	2117	21662		108	5625
<i>Public Library of Science Medicine</i>					
Total	770	22724		16	1535
<i>Public Library of Science Clinical Trials</i>					
Total	1677	8940		298	5881
<i>Journal of the American Medical Association (JAMA)</i>					
Total	858	13215		116	6622
<i>British Medical Journal (BMJ)</i>					
Total	2117	21662		108	5625
<i>New England Journal of Medicine</i>					
Total	770	22724		16	1535
<i>Lancet</i>					
Total	858	13215		116	6622
<i>Journal of the Royal Society of Medicine</i>					
Total	2117	21662		108	5625
<i>Public Library of Science Medicine</i>					
Total	770	22724		16	1535
<i>Public Library of Science Clinical Trials</i>					
Total	1677	8940		298	5881
<i>Journal of the American Medical Association (JAMA)</i>					
Total	858	13215		116	6622
<i>British Medical Journal (BMJ)</i>					
Total	2117	21662		108	5625
<i>New England Journal of Medicine</i>					
Total	770	22724		16	1535
<i>Lancet</i>					
Total	858	13215		116	6622
<i>Journal of the Royal Society of Medicine</i>					
Total	2117	21662		108	5625
<i>Public Library of Science Medicine</i>					
Total	770	22724		16	1535
<i>Public Library of Science Clinical Trials</i>					
Total	1677	8940		298	5881
<i>Journal of the American Medical Association (JAMA)</i>					
Total	858	13215		116	6622
<i>British Medical Journal (BMJ)</i>					
Total	2117	21662		108	5625
<i>New England Journal of Medicine</i>					
Total	770	22724		16	1535
<i>Lancet</i>					
Total	858	13215		116	6622
<i>Journal of the Royal Society of Medicine</i>					
Total	2117	21662		108	5625
<i>Public Library of Science Medicine</i>					
Total	770	22724		16	1535
<i>Public Library of Science Clinical Trials</i>					
Total	1677	8940		298	5881
<i>Journal of the American Medical Association (JAMA)</i>					
Total	858	13215		116	6622
<i>British Medical Journal (BMJ)</i>					
Total	2117	21662		108	5625
<i>New England Journal of Medicine</i>					
Total	770	22724		16	1535
<i>Lancet</i>					
Total	858	13215		116	6622
<i>Journal of the Royal Society of Medicine</i>					
Total	2117	21662		108	5625
<i>Public Library of Science Medicine</i>					
Total	770	22724		16	1535
<i>Public Library of Science Clinical Trials</i>					
Total	1677	8940		298	5881
<i>Journal of the American Medical Association (JAMA)</i>					
Total	858	13215		116	6622
<i>British Medical Journal (BMJ)</i>					
Total	2117	21662		108	5625
<i>New England Journal of Medicine</i>					
Total	770	22724		16	1535
<i>Lancet</i>					
Total	858	13215		116	6622
<i>Journal of the Royal Society of Medicine</i>					
Total	2117	21662		108	5625
<i>Public Library of Science Medicine</i>					
Total	770	22724		16	1535
<i>Public Library of Science Clinical Trials</i>					
Total	1677	8940		298	5881
<i>Journal of the American Medical Association (JAMA)</i>					
Total	858	13215		116	6622
<i>British Medical Journal (BMJ)</i>					
Total	2117	21662		108	5625
<i>New England Journal of Medicine</i>					
Total	770	22724		16	1535
<i>Lancet</i>					
Total	858	13215		116	6622
<i>Journal of the Royal Society of Medicine</i>					
Total	2117	21662		108	5625
<i>Public Library of Science Medicine</i>					
Total	770	22724		16	1535
<i>Public Library of Science Clinical Trials</i>					
Total	1677	8940		298	5881
<i>Journal of the American Medical Association (JAMA)</i>					
Total	858	13215		116	6622
<i>British Medical Journal (BMJ)</i>					

Using Quasi-Experimental Design and the Curriculum Vitae to Evaluate Impacts of Earmarked Center Funding on Faculty Productivity, Collaboration, and Grant Activity¹

Monica Gaughan*, Branco Ponomariov ** and Barry Bozeman*

**gaughan@uga.edu*

University of Georgia Athens 30606, Georgia (USA)

***branco.p@gmail.com*

University of Illinois at Chicago, Illinois (USA)

Abstract

Academic research conducted within multidisciplinary science centers developed rapidly during the 1980s in response to several federal initiatives in the United States. In this study, we examine the effect of a Congressionally-mandated multidisciplinary center program on the careers of affiliated scientists. Important to the research evaluation design, we incorporate a control group of researchers who are not affiliated with these centers, but who work in the same scientific area. We collect curricula vitae from both groups, a data source that has been demonstrated to provide valid and reliable longitudinal data about academic productivity. We discuss the sampling and data collection methodology in detail, concluding that our sampling methodology worked well, while our initial attempt to collect CVs via the internet was not effective. Ultimately, we concluded that internet-based collection of CVs is not feasible; therefore we collected CVs directly from respondents via mail and email requests. In the analysis, we evaluate the impact of center affiliation on publication productivity, collaboration, and grants activity. We find that center affiliation tends to promote lower grant velocity, but greater levels of collaboration. These higher levels of collaboration increase publication productivity, but gains may be offset by the lower grants velocity. Overall, we find that centers—consistent with policy objectives—tend to foster funding stability and to increase collaborative interactions in the field.

Keywords:

curriculum vitae analysis; academic career trajectories; multidisciplinary science center

Policy Background

The increasing importance of multidisciplinary science centers in academic innovation processes (Cohen et al. 1994) has captured the attention of the United States Congress. In this research, we study the outcomes from the 101st U. S. Congressional (1990) call for enhanced support for research on the diagnostic and therapeutic aspects of human infertility. Congress instructed the National Institute of Child Health and Human Development (NICHD) to make grants or contracts for the operation of two infertility research centers and three contraception research centers. As amended (2001), the purpose of the centers is to improve methods of diagnosis and treatment of infertility and to develop training protocols and programs for the education of reproductive health researchers. The NICHD responded by creating specialized and coordinated research through the National Cooperative Program for Infertility Research (NCPIR), using the NIH Specialized Research cooperative agreement mechanism. This approach was also used to develop the Specialized Cooperative Centers Program in Reproduction Research (SCCPRR).

A number of researchers have sought to specify how the Government Performance and Results Act (GPRA) applies to federally-supported research in the United States. Roessner (2002) argues that research evaluation should be able to test the policy theory that the program seeks to forward; specifically, how do the process or practices being implemented affect the desired outcome? This line

¹ The research reported here was supported by a contract from the National Institute of Child Health and Human Development (Barry Bozeman, Principal Investigator). Analysis was supported by a CAREER grant REC 0447878 from the National Science Foundation (Monica Gaughan, Principal Investigator). We thank our colleagues Barry Bozeman, Elizabeth Corley and Jan Youtie for their work throughout the project that makes this study possible. Further, we are grateful for the able research assistance of Rebecca Arce, Timothy Atkins, Jason Epstein, Mary Feeney, Euiseok Kim, Dirk Libaers, Thomas Steiner, Alejandro Suarez, and Wesley Younger. The views reported here do not necessarily reflect those of the National Institutes of Health or the National Science Foundation.

of reasoning is further developed by Feller and colleagues (2003) who show how the Office of Management and Budget's Performance Assessment Rating Tool (PART) can relate strategic priorities to production activities and the development of performance criteria in mission agencies. The strategic priorities and external criteria are established by political entities, but programs have broad authority to specify the performance criteria and indicators. They extend the OMB criteria to include leadership and scientific labor force issues as crucial considerations. In survey research that evaluated how researchers ranked 97 different research activities, Bantilan and colleagues (2004) determined that the development of partnerships and collaborations was a particularly valuable outcome of research. These collaborative activities are an explicit policy goal in the development in multidisciplinary centers, in addition to more traditional publication and funding activities.

In this analysis, we examine how affiliation with the mandated NICHD cooperative reproductive health centers affects scientists' traditional productivity outcomes such as publications and grants awards. We also examine the impact of center affiliation on propensities to engage in collaborative research, which is often an explicit goal of multidisciplinary science centers. In addition to commonly employed econometric analyses, we implement a sample design that compares center affiliates to a matched cohort control group of unaffiliated researchers working in the same research area. This evaluative design allows us to make better estimates of center impacts. In keeping with the tradition of *Research Evaluation*, we present a high level of detail in our methodology to guide colleagues wishing to use the design in their own research.

Centers and scientific outcomes

Center-based scientific research differs from principal-investigator initiated research. There is some controversy that their non-scientific objectives detract from the principal investigator initiated research tradition (Bozeman & Boardman, 2004; Gray, 2000). Because they are constituted by complex funding streams and requirements of inter and intra-organizational cooperation, centers tend to be managerially more complex than principal investigator funded research (Gray, 2000). Both Gray (2000) and Rogers and Bozeman (2001) emphasize the importance of studying processes and impacts as part of the evaluative framework for multidisciplinary science centers.

A literature is emerging that examines the effects of multidisciplinary centers on individual scientists' careers. Such work began on a purposive sample of scientists affiliated with National Science Foundation supported Science and Technology Centers and Department of Energy supported Engineering Research Centers. The curricula vitae of affiliates of those Centers were collected and coded to support several analyses of career dynamics. Lee and Bozeman (2005) examined the mutually reinforcing dynamics of collaboration, productivity, and grants. In related work, Gaughan and Robin (2004) showed how entry into the first academic career differs between French and American scientists, while Gaughan and Bozeman (2002) showed how publication productivity and grants increased rates of promotion to full professor. Dietz and Bozeman (2005) demonstrated that professors with prior industrial experience tended to have different kinds of academic careers; in particular, they are more likely to patent. Corley et al. (2003) found that women received smaller average first grants, but were otherwise similar to their male colleagues.

In later work, researchers using a representative sample survey of academic scientists affiliated with Carnegie Research Extensive universities, Corley and Gaughan (2005) found that center affiliation tended to increase researchers' research-related activity. Using the same data, Bozeman and Gaughan (forthcoming) found that center affiliates engage in more research activities with industry. This work will be extended in the future to include coded curricula vitae to develop understanding of career trajectories in a representative sample. Using only Center-based scientists from the representative sample, Boardman and Ponomariov (2007) found that junior faculty tended to have more negative views of commercial activities, seeing them as distracting from basic research.

Research Design

Curriculum vitae (CV) analysis is in many respects ideally suited to evaluating the impact of the multidisciplinary centers, especially with respect to the influence of centers on career development. Almost all researchers have a CV and, in most instances, the CV provides a valid and reliable data

source (Dietz et al., 2001). Almost all CVs include information on employment history, publications and other scholarly output, the constructs of primary interest here. Most important, this information can be viewed, for practical purposes, as longitudinal data. This attribute of CV data means that it is well adapted to understanding career trajectories and tracking career changes, especially changes in research productivity and other research related activity.

While the analysis of CVs seems conceptually straightforward, the approach has thus far been confined to studies by researchers affiliated with the Research Value Mapping Program, a research consortium of Georgia Institute of Technology, University of Georgia, Arizona State University, and University of Illinois at Chicago (Gaughan & Bozeman, 2002; Lee & Bozeman, 2005; Dietz & Bozeman, 2005; Corley, Bozeman, & Gaughan, 2003). The limited application of the approach thus far is owing to the technical challenges of using CV data as well as the enormous amount of work required to code and analyze CVs. CV analysis is uniquely well suited for evaluations focusing not just on discrete outputs but also research capacity and “scientific and technical human capital” (Bozeman, Dietz & Gaughan, 2001; Bozeman & Corley, 2004). For an overview of the strengths and weaknesses of the method see Dietz, et al. (2001).

Central to the approach of this CV analysis is the comparison of persons who have received NICHD Center support with an appropriate comparison group of researchers who have not. The focus of the comparison is publishing and grants productivity and collaboration patterns. The researchers spent a great deal of time developing a sampling frame that would adequately represent both center affiliates, and reproductive researchers who are not affiliated with centers. A comparison of the characteristics of center and comparison researcher samples shows similarities on the majority of indicators included in the study. This allows us to make robust comparisons between the groups.

As in prior work, the researchers developed indicators of career trajectories, publication histories and grant awards. In addition, the researchers developed new indicators that were particularly relevant to careers in academic medicine. These included medical training, areas of specialization, and an attention to medically-related industry work, including patents. The researchers also developed a new coding methodology that is based on an ACCESS Graphic User Interface (GUI) and database, which increased the quality and reliability of the final database.

Sampling Methodology

The problem of comparison is well-known in research evaluation, particularly in the European context in which scientists working in the same area are employed in very different employment institutions (Bonaccorsi & Daraio, 2003). In this work, we examine researchers working within the same scientific system, reducing the difficulties in making comparisons between groups.

Target Population and Sampling

We faced a complex task with respect to the definition of the target populations and construction of sampling frames. The target populations for CV collection are were all Ph.D. or M.D.-level Center researchers. We obtained lists of participants from the three NCIPR centers; 4 SCCPRR centers provided lists of senior and trainee affiliates of the centers.

Sample Frame for Control Group of Researchers:

The construction of the sampling frame for the control group of researchers was more complex than that for the center affiliates. We developed the control group of CV targets from NIH, from Community of Science public directories, and from peer nominations elicited during face-to-face semi-structured interviews. From NIH, we collected names of grantees and trainees doing infertility research, but not affiliated with any of the centers. We used the NIH grants data base from 1991 on to identify “areas of science” by the following key words: infertility, fertility, contraception, and reproductive research. These same key words were used in the COS Web of Science to develop the comparison sampling frame. The COS key word search strategy identified researchers working in areas related to human fertility, but not funded through the Centers, or identified through the NIH lists.

Finally, we augmented the comparison groups through sociometric nomination. First, we asked respondents in NCPIR semi-structured interviews to nominate peers. To create the cohort comparison group, we asked: "Name two or three peers in graduate school with whom you shared similar scientific interests." To extend the scientific peer group comparison, we asked respondents: "Please name two or three scientists in fertility research whose work is particularly close to your own." The total developed target sample frame was n=648.

Data Collection

Earlier work (Dietz et al., 2000) determined that too few researchers posted their CVs on the internet for this to be a valid or reliable means of collecting such data. Since four years had passed since our last major CV collection, we were interested to know if the Internet has improved as a source for the collection of such documents. We were interested in both the availability of target CVs, and in the completeness of them (i.e. full or truncated CV; and the last posting).

In COS, we found the CVs for 14 out of 43 that were sought, for a rate of 33%. These CVs were generally not extremely helpful, with the focus placed on current research interests over the career path information we use for developing indicators. These CVs ranged from being updated within the past year to being updated over seven years ago. On average, only 23% of the researcher's publications were listed, with a smaller percentage of grant information listed (often not listed at all.) Only 40% of those with a patent on the mailed CVs listed their patent in the Community of Science profile.

Outside the Community of Science, we attempted a general Internet search to find CVs and assess their value. This search proved to be less helpful than the Community of Science approach. Generally all that were listed on the researcher's websites were their research interests and selected journal articles, with no information on previous positions held. Furthermore, trainees (past or present) almost never had a presence on the Internet, making it difficult to track them (professionally) using the Internet. In addition, none of the information was in a standardized format, unlike the Community of Science, which was standardized.

In summary, several years have passed since the last collection of CVs for an RVM project. At that time, we determined that CVs were not available via unobtrusive means such as the Internet. Given the passage of time, we conducted a new study to determine if the Internet has improved as a source of CVs. It has not: CVs cannot be collected unobtrusively via the Internet because of their age, truncation, and missing data. Therefore, as in the past, we made written requests for CVs via email, as this is the easiest compliance method for the target scientists.

Having eliminated the internet as a viable source of CVs, we proceeded with a mail and email based strategy to request them from our target sample. We collected the CVs from center-affiliated researchers during the spring of 2004, and collected the control group sample between July and October 2004 in three successive waves of email requests to the target sample respondents. We made a fourth effort to collect CVs from center affiliates in December and January, 2005.

We achieved an overall response rate of 40%, which is less than that achieved in other samples of scientists and engineers we have studied. This may be due to a difference in the culture of academic medicine, the relationship between academic researchers and NIH, design effects, or some other factor. We did achieve an overall response rate of 56% for NCPIR Center participants and 53% for SCCPDR Center participants, which was more in line with what we expected to obtain.

Measurement

We developed the CV coding instrument to capture constructs of interest reliably and in a reasonable amount of time. We followed the reliability methodology established in previous work (Dietz et al., 2000). The initial testing helped us to conclude that ACCESS was a feasible way to code and manage the data. We further determined that the user interface resulted in more reliable data input (fewer coder errors) than the prior generation of RVM coding protocols, which were .html based. Two

versions of the coding protocol were tested using four coders on two sets of 10 CVs. For the final protocol, inter-coder reliability was above .85 (Crittenden & Hill, 1971). The final database includes the indicators on demographic, educational, employment and research productivity.

We specify three basic measurement models: grant rate, collaboration rate, and publication rate as functions of independent covariates. The grant rate concept captures the velocity with which researchers earn grants; it is the total number of grants divided by the professional age of the researcher. The collaboration rate is conceptualized as the number of coauthors since 1990 divided by the number of peer reviewed article publications since 1990. It is best thought of as a propensity to collaborate variable. It should not be used as a measure for the size of co-authorship networks, as many of the coauthors used in the numerator are duplicated (i.e. the target respondent has published with the coauthor more than once in the period). The publication rate is calculated by dividing the number of peer reviewed article publications by the years of effective publishing activity. Effective publication activity is calculated by the first year of publishing activity since 1990. For example, if a person's first publication occurs in 2000, then the years of effective publication activity is 5. All of the senior researchers (who were productive prior to 1990) have effective publication activity years of 15.

Analytic Methods

Analytic techniques for publication productivity are well established. Furthermore, recent work in the Research Value Mapping research program has validated using CVs for this type of study. Dietz (2004) examines the impact of job changes on productivity, testing four hypotheses about publication productivity: diversity, homogeneity, education and training, and precocity. Lee (2004) examines the effect of being foreign born on the productivity of scientists; most important for our purposes in the current study, he validates the use of normal counts of annual publication productivity. Together, these studies also identify the appropriate independent variables: center affiliation, postdoctoral position, publication precocity, traditional/nontraditional career track, and cohort (Dietz 2004); and collaboration, foreign born, and gender (Lee 2004). In addition, each study points to the importance of discipline. In the NICHD context, we conceptualize this variable as training type (M.D. and Ph.D.).

Stated formally, the basic econometric model is represented as follows:

$$Y' = a + B_{11} X_1 + X_2' B_{12} + X_3' B_{13} + B_{14} X_4$$

Where: Y=Vector of Publication Productivity Variables
 X₁=Center Affiliation
 X₂=Vector of Position Variables

X₃=Vector of Demographic Controls

X₄=Collaboration Pattern

Not surprisingly, grant, collaboration, and publication rates tend to be skewed with moderate kurtosis. Since we are using Ordinary Least Squares Regression (OLS) to estimate the determinants of each rate, the non-normality of the dependent variable may bias the estimate of effects, and also can affect inferences based on significance tests. Therefore, we estimate each model using a log transformation of the dependent variable, which tends to normalize the distribution, making it more suitable to the assumptions of the OLS model.

Results

Descriptive Statistics

In the Table 1, we report descriptive statistics for the sample as a whole, and the means for each sample, center-affiliated and unaffiliated researchers. We use independent samples t-tests (two-tailed) to test for differences in means between the groups. Note that roughly one-half of the obtained sample is Center affiliated, creating a balanced design between "treatment" and controls. The pooled sample is heterogeneous in its design. Respondents are drawn from Centers, NIH grantee data, and Community of Science. It is therefore quite interesting to note few statistically significant differences

between Center researchers (senior and trainees) and Control researchers (senior, junior, and trainees). More than two-thirds is male, reflecting the later career age of the sample as a whole. On average, the whole sample completed its education sixteen years ago. Sixty-one percent earned a Ph.D., and 40% are medical doctors.

Table 1. Bivariate comparison of center-affiliates with unaffiliated control researchers

	N^a	Whole Sample Mean	Center Affiliated Mean	Unaffiliated Control Mean
Demographic				
<i>Gender</i>	175	.68	.58	.77
<i>Year of Last Degree</i>	173	1989	1990	1988
<i>Have Ph.D.</i>	175	.61	.68	.55
<i>Have M.D.</i>	175	.40	.37	.43
Career Event				
<i>Center Affiliation</i>	175	.48	--	--
<i>Post-doc Start Year</i>	118	1989	1990	1988
<i>Assistant Professor Start Year</i>	107	1989	1988	1989
<i>First Grant Year</i>	120	1990	1991	1990
<i>First NIH Grant Year</i>	101	1992	1991	1993
<i>First NIH PI Grant Year</i>	72	1993	1992	1995
<i>Any Nontraditional</i>	175	.28	.24	.32
Productivity				
<i>Years of publishing</i>	175	11.55	10.13	12.12
<i>Total publications since 1990</i>	175	39.95	29.86	48.90
<i>Publication rate since 1990</i>	175	2.94	2.35	3.46
<i>X publication rate^c</i>	173	2.53	2.35	2.67
Collaboration				
<i>Total number coauthors</i>	175			
<i>Co-authorship rate</i>	177	4.58	5.03	4.13
<i>X co-authorship rate^c</i>	173	4.54	5.03	4.07
Funding				
<i>Total number of grants</i>	175	9.22	7.26	11.02

(a) Sample sizes of less than 178 are because the mean is calculated only for those respondents experiencing the event. For example, 107 of 175 respondents have ever started in an Assistant Professor position. (b) Significant differences (at the .05 level or better) between Center and non-Center affiliates are shown in bold. (c) There are two significant outliers, who publish and co-author substantially more than the rest of the sample. They are valid observations. However, the distributions of the variables are highly skewed with their inclusion. Therefore, we report results by each indicator. The "X" variable indicates that the two outliers have been removed from the distribution.

Professional characteristics are remarkably similar between the two groups. Twenty-eight percent has at least some government or industry experience. They enter postdoctoral positions (n=118) and assistant professorships (n=107) at the same time. They receive their first grant (n=120), their first NIH grant (n=101), and their first NIH grant as PI (n=72) at the same time. They are similarly productive in their publication activity, both in absolute productivity and in terms of rate.

Indeed, the only significant differences between the pooled Center and pooled Control samples are in collaboration and grant activity. Center researchers have significantly more co-authors, working with, on average, one additional researcher per year. By contrast, the Control researchers report significantly more grants for funded research, almost four on average. This may reflect the greater instability in funding experienced by unaffiliated researchers. The purpose of this bivariate analysis was to evaluate potential differences between the sample populations; we conclude that the two samples are similar in their characteristics, constituting a balanced comparative design.

Productivity Model Results

Table 2 shows the results of three models of research productivity: grant rates, collaboration rates, and publication productivity.

Grant Rates

Looking across models in the grant productivity table, the principal independent variable of interest, center

Table 2. Productivity models for the pooled sample of 175 respondents

	Grant Rate	Collaboration Rate	Publication Rate
Any Center	-0.54*	0.23*	-0.09
Demographic			
Male	0.09	-0.11	0.12
MD	0.64***	-0.22	0.18
Career Events			
<i>Ever Postdoctoral</i>	-0.10	0.22	-0.09
<i>Nontraditional</i>	-0.24	0.27*	-0.25
<i>Ever Assistant Professor</i>	0.59*	-0.38***	.44***
<i>Ever NIH Grant</i>	2.48***	-0.02	-0.10
Career Velocity			
<i>Log of Collaboration Rate</i>	-0.02	--	0.83***
<i>Log of Publication Rate</i>	0.26	0.84***	--
<i>Log of Grant Rate</i>	--	0.92***	0.08
Intercept	-3.62***	0.92***	-0.64***
R-squared	0.6	0.73	0.74

Ordinary Least Squares Regresión // Logged Productivity Rates⁺ on Career Covariates and Controls //+ Rates calculated by total number divided by number of career years since last degree//Natural logarithm of rates taken to adjust for skew and kurtosis//Note: Statistical significance levels: *= $p < .05$; **= $p < .01$; ***= $p < .001$

affiliation, has a statistically significant negative effect on grant award velocity². This persistent finding makes sense given that one of the theories of large center grants is to provide stable funding over a number of career years, thus reducing pressure on the investigator to apply continually for new grants, or for lowyield grants³. There are other consistent and interesting patterns revealed. First, in this and for all other productivity models, gender does not play a role in obtaining grants. To the extent that it is a policy choice not to discriminate against (or favor) grantees on the basis of their sex, this non-effect should be interpreted as a positive policy outcome. Those who have MD degrees are significantly more likely to seek and obtain grant funding. Again, to the extent that supporting clinical research by those with at least some clinical background is a policy objective, this is a strong indication of positive policy effect. Not surprisingly, ever having received NIH funding increases the grant rate (which is not adjusted for the number of NIH grants; therefore, not much should be made of this variable in this particular model). Overall, 60% of the variance in the log of the rate is explained by the full models. It is typical for econometric models using logged variables to inflate the variance explained, in part by shrinking the variance toward the mean (a result of normalization of the variable).

² In some research, the issue of whether grant funding is an appropriate performance indicator is raised (Laudel 2005). While we agree that the success of a proposal is affected by a number of factors, this research design controls for some of the most important bases of differentiation.

³ It would be interesting to know grant amounts, but preliminary work determined that information about funding is inconsistently reported on curricula vitae.

Collaboration

In the collaboration column, note the significant positive effects of Center affiliation. Again, to the extent that encouraging collaboration through co-authorship is an objective of the Centers, this indicates a strong positive effect of Centers on the desired outcome. Neither demographic nor career events are particularly important predictors of collaboration, but note the strong positive effect of a nontraditional background, and the strong negative effect of ever entering an academic track. Given the norms of collaboration outside academe, and of the reward structure of non-collaboration within, these results are not surprising. Also not surprising is the significant positive effect of publications on co-authorship (and vice versus). This set of models, and the subsequent set of publication productivity models give some leverage on the reciprocal nature of this relationship, although we do not model it directly. Overall, 73% of the logged rate variance is explained by the model.

Publication

In the analysis of publication productivity (right panel), center affiliation has no impact. Those who have ever entered a tenure track publish significantly more than those who have not. This makes sense in terms of the work incentives of the academic sector. Other demographic and career event variables are not interesting. As implied by the prior model, co-authorship has a significant positive effect on publication productivity. Grant rate, too, has a significant positive impact on publication productivity. Three-quarters of the log rate variance is explained by the model.

Conclusions

Multidisciplinary science centers are an important component of the nation's scientific research enterprise. They are often developed with explicit policy goals and objectives that set them apart from traditional principal investigator-initiated research. In this study of Congressionally-earmarked research centers on reproductive health, we were interested to evaluate the extent to which center affiliation affects traditional productivity indicators such as grants and publications productivity. We were further interested in the less traditional indicator of collaboration patterns, which are valued outcomes by policy makers. Our target sample was the researchers affiliated with these earmarked centers, presenting the challenge of developing a relevant control group. Using multiple sources, we drew a sample of researchers working in the same areas as the center affiliates, but who were themselves unaffiliated. The characteristics of the two groups were quite similar, allowing the use of quasi-experimental design to evaluate multivariate differences.

Our statistical results indicate that center affiliation has some results in the hypothesized direction. We find that center affiliation tends to depress grants velocity while increasing collaboration rates. Since funding stability is one important goal of multidisciplinary science centers, the negative effect on grants velocity can be interpreted as consistent with program goals. Similarly, since improving research networks is one of the explicit goals of these science centers, the positive effect of center affiliation on the tendency to collaborate through authorship should likewise be viewed as a positive program outcome. The non-effect of center affiliation on publication productivity indicates that researchers in this field are similarly productive, independent of organizational affiliation. This may allay concerns that center-affiliates are less productive researchers than those supported exclusively through principal investigator-initiated research.

Generalizability of these specific findings may be limited to those working in the field of reproductive health research. We believe, however, that we present a methodology that allows the investigation of center-specific effects on scientific work patterns in other areas as well. The use of the Community of Science and unaffiliated funded researchers in the field allows direct examination of center effects in a way that is impossible to accomplish by relying on exclusively center-based research designs. Furthermore, the use of the curriculum vitae allows for reliable collection of target indicators, and a longitudinal component that helps to specify causal mechanisms. Future research should specify other scientific fields that are stratified by multidisciplinary science centers to evaluate whether the phenomena of greater funding stability and propensity to collaborate holds.

References

- Bantilan, M.C.S., Chandra, S., Mehta, P. & Keating, J.D.H. (2004). Dealing with diversity in scientific outputs: Implications for international research evaluation. *Research Evaluation*, 13, 87-93.
- Bonacorsi, A. & Cinzia D. (2003). A robust nonparametric approach to the analysis of scientific productivity. *Research Evaluation*, 12, 47-69.
- Boardman, P. C. & Ponomariov, B. (2007). Reward systems and NSF university research centers: the impact of tenure on university scientists' valuation of applied and commercially-relevant research. *Journal of Higher Education*, 78, 51-70.
- Bozeman, B. & Boardman, P. C. (2004). The NSF Engineering Research Centers and the university-industry research revolution: A brief history featuring an interview with Erich Bloch. *Journal of Technology Transfer*, 29, 365-375.
- Bozeman, B. & Corley E. (2004). Scientists' collaboration strategies: Implications for scientific and technical human capital. *Research Policy*, 33, 599-616.
- Bozeman, B., Dietz, J. & Gaughan, M. (2001). Models of scientific careers: Using network theory to explain transmission of scientific and technical human capital. *International Journal of Technology Management*, 22, 716-740.
- Bozeman, B. & Gaughan, M. (In Press). Impacts of grants and contracts on academic researchers' interactions with industry. *Research Policy*.
- Cohen, W., Florida, R. & Goe, R. (1994). *University-Industry Research Centers in the United States*. Pittsburgh: Carnegie Mellon University.
- Corley, E. & Gaughan, M. (2005). Scientists' participation in university research centers: What are the gender differences? *Journal of Technology Transfer*, 30, 371-381.
- Corley, E., Bozeman, B. & Gaughan, M. (2003). Evaluating the impacts of grants on women scientists' careers: The curriculum vita as a tool for research assessment. In P. Shapira and S. Kuhlmann (Eds.), *Learning from Science and Technology Policy Evaluation: Experiences from the U.S. and Europe*. Cheltenham, UK: Edward Elgar Publishing.
- Crittenden, K. & Hill, R. (1971). Coding reliability and validity of interview data. *American Sociological Review*, 36, 1073-1080.
- Dietz, J. & Bozeman, B. (2005). Academic careers, patents, and productivity: Industry experience as scientific and technical human capital. *Research Policy*, 34, 349-367.
- Dietz, J. (2004). Scientists and Engineers in Academic Research Centers: An Examination of Career Patterns and Productivity. Dissertation at Georgia Institute of Technology.
- Dietz, J., Chompolov, I., Bozeman, B., Lane, E. & Park, J. (2001). Using the curriculum vita to study the career paths of scientists and engineers: An exploratory assessment. *Scientometrics*, 49, 419-442.
- Feller, I., Gamota G. & Valdez, W. (2003). Developing science indicators for basic science offices within mission agencies. *Research Evaluation*, 12, 71-79.
- Gaughan, M. & Bozeman, B. (2002). Using curriculum vitae to compare some impacts of NSF research center grants with research center funding. *Research Evaluation*, 11, 17-26.
- Gaughan, M. & Robin, S. (2004). National science training policy and early scientific careers in France and the United States. *Research Policy*, 33, 569-581.
- Gray, D. O. (2000). Government-sponsored industry-university cooperative research: an analysis of cooperative research center evaluation approaches. *Research Evaluation*, 9, 57-67.
- Laudel, G. (2005). Is external research funding a valid indicator for research performance? *Research Evaluation*, 14, 27-34.
- Lee, S. & Bozeman, B. (2006). The impact of research collaboration on scientific productivity. *Social Studies of Science*, 35, 673-702.
- Lee, S. (2004). Foreign-born scientists in the United States: Do they perform differently than native-born scientists? Dissertation at Georgia Institute of Technology.
- Lin, M. & Bozeman, B. (2006). Researchers' industry experience and productivity in university-industry research centers: A scientific and technical human capital explanation. *Journal of Technology Transfer*, 31, 269-290.
- Roessner, D. (2002). Outcome measurement in the US: the state of the art. *Research Evaluation*, 11, 85-93.
- Rogers, J. D., & Bozeman, B. (2001). Knowledge value alliances: an alternative to R&D project evaluation. *Science, Technology and Human Values*, 26, 23-55.
- United States Congress. (2001). Compilation of selected acts within the jurisdiction of the committee on energy and commerce: health law as amended through December 31, 2000.
- United States Congress.(1990). Departments of Labor, Health and Human Services, and Education, and related agencies appropriation bill. Report 101-127.

Mapping the Changing Centrality of Physicists (1900-1944)

Yves Gingras*

*gingras.yves@uqam.ca

Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, CP 8888, Succursale Centre-ville, Montréal, Québec, H3C 3P8 (Canada)

Abstract

This paper contributes to historical bibliometrics in proposing to apply social networks methodology and the concept of centrality to co-citation networks. Using the case of physics for the period 1900-1944 we show that measuring the centrality of authors in co-citation networks provides a useful index of the evolution of a scientific field (in this case physics) and the changing focus of research over time. We divided the 45-year period into seven sub-periods related to major events in physics (1900-1904; 1905-1911; 1912-1918; 1919-1924, 1925-1930; 1931-1936; 1937-1944) and calculated the centrality of actors present in the co-citation network for each of these periods. The results obtained reflect the major changes in the disciplines: we see the evolving rise and decline in centrality of major physicists like H.A. Lorentz, A. Einstein, N. Bohr, E. Rutherford as well as others less known figures as the “hot” topics move from black body and electron theory to relativity and spectroscopy and from quantum mechanics to nuclear physics.

Keywords

history of physics; co-citation network; centrality, network analysis.

Introduction

During the First half of the 20th century, the field of physics has been transformed by many conceptual revolutions: quantum physics in 1900, relativity in 1905 and 1916 and quantum mechanics in 1925-1926. The centers of interests of physicists have thus moved from black body theory to atomic physics and spectroscopy and then nuclear physics in the mid-1930s. Historians of science have shown us in detail that the main actors of these transformations were Max Planck, Albert Einstein, Niels Bohr, Ernest Rutherford, Arnold Sommerfeld, Werner Heisenberg, Erwin Schrödinger, Wolfgang Pauli and Paul Dirac to name only the major ones, all of them but Sommerfeld being Physics Nobel Prize winners. Though the traditional tools of historians of science are well suited to identify and follow individuals in their career and scientific production, they often leaves in the dark the more global trends affecting the field as a whole. And it is worth noting that despite frequent use of the term “scientific community” and “discipline” in many historical papers, most of them use an individualistic approach and concentrate on a few scientists using the usual resources of analyzing the content of their papers (and sometimes the correspondence), taking for granted that they probably represent the trend of the whole community and that empathy and immersion make it possible to read the papers as scientists did at the time. Bibliometrics in relation with methods developed for the analysis of social network now provide tools for an historical bibliometrics that can capture the evolving structure of a scientific field.

Already in 1964, Eugene Garfield had made a first attempt at historical bibliometrics using citations to reconstruct the history of scientific specialties (Garfield, Sher & Torpie, 1964). In 1981, ISI, under the leadership of Henry Small proposed a first historical citation index covering physics in the 1920s, but it did not give rise to many detailed analysis (Small, 1986). More recently Garfield came back to his original project and developed the software *Histcite* that automatically maps the genealogy of papers related through a citation history (Garfield, 2004). With the newly available *Century of Science* database covering the period 1900-1944, we are now in a position to have a fresh look at the evolution of science in the 20th century. What we propose here is to combine bibliometrics and the methodology of social networks in order to provide a way to analyze the structure of the community as a whole, instead of following the trajectory of a particular paper through its citations. The field can thus be mapped using co-citations (Small, 1977, 1978; Gmür, 2003), a tool adapted to the analysis of a whole

community, as opposed to the biographical and conceptual analysis of a single or a few scientists. Also, calculating the centrality of actors in the co-citation network for a series of time intervals makes possible the analysis of the global transformation of the field of physics over a long time period (here the period 1900-1944).

Using data from Thomson Scientific *Century of Science* database, we will map the evolving co-citation networks and use the degree of centrality as an indicator of the changing positions of the physicists in theses networks. By ranking physicists according to their centrality for each period provides an index of their rise and decline in the field over time, which is complementary to the ranking in terms of total citations per period¹. Whereas centrality is usually associated with *social* networks we use it here for *conceptual* networks where the link is a co-citation instead of a social relation (Freeman, 1978/1979; Wasserman & Faust, 1994)².

After briefly presenting the source data, we will comment on the main the results obtained using centrality as an indicator of eminence in the field of physics.

Methods

Thomson Scientific *Century of Science* database for the period 1900-1944 comprises papers and references published in 266 journals³. We have grouped them by disciplines and assigned 27 journals to physics largely defined (including astrophysics and astronomy), 14 to mathematics, 34 to chemistry and 10 to “general” since they cover magazines like *Science* and *Nature* as well as Academy journals that are multidisciplinary and cover physics, mathematics as well as other disciplines. As Table 1 shows, our analysis of the citations and co-citations in physics over the period 1900-1944 is based on more than half a million references (555 123) contained in nearly sixty thousands (59 950) papers. In Journals like *Nature*, many articles have no author names though they may contain references. There are also many papers without references and the average number of references per paper containing at least one reference for all fields is 10.4 and varies according to discipline between 5,3 for general journals to 13,2 for chemistry. In this paper, we will only look at co-citations to authors in physics journals. The same method could be applied to the other fields or to all of them together in order to see the relations between disciplines. This will be done in future papers.

Table 1. Papers and References in Selected Source journals (1900-1944)

Field	Papers	Papers with references	References	References/ Paper
Chemistry	122 902	97 935	1 295 014	13,2
General	112 087	57 188	304 884	5,3
Mathematics	13 496	10 429	82 562	7,9
Physics	59 950	50 275	555 123	11,0
Total	308 435	215 827	2 237 583	10,4

In order to follow changes in the field of physics, we have divided the data into seven periods (1900-1904; 1905-1911; 1912-1918; 1919-1924, 1925-1930; 1931-1936; 1937-1944), which correspond roughly to major periods in the history of physics before 1945. Although there is some arbitrariness in

¹ Comparing the rankings from both methods show a very strong rank correlation of 0.8 for the first two periods (1900-1911); it diminishes to less than 0.5 in the periods 1912-1936 to rise again to 0.74 in the last period (1937-1944). This is probably an effect of the dispersion of the field due to the rising number of actors and citable papers. As more specialties exist which are not all closely connected, one can be highly cited without being strongly connected in a network through high co-citations, hence the decline of rank correlation between centrality and citation. This suggests that both indicators provide useful and different information on the dynamic of the field.

² On co-citation see the recent review by Gmür (2003).

³ For details on the *Century of science* database and the list of journals see <http://scientific.thomson.com>. Some journals have merged over time or changed titles.

this process we have tried to keep periods of about the same duration while also taking into account more active periods like that surrounding the emergence of quantum mechanics (1925-1929). We have also manually standardized the most cited authors who appear under different forms (like Einstein and Einstein A. and other minor orthographic errors). Finally, we have applied a variable cut-off of minimal co-citations in order to include in the map only the most co-cited authors. As could be expected the number of co-cited authors rise rapidly over the period from about 1500 in 1900-1904 to more than 8000 in 1937-1944. In the first map we used a cu-off of 9 co-citations whereas in the last period we rised the cut-off to 17 co-citations. The maps have been constructed with UCINET (Borgatti, Everett & Freeman, 2002) and NETDRAW (Borgatti, 2002) social network analysis programs developed by Steve Gorgatti. The UCINET program also computes the centrality of each node as the total number of links of that node with all the others nodes in the map.

Results

Space does not permit us to show all maps and we reproduce here only four of them showing the most important changes. Figures 1, 2, 3 and 4 show the co-citation network for the period 1900-1904; 1912-1918, 1925-1930 and 1937-1944 respectively. For all the periods we have calculated the centrality of all actors present in the maps. Table 2 provides the names of the most central physicists for each time period. The shaded regions show authors who have ranked among the top ten for two consecutive periods or more.

Table 2. Centrality Rankings of Physicists Over Time (1900-1944)

Names	1900-04	1905-11	1912-18	1919-24	1925-30	1931-36	1937-44
ABRAHAM, M	26	7	19	96	488	-	-
BETHE, HA	-	-	-	-	-	34	1
BHABHA, HJ	-	-	-	-	-	438	2
BIRGE, RT	-	-	-	-	7	15	123
BLACKETT	-	-	-	-	-	75	4
BOHR, N	-	-	16	2	16	151	47
BORN, M	-	36	42	4	4	29	35
BREIT, G	-	-	-	-	-	25	3
COMPTON, AH	-	-	-	17	9	1	28
DEBYE, P	-	-	3	3	10	2	15
DIRAC, PAM	-	-	-	-	17	5	33
DRUDE, P	2	2	15	101	156	268	410
EINSTEIN, A	-	10	4	6	38	-	-
FERMI, E	-	-	-	-	185	9	7
FRANCK, J	-	-	18	5	11	152	-
HARTREE, DR	-	-	-	-	-	8	143
HEISENBERG, W	-	-	-	131	1	4	6
HUND, F	-	-	-	-	3	41	-
JAUNCEY, GEM	-	-	-	172	87	10	204
JOHNSON, TH	-	-	-	-	-	17	10
LENARD, P	5	4	8	9	37	91	181
LORENTZ, HA	13	8	12	13	165	-	-
LUMMER, O	7	28	106	-	-	-	-
MOTT, NF	-	-	-	-	-	54	5
MULLIKEN, RS	-	-	-	-	5	6	36
NERNST, W	24	24	6	-	143	-	-
PASCHEN, F	10	16	13	16	22	186	-
PAULI, W	-	-	-	47	6	129	30
PLANCK, M	6	5	5	10	108	-	-
ROSSI, B	-	-	-	-	-	21	8
RUTHERFORD, E	21	13	11	14	118	42	185
SCHR...DINGER, E	-	-	-	86	8	63	-
SLATER, JC	-	-	-	459	84	3	12
SOMMERFELD	-	38	9	1	2	27	77
STARK, J	4	3	1	8	43	-	-
THOMSON, JJ	1	1	2	7	34	-	-
VOIGT, W	22	9	10	29	-	92	-
WALLER, I	-	-	-	454	75	7	78
WARBURG, E	9	6	26	50	-	-	-
WIEDEMANN, E	8	20	138	109	-	-	-
WIEN, W	3	14	7	19	70	-	-
WIGNER, E	-	-	-	-	150	38	9

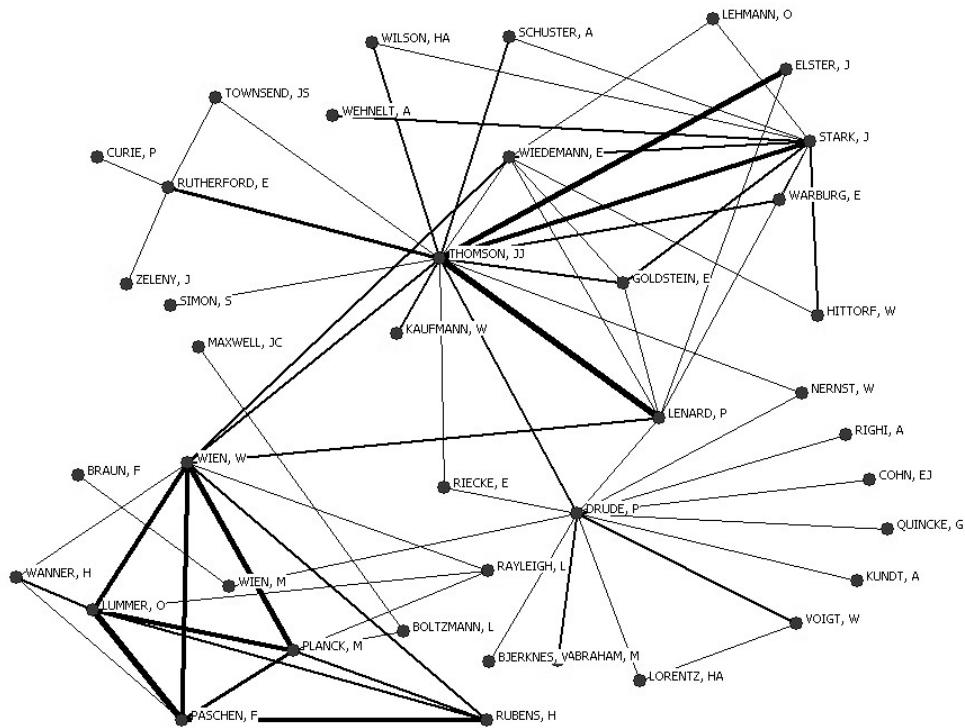


Figure 1. Co-citation network of physicists, 1900-1904 (More than 8 co-citations).

A striking feature of Table 2 is that authors who were central in the first period (1900-1904) disappear from the “top ten” at most 25 years later. The declining centrality of authors suggests an important restructuring of the field at most every generation (25 years). Over the first quarter of the 20th century, only Philip Lenard, Max Planck, Johannes Stark and Joseph J. Thomson continuously remained central actors. All were closely related to some aspects of radiation and electric conductivity through gases and they all received a Physics Nobel Prize⁴. Einstein emerged in the period 1905-1911 and then rapidly declines after 1925. The complete renewal of central actors after 1925 is a generational shift and also a complete renewal of the conceptual landscape with the emergence of quantum mechanics followed by nuclear physics. As the distribution shows, it is impossible to stay among the ten most central actors more than 25 years. In addition to those already mentioned, only Peter Debye (Nobel Prize of Chemistry in 1936) managed such a feat over the period 1912-1936 among the 42 authors included in Table 2.

The rise of Einstein is directly visible in the evolution of his degree of centrality: while he was absent in the period 1900-1904, he jumps to the 10th position in the 1905-1911 map and to the 4th in the 1912-1918 map and slightly down to the 6th in the next period (1919-1924). The decline in centrality of Lorentz in the period 1912-1918 suggests that electron theory has become obsolete with the rise of Einstein theory of relativity. Einstein continued to be a central actor through his development of general relativity as well as his contributions to quantum theory (Pais, 1982). Like many of his generation, he rapidly declines in the period 1925-1930 and disappear from the network of highly co-cited authors after 1930.

Arnold Sommerfeld is the most central actor in the period 1919-1924 and second in the following one (1925-1930) and covers three consecutive periods. This is essentially explained by the centrality of his classic textbook: *Atomic structure and spectral lines*. With Peter Debye, he provides a bridge between the generation active in the 1910s and the one that emerges with quantum mechanics in the mid-1920s. The rise of Debye in the period 1931-1936 is here also due to an important textbook he published in 1929 on *Polar molecules*. These examples suggest that in many cases textbooks more than papers keep

⁴ In this paper, the content of research as not been identified using automatic analysis of titles; we use instead information on authors in the *Dictionary of Scientific Biographies*.

an author central in a network over many periods, as the half-life of books is much larger than that of papers.

Unsurprisingly Werner Heisenberg is among the most central actors for the period 1925-1944. It is also interesting to observe that some major figures like P.A.M. Dirac and E. Schrödinger have been central actors only during the heroic period of 1925-1930. The co-citation maps make visible the different specialties making up physics and how they change or even disappear with time. In Figure 1, J.J. Thomson, the master of the study of electron and ionization of gases is clearly at the center of action and closely related to P. Lenard, J. Stark and E. Rutherford, while a second center of interest is the network around M. Planck (Rubens, Paschen, Lummer W. Wien) on the study of black body radiation.

The networks in Figure 2 show that in addition to the works centered around Thomson and Planck, new actors like Einstein and W. Nernst appear. Einstein ranks 4th in the period 1912-1918 while Nernst is 6th. The disconnectedness of many co-citations pairs also suggest that many active topics co-exist without being strongly linked. When we jump to the 1925-1930 map, we observe a complete restructuration of the fields of interest where Einstein has become marginalized. As could be expected, strong links unite Heisenberg to M. Born, Dirac, A. Sommerfeld and other active contributors to quantum mechanics. Specialties like relativistic quantum mechanics, which link Dirac to O. Klein and W. Gordon and Schrödinger, are also visible (Figure 3). Finally, Figure 4, covering the period 1937-1944, shows a less dense and more diversified picture with a very strong center around Hans Bethe, a focal point of the first nuclear theory with new actors like E. Fermi and B. Rossi. Also visible is the beginning of solid-state physics around physicists like R. Peierls, F. Seitz, N. Mott and J.C. Slater. It is probable that this network will become stronger in the period 1945-1950 not yet analyzed.

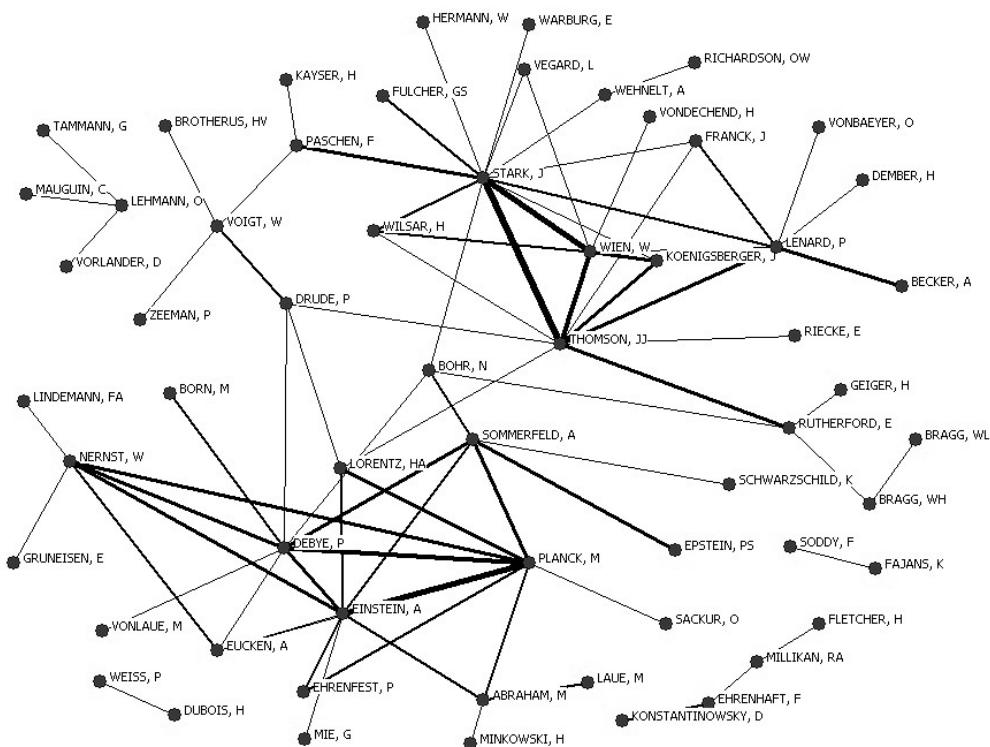


Figure 2. Co-citation network of physicists, 1912-1918 (More than 10 co-citations).

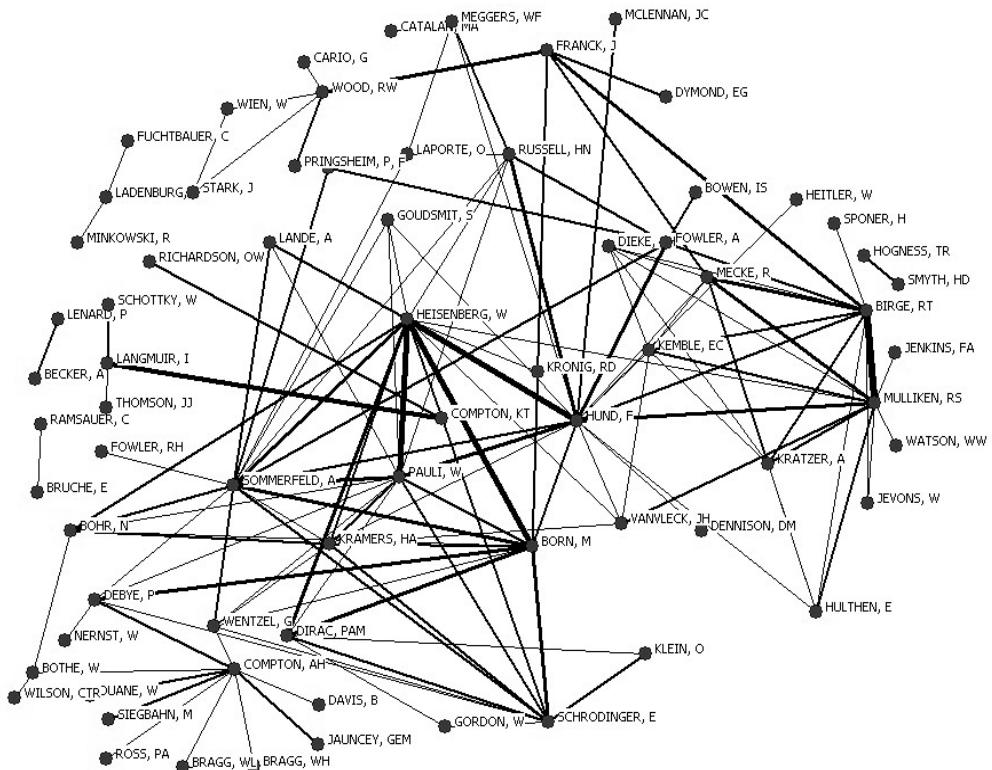


Figure 3. Co-citation network of physicists, 1925-1930 (More than 16 co-citations).

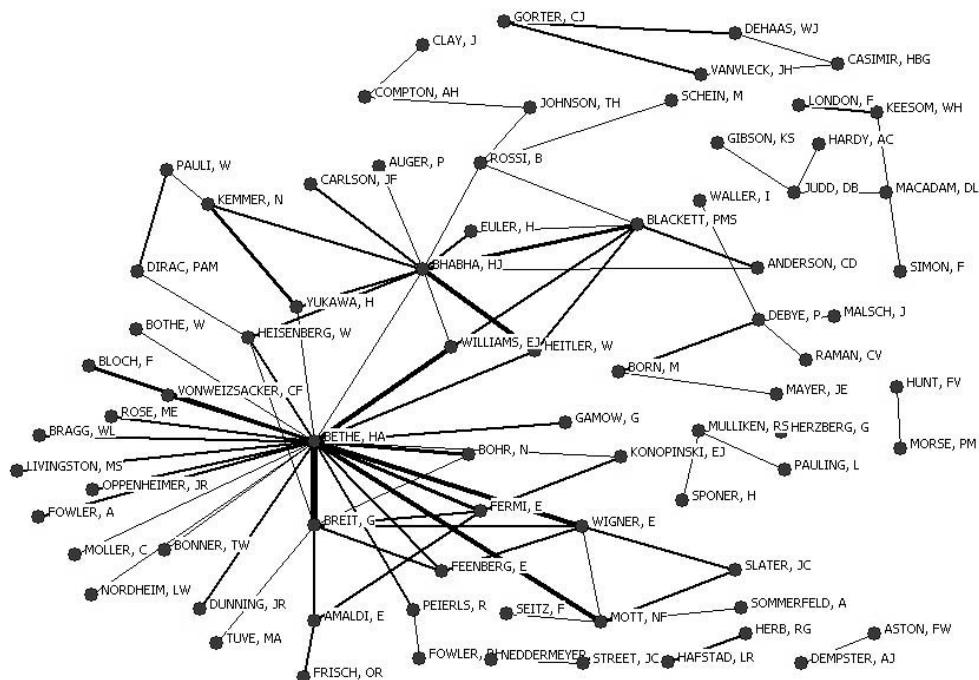


Figure 4. Co-citation network of physicists, 1937-1944 (More than 21 co-citations)

Conclusion

We have shown that calculating the centrality of actors in co-citation networks provides a useful indicator of their changing place in the scientific field. Comparing maps of co-citations and ranking of centrality make visible the dynamic of science. The value of the indicator, and thus the relative ranking, is affected by the choice of the lower limits for co-citation, but the major actors are still present when we calculate centrality using all links instead of only the stronger ones. Also, actors can have the same centrality by being weakly related to many different actors or by having strong links with only one or two actors. But despite these caveats, we have seen that using this simple measure of centrality gives us a representation of the evolution of physics that is coherent with the existing historical literature for the period 1900-1944. Comparison of the networks over time suggest that as the density of the field increases with time and more scientists are active, the field is more diverse and seems to contain more subgroups. Whereas Figure 1 contains few isolated groups, Figures 4 contains some unconnected clusters. The broadening of the networks over time is confirmed by the calculation of the average minimum distance between nodes (closeness) for the 50 most cited authors, which grows from 10.3 in 1900-1904 to 13.3 in 1937-1944 with a peak at 17.6 in 1925-1930. In future works, we will test other social network methods to see if they can be used to define specialties.

References

- Borgatti, S. P., Everett, M. G. & Freeman, L. C. (2002). Ucinet for Windows: Software for Social Network Analysis. Harvard: Analytic Technologies.
- Borgatti, S. P. (2002) NetDraw: Graph Visualization Software. Harvard: Analytic Technologies.
- Freeman, L. C. (1978/1979). Centrality in Social Networks. Conceptual Clarification. *Social Networks*, 1, 215-239.
- Garfield, E., Sher, I.H. & Torpie, R.J. (1964). The use of citation data in writing the history of science. Unpublished report.
- Garfield E. (2004). Historiographic mapping of knowledge literature, *Journal of information Science*, 30 (2), 119-145.
- Pais, A. (1982). ‘Subtle is the Lord...’ The Science and the Life of Albert Einstein. New York: Oxford University Press.
- Wasserman, S. & Faust, K. (1994). *Social Networks Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
- Gmür, M. (2003). M. Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, 57, 27-57.
- Small, H.G. (1977). Co-citation model of a scientific specialty – Longitudinal study of collagen research. *Social Studies of Science*, 7, 139-166.
- Small, H. (1978). Cited documents as concept symbols”, *Social Studies of Science*, 8, 327-340.
- Small, H. (1986). Recapturing Physics in the 1920s through citation analysis, *Czechoslovakian Journal of Physics B*. 36, 142-147.

A Comparative Analysis of Publication Activity and Citation impact Based on the Core Literature in Bioinformatics

Wolfgang Glänzel^{* ***}, Frizo Janssens^{***} and Bart Thijs^{*}

^{*}*Wolfgang.Glanzel@econ.kuleuven.be, Bart.Thijs@econ.kuleuven.be*

K.U. Leuven, Steunpunt O&O Indicatoren, Dekenstraat 2, B-3000 Leuven (Belgium)

^{**}*glaenzw@iif.hu*

Hungarian Academy of Sciences, ISPR, Nádor u. 18, H-1051 Budapest (Hungary)

^{***}*Frizo.Janssens@esat.kuleuven.be*

K.U. Leuven, Dept. of Electrical Engineering ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)

Abstract

A novel subject-delineation strategy has been developed for the retrieval of the core literature in bioinformatics. The strategy combines textual components with bibliometric, citation-based techniques. This bibliometrics-aided search strategy is applied to the 1980-2004 annual volumes of the Web of Science. Retrieved literature has undergone a structural as well as quantitative analysis. Patterns of national publication activity, citation impact and international collaboration are analysed for the 1990s and the new millennium.

Keywords

Bioinformatics; Subject delineation; mapping of science; citation analysis; scientific collaboration

Introduction

Bioinformatics is an interdisciplinary field that emerged from the increasing use of computer science and information technology for solving problems in biomedicine, mostly at the molecular level. Ouzounis and Valencia have provided a review of the early stages of the long history of the bioinformatics discipline (Ouzounis and Valencia, 2003). In recent studies by Patra and Mishra (2006) and Perez-Iratxeta et al. (2006), evolution and trends in bioinformatics research have been studied. The field has been characterised as an emerging discipline with astonishing growth dynamics. The studies were based on the MEDLINE database and partially on NIH-funded project grants. In both cases, bioinformatics was analysed in a broader biomedical context. In our present paper, we will strictly focus on the *core literature* in bioinformatics. An earlier structural analysis of the domain by Janssens et al. (2006) was strongly based on text mining and bibliometrics aided techniques, and aimed at improving classification of literature through the combination of linguistic and bibliometric tools. The aim of the present study, however, is to analyse the bibliometric core literature and its structure from the bibliometric point of view. We analyse retrieved bioinformatics literature along the following research tasks. First, we will have a look at growth dynamics of the field and the sub-discipline representation of the found clusters. In a second step, national publication activity and citation impact will be studied. Finally, patterns of international co-authorship and its citation impact are analysed. Unlike the above-mentioned studies by Patra and Mishra, (2006) and Perez-Iratxeta et al. (2006), the present paper is based on literature extracted from the Web of Science of *Thomson Scientific*. Only part of the computational linguistic analysis was conducted on MeSH terms taken from the MEDLINE database. Methodological background and data processing of this novel approach are summarised in the following sections.

Data sources and data processing

All bibliometric results are based on raw bibliographic data extracted from the 14-year annual volumes (1991-2004) of the Web of Science Edition of the *Science Citation Index Expanded™* (SCIE) of *Thomson Scientific* (Philadelphia, PA, USA). Publication data have been matched with *MEDLINE* which has been used as auxiliary data source for the determination of search terms. Only papers recorded as *article*, *note* or *review* in the SCIE were taken into consideration. Papers recorded as *Letter to the Editor* were excluded since this document type tends to cause biases in the application of bibliographic coupling and co-citation analyses (see Glänzel and Czerwon, 1996). The papers were

assigned to countries based on the corporate address given in the by-line of the publication. All countries and institutions indicated in the address field were thus taken into account. Co-authorship was counted for the corresponding address pairs (countries and institutions) if the names of the concerning entities occurred simultaneously. It has to be stressed here that publication counts and citation frequencies cannot be summed up over co-authorship links to the total. For the meso study, addresses were cleaned-up, unified and accordingly de-duplicated at the level of main institution.

Citation counts have been determined on basis of an item-by-item procedure using special identification keys made up of bibliographic components of the author and source fields. Citations were counted in a three-year period: in the year of publication and the two subsequent years, that is, for instance, if papers published in 1999 were considered, all citations received by them in the period 1999-2001 have been counted. The choice of the citation window is in line with recent practice in the field of scientometrics. Because of the use of 3-year citation windows, citations could be counted for papers published up to 2003 (citations received in 2003-2005).

The delineation of the research field bioinformatics

An earlier bibliometric study by Patra and Mishra (2006) was based on MEDLINE and the use of MeSH terms resulted in a rather broad coverage of the field. In the present study we apply a much stricter strategy resulting in a *core set* of bioinformatics literature. This strategy with strong bibliometric component is based on bibliographic coupling (“horizontally” searching at the same time level) as well as on references and citations (“vertically” searching in the past and future, respectively), a further data source has been used, namely the subject headings annotated to *MEDLINE* records that were matched with the ISI dataset. The MeSH terms are also used in part for validation and to refine the retrieval made in the SCIE database. This complex strategy applied in (Janssens, 2006) consists logically of two parts which, in turn, have several components each. The first part comprises two *unconditional criteria* (*UC1* and *UC2*), which include core journals covered by the Web of Science (*UC1*) and the MEDLINE database (*UC2*), respectively.

UC1: Journal in WoS = BIOINFORMATICS (formerly COMPUTER APPLICATIONS IN THE BIOSCIENCES), JOURNAL OF COMPUTATIONAL BIOLOGY, BRIEFINGS IN BIOINFORMATICS, BMC BIOINFORMATICS.

UC2: Journal in MEDLINE = IN SILICO BIOLOGY, PSB ON-LINE PROCEEDINGS, APPLIED BIOINFORMATICS, PLOS COMPUTATIONAL BIOLOGY

UC3: Keywords in title = BIOINFORMATICS, COMPUTATIONAL BIOLOG*, SYSTEMS BIOLOGY

In other terms, all papers meeting at least one of the criteria UC1 – UC3 are considered relevant. This set has been extended by two *conditional criteria* (*CC1* and *CC2*), each of which results in related but not necessarily in core literature. In particular, the conditional criteria comprise conditions for reference (*CC2*) and citation links (*CC3*).

CC1: Records cited by UC1

CC2: Records citing UC1

All papers meeting at least one of the criteria CC1 and CC2 are considered potentially relevant, but might not directly be concerned with bioinformatics. Only that part of literature, which meets further restrictive criteria, will be considered truly relevant. In order to reduce or even exclude noise, thresholds T_i for the strength of citation and reference links were used for fine-tuning. The bibliometrics aided retrieval strategy (BR) for identifying relevant papers in bioinformatics is thus obtained by the following formula.

$$BR_{bioinf} = (UC1 \vee UC2 \vee UC3) \vee ((CC1 \wedge T_i) \vee (CC2 \wedge T_i)).$$

In particular, we used four different thresholds T_i based on the absolute number i of citations and references, respectively. Table 1 presents the effect of adjusting the strength of citation/reference links for $i = 1, 2, \dots, 4$ on the number of retrieved documents. In addition, the results of the first

unconditional criterion as well as the ‘OR’ combination of UC1 with the third unconditional criterion is shown. Since T_1 and T_2 still resulted in perceptible noise, we decided to use T_3 for the study.

The retrieval has first been made for the period 1981-2004; all papers indexed for the sub-period 1991-2004 have then been selected for the bibliometric analysis. All retrieval related statistics are calculated for the full 26-year time span. The publication output in the field in 1980-1990 is, however, minute and from the statistical viewpoint not decisively.

Records retrieved from WoS were matched against MEDLINE in order to obtain the Medical Subject Headings (MeSH). Matching was based on an item-by-item procedure using special identification-keys made up of bibliographic components such as publication year, volume, first page, first characters of author names and substrings of the title.

Table 1. Number of records retrieved for different combinations of criteria.

Strategy	Threshold	Documents retrieved
UC1	---	3,386
UC1 \square UC3	---	9,620
BR	T_1	41,995
	T_2	13,239
	T_3	7,655
	T_4	5,470

Table 2 shows the 20 most frequent MeSH terms where we have excluded those terms acknowledging research support.

Table 2. The most frequent 20 MeSH terms (excluding terms acknowledging funding).

Rank	MeSH term	Rank	MeSH term
1.	Algorithms	11.	Sequence Alignment/methods
2.	Software	12.	Proteins/chemistry
3.	Humans	13.	Base Sequence
4.	Sequence Alignment	14.	Sequence Analysis, DNA
5.	Comparative Study	15.	Gene Expression Profiling
6.	Animals Proteins	16.	Models, Genetic
7.	Molecular Sequence Data	17.	Internet
8.	Computational Biology	18.	Computer Simulation
9.	Amino Acid Sequence	19.	Oligonucleotide Array Sequence Analysis
10.	Databases, Factual	20.	Information Storage and Retrieval

The TF-IDF weight (*term frequency-inverse document frequency*) was used as a relative measure to evaluate how important a term is to a document relative to the collection. In particular, the relative frequency of the term in a document is gauged against the frequency of the term in the collection. The *term frequency* is often defined as the relative frequency of a word in a document, that is, $tf = n_j / \sum_j n_j$. The *inverse document frequency*, in turn, is a measure of the general importance of the term. It is defined as $idf = -\log(d_j/d)$, that is, the negative logarithm of the share of documents where the term T_i appears in. Finally, the TF-IDF weight is defined as the product of the two previous measures ($tfidf = tf \cdot idf$). The 20 best TF-IDF terms in titles and abstracts are presented in Table 3.

Journal coverage of bioinformatics literature in the SCIE database

In total 7401 articles, notes or reviews in bioinformatics could be retrieved for the period 1981-2004. Patra and Mishra (2006) selected 14563 journal articles, that is, about twice as many as we have found.

The main reason is the broad interpretation of bioinformatics resulting in a somewhat more liberal search strategy. The other reason is the broader coverage of the underlying database. As mentioned above, we aimed at a very strict interpretation of the field, at retrieving the very core of bioinformatics with practically no noise. This was essential for having a solid groundwork for the cluster analysis of the retrieved literature. Nonetheless, the list of most relevant journals in bioinformatics of our exercise by and large coincides with that by Patra and Mishra. Core journals, of course, can be found at the top of the list (see Table 4).

Table 3. The 20 best TF-IDF terms in titles and abstracts.

Rank	TF-IDF term	Rank	TF-IDF term
1	protein	11	function
2	sequenc*	12	cluster
3	align	13	interdisciplinar*
4	gene	14	applic*
5	structur*	15	program
6	predict*	16	set
7	databas*	17	base
8	genom	18	domain
9	algorithm	19	interact*
10	model	20	famili*

Table 4. The 25 most frequently used journals for publishing bioinformatics literature.

Rank	Journal	Frequency
1.	BIOINFORMATICS	1900
2.	COMPUTER APPLICATIONS IN THE BIOSCIENCES	724
3.	NUCLEIC ACIDS RESEARCH	594
4.	JOURNAL OF COMPUTATIONAL BIOLOGY	397
5.	JOURNAL OF MOLECULAR BIOLOGY	241
6.	BMC BIOINFORMATICS	239
7.	GENOME RESEARCH	203
8.	PNAS USA	189
9.	NATURE	116
10.	MOLECULAR BIOLOGY AND EVOLUTION	107
11.	SCIENCE	107
12.	PROTEIN SCIENCE	92
13.	PROTEINS-STRUCTURE FUNCTION AND GENETICS	88
14.	PROTEIN ENGINEERING	84
15.	MOLECULAR PHYLOGENETICS AND EVOLUTION	63
16.	NATURE GENETICS	56
17.	JOURNAL OF MOLECULAR EVOLUTION	54
18.	CURRENT OPINION IN STRUCTURAL BIOLOGY	51
19.	GENOMICS	46
20.	PROTEINS-STRUCTURE FUNCTION AND BIOINFORMATICS	44
21.	FEBS LETTERS	37
22.	GENOME BIOLOGY	37
23.	TRENDS IN BIOCHEMICAL SCIENCES	33
24.	GENETICS	30
25.	TRENDS IN GENETICS	30

Journals in computational and molecular biology as well as the important multidisciplinary journals *PNAS USA*, *Nature* and *Science* are the most important publication channels for bioinformatics research. The huge number of journals in which the papers were scattered in the paper by Patra and Mishra could be confirmed by us as well.

Table 5. The nine bioinformatics clusters obtained from the hybrid hierarchical cluster algorithm.

Cluster	Name	Papers	Best author keyword	Best stem or phrase
1	RNA structure prediction	205	rna secondary structure	RNA
2	Protein structure prediction	1167	protein structure prediction	protein
3	Systems biology & molecular networks	694	bioinformatics	network
4	Phylogeny & evolution	749	phylogeny	phylogenetic
5	Genome sequencing & assembly	640	sequencing hybridization	base sequencing
6	Gene/promoter/motif prediction	995	gene regulation	gene
7	Molecular DBs & annotation platforms	1091	genome analysis	databases
8	Multiple Sequence alignment	713	sequence alignment	align
9	Microarray analysis	1147	microarray	microarray
<i>Total</i>	Bioinformatics	7401	bioinformatics	protein

In order to depict the cognitive structure of the field represented by its core literature, the agglomerative hierarchical, hard cluster algorithm using Ward's method (cf., Jain and Dubes, 1988; Berkhin, 2002; Kaufman and Rousseeuw, 1990) was used. In total, we obtained nine clusters, the algorithm, which is based on the integration of both textual information and citation links, is described by Janssens et al. (2005, 2006). The cognitive structure of the field as reflected by term networks using the best 10 terms from titles and abstracts according to mean TF-IDF scores for each of nine clusters is shown in Figure 1. In addition, Table 5 presents the clusters, their size and their characterisation by best author keywords and best terms from titles and abstracts.

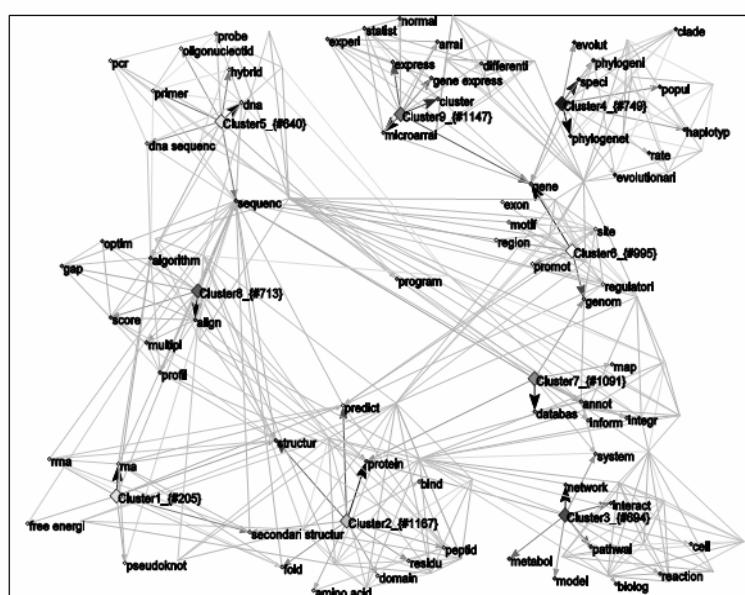


Figure 1. Visualisation of the cognitive structure of bioinformatics based on term networks using Pajek (Batageli, 2002).

Evolution of publication output and citation impact in bioinformatics in the period 1991-2004

Evolution of publication output and citation impact of the field

Figure 2 visualises the evolution of the cumulative number of papers in bioinformatics. The growth of publications lies in between the linear model in the first half of the period and the exponential model for the second half (similarly as observed in nanoscience and -technology, Glänzel et al. 2003). Literature growth clearly characterises the field as a young, emerging and dynamically evolving discipline.

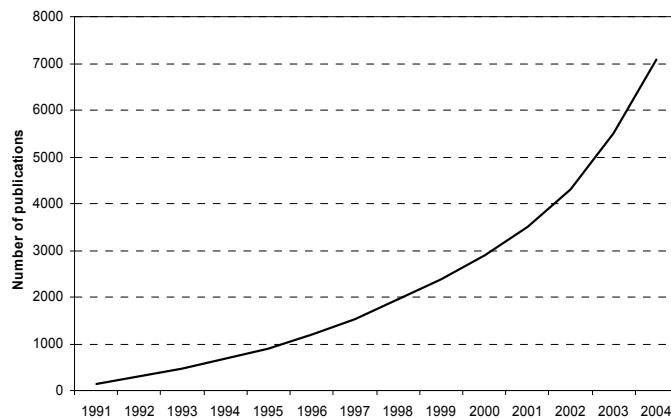


Figure 2. Evolution of cumulated publication output in the Period 1991-2004 (world total).

The dynamic growth of literature in bioinformatics is outrun by an even more powerful increase of citations. The patterns are shown in Figure 3. Citations have, as already mentioned in the outset, been determined on a basis of three-year citation windows. Before we have a closer look at citation patterns, we will introduce the indicators used for the analysis.

The *Mean Observed Citation Rate* (MOCR) is defined as the ratio of citation count to publication count. It reflects the factual citation impact of any unit like a country, region, institution, research group etc. Since the underlying paper set is restricted to a single, however cross-disciplinary subject, we can use the subject-standardised Mean Observed Citation Rate ($MOCR|_s$) which is simply the ratio of the unit's MOCR value and the world standard of the field. In addition, we use the share of author self-citations and the citation impact of internationally co-authored papers.

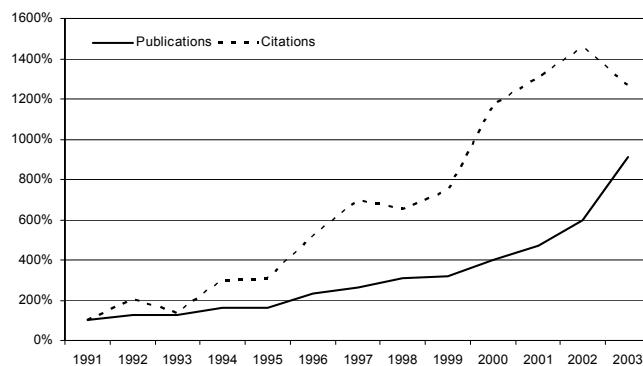


Figure 3. Annual change of citations compared with that of publications in bioinformatics for 1991-2003 (1991 = 100%).

The evolution of the field's mean observed citation impact is presented in Figure 4. The strong linear increase of citation impact in the 1990s is followed by a sharp decline in the new millennium. The reasons for this phenomenon are not clear. However, a decline of citation impact has been observed in the case of nanoscience and -technology (Glänzel et al., 2003). It seems that emerging fields are

characterised first by a growth of citation impact exceeding that of the publication output, then by stagnation and later on by the decrease of impact while the powerful increase of publication activity continues.

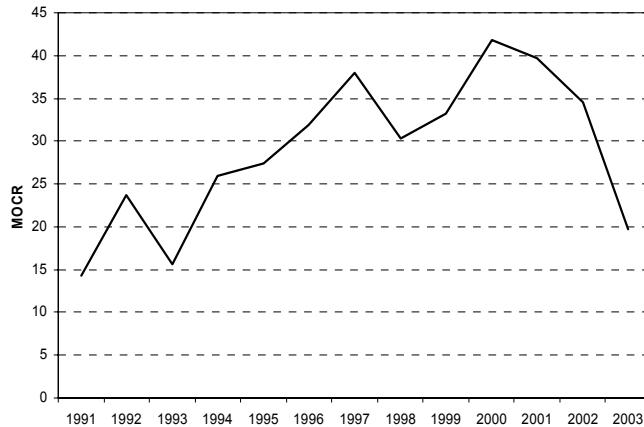


Figure 4. Evolution of mean observed citation impact in the period 1991-2003 (world total).

In order to gain detailed information about the evolution of citation means, we analyse the distributions of citations over individual papers, one each for the beginning and the end of the period under study. The diagram is presented in Figure 5.

Although the citation impact decreases from 2001, the MOCR values for the second sub-period are still distinctly higher than the corresponding values for the first one. The distributions for 1991-1995 and 2000-2003 are quite similar except for the shares of poorly and frequently cited papers. The distribution has evolved into a slightly less skewed one. The moments of the two distributions are high: The mean in 1991-1995 amounts to 22.2, that in 2000-2003 to 31.1. The share of less cited and uncited papers decreased while that of frequently cited papers increased. In verbal terms, the frequency distributions of citations over publications characterise the field as a specialty with high citation impact the citation patterns of which, however, are quite polarised although in the second sub-period less distinctly.

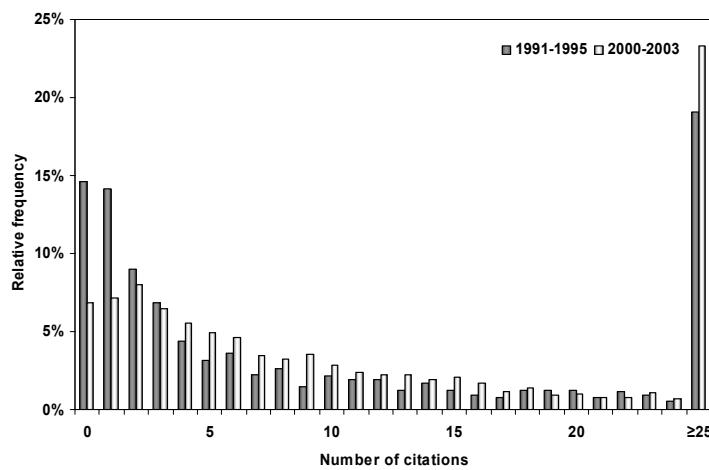


Figure 5. Distribution of citations over papers published in 1991-1995 and 2000-2003.

Table 6 presents the cluster size and citation impact of the nine clusters for the period 1991-2003. Citations have been collected for 3-year time windows each (beginning with the publication year). Cluster #1 labelled “RNA structure prediction” with 152 publications is the smallest one; all other clusters have more than 400 and less than 1000 papers. The citation impact of most clusters lies around that of the field of about 30 citations per paper. The impact of Clusters #5 (Genome sequencing

& assembly) and #8 (Multiple Sequence alignment) lies far below, that of Cluster #6 (Gene/promoter/motif prediction) distinctly above the field standard.

Table 6. Publication output and citation impact of the nine bioinformatics clusters for 1991-2003.

Cluster	Name	Papers	MOCR
1	RNA structure prediction	152	24.22
2	Protein structure prediction	923	27.05
3	Systems biology & molecular networks	476	26.89
4	Phylogeny & evolution	480	29.92
5	Genome sequencing & assembly	432	11.78
6	Gene/promoter/motif prediction	781	47.24
7	Molecular DBs & annotation platforms	907	33.18
8	Multiple Sequence alignment	558	16.60
9	Microarray analysis	798	37.06
<i>Total</i>	Bioinformatics	5507	30.27

Publication output and citation impact of the 30 most active countries

For the analysis of national publication activity and citation impact, the 30 most active countries in the period 1991-2004 have been selected. Countries with less than 30 papers in the 14-year period have not been selected by reasons of statistical reliability. The publication output of the 30 most active countries in bioinformatics and their share in the world total in this field are presented in Table 7. In order to provide information about the evolution of national publication activity in the field, the period 1991-2004 has been split into two sub-periods, particularly, 1991-1997 and 1998-2004. National data in Table 7 are ranked in descending order by their publication output in the whole 14-year period. If we compare the list with similar lists on national publication output in all fields combined, we can conclude that those countries that are most active in scientific research in all fields combined have top activity in bioinformatics research, too.

However, the three “leading” countries, USA, UK and Germany rank distinctly higher in bioinformatics than in all fields combined (cf., Glänzel et al., 2002). Although publication counts of most countries for the first period are small, we can observe the same powerful growth of publication activity of China and other emerging scientific nations like South Korea, Taiwan and Brazil (see Glänzel et al., 2007). National representation also confirms the findings by Patra and Mishra (2006).

The citation impact of the 30 most active countries with at least 25 papers in 1991-2003 in the two sub-periods 1991-1997 and 1998-2003 is shown in Table 8. The overall high impact is partially a consequence of the citation-based component of the retrieval strategy. A study of bibliographic coupling by Glänzel and Czerwon (1996) has shown that retrieval based on strong coupling links results on higher-than-average citation impact. Citation aided tools in information retrieval and data mining necessarily imply a certain bias concerning visibility of the literature. The better depiction of the structure of the information space is to the detriment of loosely linked and less visible documents. The high relative citation impact of Canada, Switzerland, Australia and the Netherlands (more than twice the world standard) is worth mentioning. This is contrasted by the relatively low impact of Russia and Italy in all sub-periods although their publication activity is quite high. The share of author self-citations f_S of about 10% is low in this field; national deviation from this standard follows the patterns observed from other science fields (Glänzel et al., 2003). In general, self-citation rates are rather low, even for a biomedicine related field.

Table 7. Publication output of the 30 most active countries in sub-periods 1991-1997 and 1998-2004.

Country	1991-1997		1998-2004		1991-2004	
	Papers	Share	Papers	Share	Papers	Share
USA	721	46.8%	2923	52.8%	3644	51.5%
GBR	235	15.3%	767	13.9%	1002	14.2%
DEU	189	12.3%	594	10.7%	783	11.1%
FRA	121	7.9%	331	6.0%	452	6.4%
JPN	74	4.8%	232	4.2%	306	4.3%
CAN	49	3.2%	223	4.0%	272	3.8%
ITA	60	3.9%	150	2.7%	210	3.0%
ESP	39	2.5%	146	2.6%	185	2.6%
ISR	33	2.1%	144	2.6%	177	2.5%
SWE	14	0.9%	161	2.9%	175	2.5%
RUS	56	3.6%	118	2.1%	174	2.5%
AUS	21	1.4%	134	2.4%	155	2.2%
CHE	47	3.1%	100	1.8%	147	2.1%
CHN	7	0.5%	139	2.5%	146	2.1%
BEL	24	1.6%	108	2.0%	132	1.9%
DNK	12	0.8%	83	1.5%	95	1.3%
NLD	18	1.2%	77	1.4%	95	1.3%
IND	16	1.0%	72	1.3%	88	1.2%
SGP	6	0.4%	73	1.3%	79	1.1%
POL	5	0.3%	53	1.0%	58	0.8%
NOR	6	0.4%	45	0.8%	51	0.7%
IRE	7	0.5%	43	0.8%	50	0.7%
TWN	1	0.1%	47	0.8%	48	0.7%
AUT	5	0.3%	42	0.8%	47	0.7%
FIN	5	0.3%	41	0.7%	46	0.7%
KOR	1	0.1%	44	0.8%	45	0.6%
BRA	0	0.0%	44	0.8%	44	0.6%
NZL	6	0.4%	36	0.7%	42	0.6%
HUN	11	0.7%	27	0.5%	38	0.5%
GRC	5	0.3%	30	0.5%	35	0.5%
WORLD	1540	100.0%	5536	100.0%	7076	100.0%

Table 8. Citation impact and self-citation rate of the 30 most active countries in 1991-2003 in the two sub-periods 1991-1997 and 1998-2003.

Country	1991-1997			1998-2003			1991-2003		
	Papers	MOCR _f	f _s	Papers	MOCR _f	f _s	Papers	MOCR _f	f _s
USA	721	1.28	10.1%	2162	1.37	9.1%	2883	1.35	9.3%
GBR	235	1.17	12.0%	594	1.47	11.0%	829	1.39	11.2%
DEU	189	1.24	13.8%	429	1.48	11.2%	618	1.41	11.8%
FRA	121	2.09	12.5%	247	1.66	9.6%	368	1.78	10.6%
JPN	74	1.01	17.3%	157	1.97	10.9%	231	1.68	12.0%
CAN	49	2.96	11.1%	140	2.15	10.1%	189	2.34	10.4%
ITA	60	0.90	19.4%	103	0.73	19.9%	163	0.78	19.7%
RUS	56	0.16	26.7%	94	0.52	17.6%	150	0.39	18.9%
ISR	33	0.40	21.5%	112	2.06	9.1%	145	1.73	9.7%
ESP	39	1.20	17.5%	99	2.18	10.3%	138	1.93	11.5%
SWE	14	—	—	105	1.63	9.5%	119	1.85	8.8%
CHE	47	2.24	12.0%	68	3.04	8.6%	115	2.69	9.7%
AUS	21	—	—	90	2.49	8.9%	111	2.13	9.2%
BEL	24	—	—	71	0.88	14.2%	95	1.14	17.5%
CHN	7	—	—	79	1.96	8.9%	86	1.90	9.2%
DNK	12	—	—	61	1.64	8.0%	73	1.78	8.5%
NLD	18	—	—	47	2.56	8.5%	65	2.42	10.5%
IND	16	—	—	42	0.29	19.4%	58	0.23	20.1%
SGP	6	—	—	42	0.59	22.9%	48	0.54	23.2%
NOR	6	—	—	35	1.65	10.4%	41	1.50	10.9%
POL	5	—	—	34	0.52	27.3%	39	0.69	25.2%
IRE	7	—	—	30	4.31	7.2%	37	4.40	9.1%
FIN	5	—	—	31	0.56	13.3%	36	0.55	13.9%
HUN	11	—	—	22	—	—	33	0.42	16.8%
NZL	6	—	—	24	—	—	30	0.83	13.3%
AUT	5	—	—	24	—	—	29	0.60	17.6%
BRA	0	—	—	27	0.26	32.9%	27	0.26	32.9%
GRC	5	—	—	21	—	—	26	2.33	10.6%
TWN	1	—	—	25	0.21	26.7%	26	0.21	28.0%
KOR	1	—	—	20	—	—	21	—	—
WORLD	1540	1.00	11.3%	3967	1.00	10.2%	5507	1.00	10.5%

Cluster representation of the five most active countries

The breakdown of national publication output by clusters does not allow any reliable quantitative analysis for most of the 30 selected countries because of the often too small publication sets. We restrict the analysis to the five leading countries, particularly, the USA, UK, Germany, France and Japan. Their share in the nine individual clusters is presented by Figure 5.

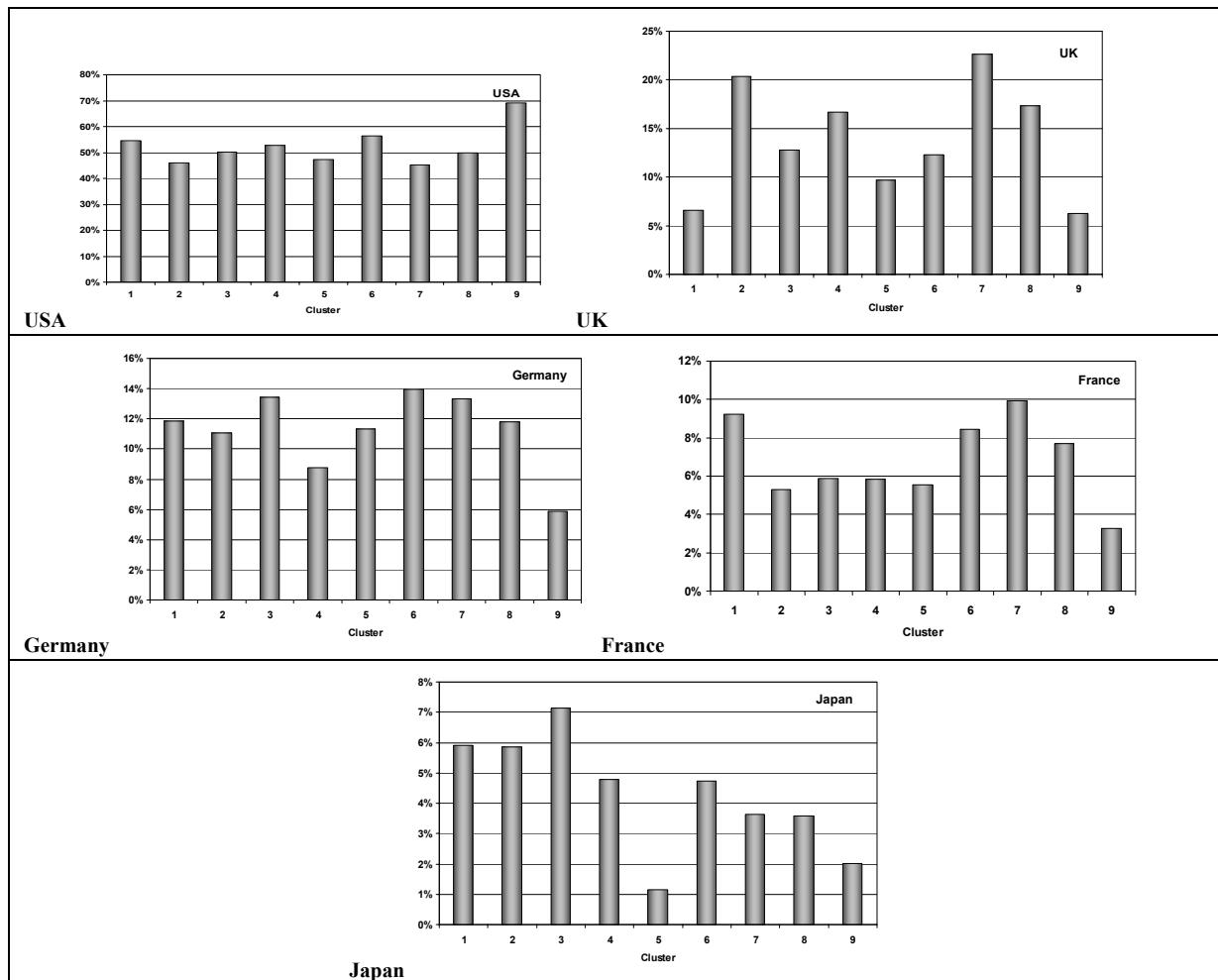


Figure 5. National representation of the five most active countries by clusters.

The US with share of about 50% and more are predominant in most sub-disciplines. Above all, Cluster #9 (Microarray analysis) is dominated by the USA with 70% of all papers. Also Germany has a well-balanced high share in all clusters, except for Cluster #9. The other three countries reflect a rather heterogeneous picture; the UK contribution to Cluster #2 (Protein structure prediction) and #6 (Gene/promoter/motif prediction) is worth mentioning, however, the contribution to Cluster #1 (RNA structure prediction) and #9 (Microarray analysis) is rather small. The situation in France is similar: the strong contribution to Cluster #1 and #7 (Molecular DBs & annotation platforms) is contrasted by a low share in Cluster #9. The extremes in the Japanese publication output can be found in Cluster #3 (Systems biology & molecular networks) with 7% of the world total and Cluster #5 (Genome sequencing & assembly) with 1%. The results of further analysis of cluster dynamics and structural changes will be the subject of a forthcoming study.

International co-authorship patterns in bioinformatics

The global collaboration network of research in bioinformatics

Beyond individual interests and motivation of individual scientists, teamwork and scientific collaboration is one of the characteristics of “big science” (Price, 1966). Of course, in inter- and cross-disciplinary areas, where scientists from different fields are jointly doing research, intensive collaboration is expected (see Glänzel et al., 2003).

It is clear that a variety of different purposes and motivations, the manifold of factors influencing (international) collaboration, must have at least in part a measurable impact on the published results of joint research work. National characteristics in international scientific co-authorship patterns have been studied by *Glänzel* (2001). The results often confirmed but sometimes contradicted widespread notions on the efficiency of international collaboration. Furthermore, an interesting observation has been made concerning the re-integration of EIT countries into the scientific collaboration structures of Europe and the Western world.

The absolute number of international papers and their share in the total national publication output serve as basic indicators of international co-authorship relations and scientific collaboration. International collaboration depends on the country's 'size' (cf., for instance, *Schubert* and *Braun*, 1990 and *Katz*, 2000). At the national level, the share of international collaboration in large countries is necessarily lower than that of medium-sized or even small countries. The share of all international papers in the world can, in principle, be determined as the complementary share of the ratio of all countries' domestic papers and the total world publication output. Such 'world average' is, however, not an appropriate reference standard for international collaboration activity (cf. *Schubert* and *Braun*, 1990), and is therefore not used here.

Table 9. Share and citation impact of international co-publications of 19 selected countries in 1991-2003.

Country	Co-publ.	Share	MOCR	MOCR _i
HUN	25	75.8%	12.64	11.92
NLD	46	70.8%	73.38	97.02
DNK	49	67.1%	53.90	73.20
ISR	85	58.6%	52.41	80.12
SGP	28	58.3%	16.38	24.11
SWE	68	57.1%	55.93	89.96
CHN	49	57.0%	57.38	95.55
CHE	65	56.5%	81.57	113.18
CAN	97	51.3%	70.91	116.03
RUS	76	50.7%	11.78	20.42
AUS	55	49.5%	64.44	120.73
ESP	68	49.3%	58.41	106.53
BEL	43	45.3%	34.37	56.35
ITA	70	42.9%	23.63	49.37
DEU	265	42.9%	42.63	73.43
GBR	324	39.1%	42.05	78.02
FRA	139	37.8%	53.77	120.79
JPN	73	31.6%	50.98	131.66
USA	707	24.5%	40.91	61.53

Table 9 presents number and share of internationally co-authored publications of those of the 19 most active countries that have at least 25 international papers each in the period 1991-2004. Countries have

been ranked by the share of international co-publications in the total national publication output. In addition, both the national MOCR values and the corresponding indicator for international co-publications ($MOCR_i$) is presented.

Hungary, the Netherlands and Denmark have the highest share of international co-publications. More than two thirds of their papers have been published in international collaboration. Among the countries with high share of international co-publications, we also find Israel, Sweden, Singapore, China, Switzerland, Canada and Russia with more than 50% international papers. Even US scientists publish one quarter of their papers jointly with colleagues abroad. In all, the shares of international co-publications are roughly in line with those found in other interdisciplinary fields like, for instance, nanoscience and -technology (cf., Glänsel et al., 2003).

The citation indicators are even more impressive. The figures confirm that international collaboration in general results in higher visibility and impact, but as mentioned above, there are also exceptions to the rule. The already very high citation scores are outrun by the reception of the international papers in our set. Almost incredible values are reached by Switzerland, Canada, Australia, Spain, France and Japan. On the other hand, collaboration seems not to pay off for Hungary; despite the huge share of collaboration, domestic papers have a better reception here.

Mapping mutual co-authorship links

In order to measure the strength of mutual collaboration links, an appropriate similarity measure based on country pairs is used. Multinational collaboration is therefore split up to a group of bilateral relations.

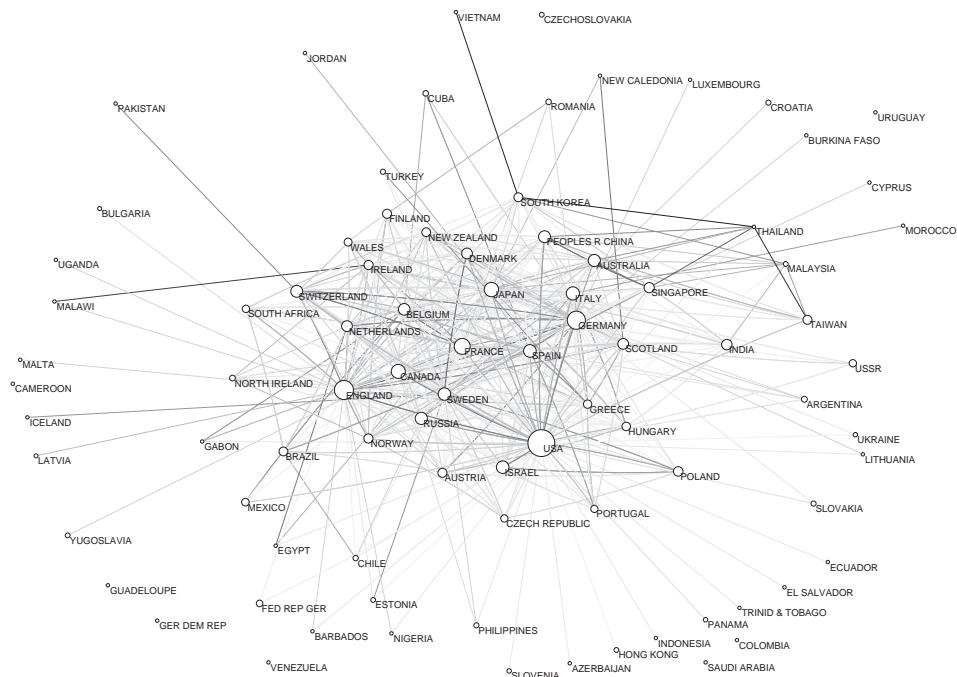


Figure 6. International collaboration network based on Salton's cosine measure with Kamada-Kawai layout (Batagelj, 2002).

A frequently used measure for the strength of co-publication links is the cosine measure according to Salton. It is defined as the number of joint publications divided by the square root of the product of the number (i.e., the geometric mean) of total publication outputs of the two countries, that is,

$$r = \frac{p_{ij}}{\sqrt{p_i \cdot p_j}},$$

where p_{ij} is the number of links between the countries i and j , and p_i (p_j) the total number of publications of the country i (j). As a consequence of this practice one has to distinguish between the number of *co-publications* and of *co-authorship links*. The results are presented in Figure 6. The “big” countries, USA, UK, Germany, France and Japan can be found in the very centre of the diagram. These countries are the real nodes of this global network. Since the figure is based on all bioinformatics papers retrieved for 1980-2004, countries like Czechoslovakia, GDR and FRG still appear in the diagram; because of the dynamical growth of the field, their role in the complete set is marginal. The appearance of the emerging nations like China, Singapore, Korea and Brazil as nodes in the collaboration network is again worth mentioning.

Conclusions

The field of bioinformatics proved a young, emerging field characterised by a powerful, from the late 1990s on, by an almost exponential growth of literature. Beyond several core journals, important periodicals in molecular biology as well as the multidisciplinary journals *Science*, *Nature* and *PNAS USA* proved to be the most important publication channels. Although we focussed on the bioinformatics core literature, our study has confirmed findings by other recent studies concerning publication patterns.

The structural analysis resulted in the identification of nine sub-disciplines with individual national profiles. The partially citation-based subject delineation supported the identification of rather visible publications; the citation analysis characterised bioinformatics as a field with very high overall citation scores. According to our expectations, the extent of international collaboration is in keeping with that of other emerging interdisciplinary fields. The “big” countries form the nodes of the global co-publication network. International collaboration resulted in general to a powerful increase of the otherwise already high citation impact.

References

- Batagelj, V. & Mrvar, A. (2002). Pajek - Analysis and visualization of large networks. *Graph Drawing*, 2265, 477-478.
- Berkhin, P. (2002). *Survey of clustering data mining techniques*. Technical report (Accrue Software). Retrieved November 15, 2006 from: <http://citeseer.ist.psu.edu/berkhin02survey.html>.
- Glänzel, W. & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level, *Scientometrics* 37, 195-221.
- Glänzel, W. (2001). National Characteristics in International Scientific Co-authorship, *Scientometrics*, 51, 69–115.
- Glänzel, W., Schubert, A. & Braun, T. (2002). A relational charting approach to the world of basic research in twelve science fields at the end of the second millennium. *Scientometrics*, 55 (3), 335-348.
- Glänzel, W., Meyer, M., Du Plessis, M., Thijs, B., Magerman, T., Schlemmer, B., Debackere, K. & Veugelers, R. (2003). *Nanotechnology - Analysis of an Emerging Domain of Scientific and Technological Endeavour*. Retrieved November 15, 2006 from: http://www.steunpuntoos.be/nanotech_domain_study.pdf
- Glänzel, W., Thijs, B. & Schlemmer, B. (2004). A bibliometric approach to the role of author self-citations in scientific communication. *Scientometrics*, 59, 63-77.
- Glänzel, W., Debackere, K. & Meyer, M. (2007). ‘Triad’ or ‘Tetrad’? On global changes in a dynamic world. *Scientometrics*, to be published.
- Janssens, F., Glenisson, P., Glänzel, W. & De Moor, B. (2005). Co-clustering approaches to integrate lexical and bibliographical information. In P. Ingwersen and B. Larsen (Eds.), *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics* (ISSI) (pp. 284-289). Stockholm, Sweden.
- Janssens, F., Tran Quoc, V., Glänzel, W. & De Moor, B. (2006). Integration of textual content and link information for accurate clustering of science fields. In V. P. Guerrero-Bote (Ed.), *Proceedings of the I International Conference on Multidisciplinary Information Sciences and Technologies* (InSciT2006) (pp. 615-619). Mérida, Spain.
- Jain, A. & Dubes, R. (1988). *Algorithms for clustering data*. New Jersey: Prentice Hall.
- Katz, J. S. (2000). Scale-independent indicators and research evaluation, *Science and Public Policy*, 27 (1), 23-36.
- Kaufman, L. & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley and Sons Inc.
- Ouzounis, C. A. & Valencia, A. (2003). Early bioinformatics: the birth of a discipline - a personal view. *Bioinformatics*, 19, 2176-2190.

- Patra, S. K. & Mishram S. (2006). Bibliometric study of bioinformatics literature. *Scientometrics*, 67, 477–489.
- Perez-Iratxeta, C., Andrade-Navarro, M. A. & Wren, J. D. (2006). Evolving research trends in Bioinformatics. *Briefings in Bioinformatics*, in press.
- Price, D. deSolla (1966). *Little Science, Big Science*. New York: Columbia Univ. Press.
- Schubert, A. & Braun, T. (1990). World Flash on Basic Research: International Collaboration in the Sciences, 1981–1985, *Scientometrics*, 19, 3–10.

Structure and Research Performance of Spanish Universities¹

Isabel Gómez, María Bordons, M.Teresa Fernández, Fernanda Morillo

igomez@cindoc.csic.es, mbordons@cindoc.csic.es, mtf@cindoc.csic.es, fmorillo@cindoc.csic.es
Centro de Información y Documentación Científica (CINDOC), Spanish Research Council (CSIC),
Joaquín Costa 22, 28002 Madrid (Spain)

Abstract

The aim of this paper is to describe Spanish universities by means of structural, input and output indicators, to explore the relationship between those indicators and to analyse university behaviour in different dimensions. Age of the universities and environmental conditions are taken into account, together with input and output indicators, as well as others related to the networks and links established. Our results will contribute to the knowledge of the research university system in Spain, producing data that could be useful for research management both at the institutional, regional and national level.

Keywords

research performance; universities; spain; bibliometric indicators

Introduction

Universities play a decisive role in the current advancement of science in most countries. Together with their twofold traditional vocation of teaching and research, a third mission oriented to solve societal needs is increasingly demanded at present. In fact, universities contribute to the production of new knowledge, its transmission, its dissemination and its use in technical innovation. University research can improve the competitiveness of local industry, but also contribute to social cohesion and to the promotion of cultural values.

In this context, the study of the structure and dynamics of universities gains importance, as well as research performance assessments that try to guarantee an efficient use of resources and promote an internal quality culture. Due to the central role of universities in the education and research system, both policy makers and the society as a whole are interested in the results of the evaluation processes of universities. National and international rankings of universities (SJTU, 2005; THES, 2005), based on prestige indicators, but also on different structural, input and output indicators have emerged during the past years creating a healthy competence among universities to be on the top.

Spain's involvement in scientific research has increased steadily over the last twenty years, due to different scientific policy measures and to its integration in the EU in the mid 80s. According to the number of papers in Thomson Scientific databases Spain improved from a 15th rank position in 1982 to a 10th position in 2004, contributing at present to 3% of world's papers. Nevertheless, Spanish papers receive only 0.7% of world's citations, indicating a low international impact of its research in most fields. The public sector is the main producer of ISI papers through universities, the Spanish Research Council (CSIC) and hospitals. Enterprises are nearly invisible in ISI databases, as they traditionally had a low involvement in basic research in Spain.

Spanish universities have faced different reforms in the past years. Since the late seventies a trend towards de-centralization of the Spanish government into autonomous regions (NUTS2) has originated the creation of new universities together with the change of their administrative dependence to the regional governments. Consecutive reforms tried to increase the quality and efficiency of the Spanish universities. At present there are different types of universities in Spain: public/private, generalist/specialised, small/large, central/peripheral, etc. We consider it is interesting to face the study of the behaviour of the university system in Spain including both input and output indicators and exploring the influence of the above mentioned features on research performance.

¹ This study has been supported by the Spanish Ministry of Education and Science through the research project SEJ2004-08052-C02-02 /SOCI. We want to thank Luis Sanz and Laura Cruz for their valuable contributions to the document and Laura Barrios for her help in statistical analyses.

This study is part of a research project on dynamics of change in research organisations which focuses on two main dimensions of their activity: the production of new knowledge and the way universities manage their inputs, determine their competitive research strategies and how it is connected with the labour markets of researchers.

Objectives

The objective of this paper is to describe Spanish universities by means of structural, input and output indicators, to explore the relationship between those indicators and to describe university behaviour in different dimensions. We would like to contribute to the knowledge of the research university system in Spain, producing data that could be useful for research management both at the institutional, regional and national level.

Different questions are addressed: How are productivity and performance related to structural indicators of universities? Are they influenced by the size and age of the institutions? Is there any relation between university size and productivity or visibility of research? Can we distinguish generalist and specialised universities? How are productivity and performance related to relative specialisation of academic human resources across universities? Can we characterise teaching and research universities? Have public and private universities different interests and behaviours?

Methodology

In 2005, Spain had 48 public and 21 private universities. Two universities providing only “distance education” (one public and one private) were only considered for the thematic analysis of universities.

The following selection of structural input and output indicators for Spanish universities are included in the study.

- a) Structural and organizational indicators. Different indicators for the Spanish Universities were collected from the Spanish Institute of Statistics (INE, <http://www.ine.es>).
 - Age: number of years since the creation of the university
 - GDP per NUTS-2 region relative to EU-25 average, as indicator of the regional development that may influence the activity of the university
 - Administrative type: public (PB) or private (PV).
 - Total number of professors and number of them with PhD degree
 - Number of students (not including doctorate students)
 - Input specialisation: different specialisation profiles of universities were described by means of the thematic distribution of PhD professors over nine thematic areas. The following areas were considered: Agriculture-Biology-Environment, Biomedicine, Chemistry, Clinical Medicine, Engineering, Humanities, Mathematics, Physics and Social Sciences. The concentration of professors over areas was analysed through the Pratt index. This index ranged from 0 (low concentration) to 1 (high concentration).
- b) Scientific output. Scientific publications of the Spanish Universities during 1996-2004 were downloaded from the SCI, SSCI and AHCI (produced by Thomson Scientific, previously ISI) in order to analyse and describe different aspects of the activity of these institutions.
 - Activity: measured through the number of publications
 - International orientation: percentage of ISI documents as compared with total production (including Spanish databases ICYT and ISOC)
 - Impact: analysed by means of the number of citations per document, percentage of non-cited documents and percentage of publications in multidisciplinary high impact factor journals (*Science*, *Nature* and *PNAS*)
 - Output specialisation: thematic distribution of ISI publications over the nine thematic areas previously described was analysed. Different types of universities were identified

according to their thematic profile and the specialisation/diversity of documents by areas was quantified by means of the Pratt index.

- Collaboration practices: involvement of the universities in national and international collaboration was analysed in relation to the average values for the whole country.
- Collaboration with the private sector, as an indicator of the involvement of the university in solving societal needs.

Different techniques of multivariate statistical analysis were used to group universities with a similar thematic profile (k-means clustering) and to explore the relationship between structural, input and output variables (factor analysis). The software used was SPSS (version 13).

Results

Public vs. private universities

In Spain there is a long tradition of public universities: the University of Salamanca was created in 1218, followed by ten more up to 1700. The first private university was a catholic one, and dates from 1886 (University of Deusto). Along the twentieth century and first years of the twenty-first many other public and private universities were created.

Public and private universities differ in their structural and main output features (table 1). Public universities are older than private ones and show a larger size as measured through the number of students and professors. Professors with a PhD degree were compared to the total number of university tenured professors, but significant differences were not found between the two types of universities in the percentage of professors with PhD degree. Private universities are associated with a higher GDP, since they have been set up mainly in the rich regions while public ones are highly distributed geographically. Concerning the output, a higher productivity per professor in ISI publications and thesis awarded is observed for PB universities, which also obtain a higher impact.

The thematic profile of public and private universities is analysed through the distribution of documents over nine scientific areas. The Pratt index (PI) of publications was calculated to enable inter-university comparisons as to concentration/diversity of publications over thematic areas. It showed higher thematic concentration for PV than for PB universities (0.59 vs. 0.45; $p=0.000$). The highest concentration (high PI) corresponded to the newest universities as well as to the technical ones. Figure 1 shows the relationship between age and output specialisation of public and private universities. Old universities, created before 1800, are general universities, with PI values under 0.45. A wide range of PI values was found for PV universities, located at the right of the figure. According to the output specialisation, three different groups of universities with similar thematic profiles were identified: four universities which show a high specialisation in Social Sciences, Humanities and Mathematics; Technical universities with high activity in Engineering; and the rest, more or less generalist universities, not so well defined.

As most private universities were created very recently and only 3 of them produced more than 250 ISI documents in the last decade (U. Navarra 1889 docs; U. S. Pablo CEU 354 docs; U. Ramon Llull 273 docs), only data of public universities are studied in the next section.

Public universities

The behaviour of 47 public universities distributed over 17 NUTS2 regions is analysed in this section. The highest number of universities was found in Andalucía (9), Cataluña (7), Madrid (6) and Comunidad Valenciana (5). On the other hand, 8 regions had only one university.

Table 1. Structure and scientific output of public and private universities

	Private (N=17)	Public (N=47)	Total (N=64)
Age	26.5±33.32 (4-120)	141.06±221.75 (7-788)	111.97±198.19 (4-788)
No.Students	6014±3646 (723-11984)	25870±16917 (6021-84293)	20827±17073 (723-84293)
GDP per NUTS region	113.65±14.94 (79.4-124.7)	92.50±20.11 (59.9-124.7)	98.12±20.99 (59.9-124.7)
Tot.No.Prof.	528.76±347.45 (98-1317)	1893.04±1195.48 (477-5989)	1530.66±1201.26 (98-5989)
No.PhD prof.	207.8±188.80 23-762	824.45±636.88 109-2969	660.65±616.90 23-2969
% PhD prof/tot.prof.	36.68±13.21 16.06-58.16	40.40±10.57 19.19-61.29	39.41±11.34 16.06-61.29
No.Students/tot.prof.	10.77±3.59 6.23-18.74	13.55±2.11 7.83-18.83	12.84±2.81 6.23-18.83
No.Students/PhD prof.	33.27±14.70 13.07-62.40	35.38±8.92 23.15-62.35	34.85±10.59 13.07-62.40
No.PhD thesis/1000 students	2.68±4.56 (0-18.63)	5.53±4.06 (2.49-24.47)	4.74±4.36 (0-24.47)
No. ISI Publications	220 ± 573 (1-2403)	3297 ± 334 (186-14570)	2479±3174 (1-14570)
No.ISI Doc./PhD.Prof.	0.54±0.77 (0-3.15)	3.72±1.57 (1.12-8.86)	2.91±1.98 (0-8.86)
No.Cit./artic.	1.73±1.33 (0-4.08)	2.40±0.63 (1.29-3.95)	2.22±0.92 (0-4.08)
% Non-cited artic.	58.95±20.08 (32.19-100)	45.62±5.71 (32.7-59.41)	49.31±12.91 (32.19-100)
Output Specialisation	0.59±0.15 (0.39-0.92)	0.45±0.09 (0.29-0.68)	0.48±0.12 (0.29-0.92)

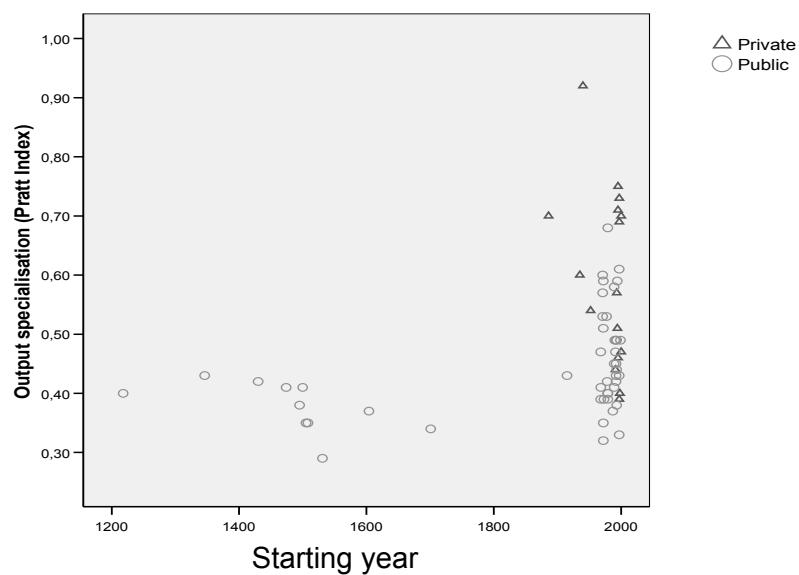


Figure 1. Age and output specialisation of public and private universities

Age. Four groups of Spanish universities can be distinguished according to their age. A total of 11 public universities (23%) were created along the 13th to 16th centuries. A second group of 17 universities (36%) were set up from 1900 to 1979. Finally, 5 universities (11%) started in the period 1980-1989, and 14 (30%) from 1990 up to the present.

Size. The largest university is the Complutense in Madrid (UCM), with around 84,000 undergraduate students, while the smallest (Cartagena Technical University, in Murcia) has only 6,000 students.

Input and output specialisation. The input specialisation of universities was analysed through the distribution of PhD professors over nine scientific areas. To make inter-university comparisons as to the concentration/diversity of the scientific effort over areas, the Pratt index was calculated. The oldest universities show low concentration, (high distribution of professors by areas), while a wide range of PI values were observed for the most recent ones (figure 2).

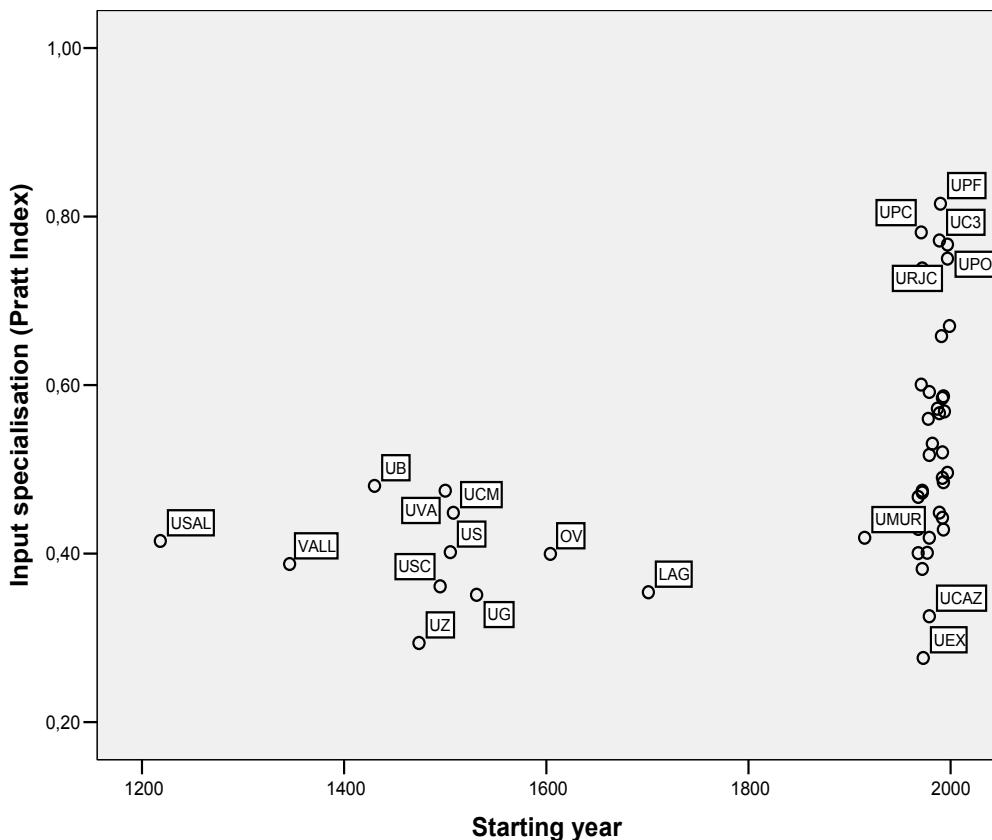


Figure 2. Age and input specialisation of public universities

The most specialised young university is the UPF, specialised in Social Sciences-Humanities. The Poly-technical universities, with a clear technical profile, are a special case, as they existed as Engineering Technical Schools since the 19th century, but were recently organised into Poly-technical universities (figure 2).

On the other hand, five different groups of universities with similar input specialisation were identified through cluster analysis: 1) 4 universities specialised in Social Sciences (UPF, UPO, UNED, URJC), which show a high Pratt index (0.77); 2) 4 universities specialised in Engineering and Mathematics (UPM, UPCT, UPV, UPC) (PI=0.69); 3) 19 general universities in which Social Sciences and Humanities predominate (PI= 0.40); 4) 4 universities in which Agriculture-Biology-Environment and Biomedicine predominate (PI=0.48); and 5) 17 universities with high activity in Social Sciences, Engineering and Humanities (PI=0.54).

Table 2. Thematic specialisation of the five clusters of universities

	Clusters				
	1	2	3	4	5
<i>Agriculture-Biology-Environmental Sciences</i>	1.70	10.71	6.57	24.30	6.03
<i>Biomedicine</i>	6.22	1.29	10.73	17.70	5.00
<i>Physics</i>	1.29	7.29	5.96	2.50	5.86
<i>Humanities</i>	17.26	6.62	19.03	14.24	15.19
<i>Engineering</i>	7.63	53.14	8.83	5.43	17.34
<i>Mathematics</i>	2.78	11.16	5.17	4.64	5.51
<i>Clinical Medicine</i>	1.24	0.24	7.70	10.24	2.41
<i>Chemistry</i>	3.22	2.01	7.88	2.91	6.61
<i>Social Sciences</i>	58.10	7.33	27.96	17.85	35.98

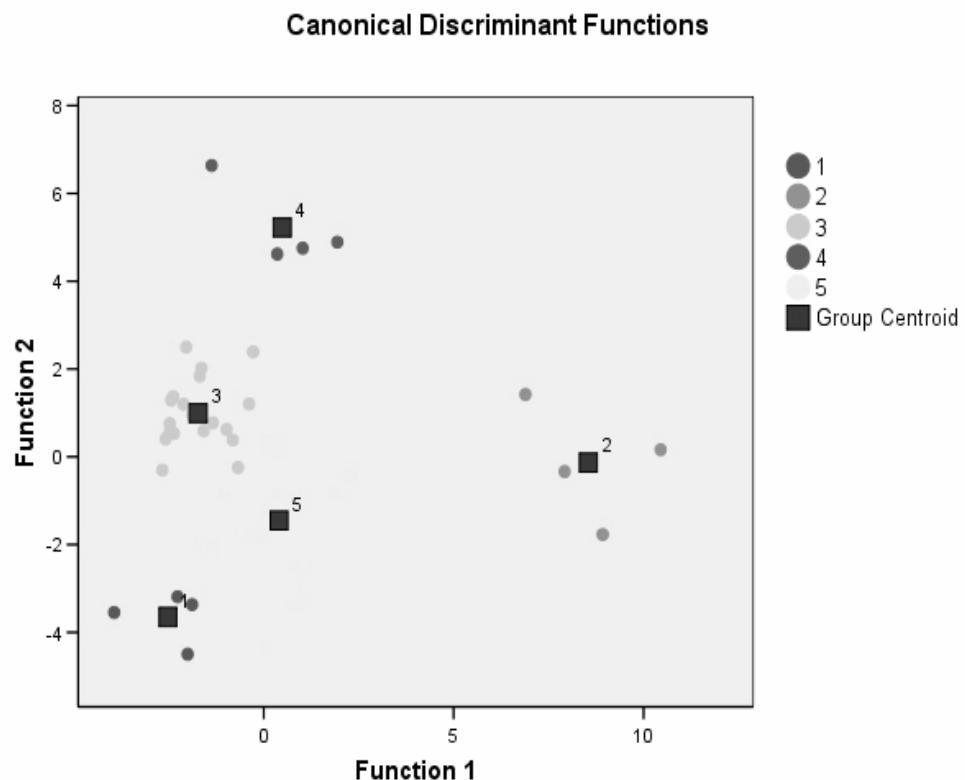


Figure 3. Clusters of universities by thematic specialisation of PhD professors

The different clusters are shown in figure 3, in which the axes correspond to two canonical functions that represent 87% of the variance. To the right is the Engineering-Mathematics cluster, while bottom left is cluster 1, specialised in Social Sciences and Humanities. In the middle are generalist universities of cluster 3. Although clusters 4 and 5 have low PI, 4 on the top shows a predominance of Agriculture and Biomedicine and cluster 5 is a bridge between Engineering and Social Sciences and Humanities.

Relation between structure and research performance

To explore the behaviour of universities, the relationship among structural, input and output variables (described in the Methodology) is analysed through factor analysis. Four different factors are found which explained 74% of the variance (table 3 and 4).

Table 3. Total variance explained

Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	4.0	26.49	26.49
2	3.95	23.22	49.71
3	2.12	12.49	62.20
4	1.95	11.45	73.65

Table 4. Rotated Component Matrix

	Component			
	1	2	3	4
No. PhD professors	0.969			
No. students	0.921			
No. ISI publications	0.874	0.404		
No. PhD thesis	0.835			
University age	0.673			-0.422
No. students/PhD prof.	-0.620			
No. citations/article		0.919		
No. ISI doc/PhD prof.		0.868		
% internat. vs. Spanish publications		0.818		
% non-cited articles		-0.815		
% doc. in top journals*10		0.608	0.398	
Input specialisation (Pratt-PhD prof)			0.858	
GDP of NUTS2 regions			0.689	0.354
International collaboration rate		0.535	0.627	-0.415
University-industry collab. rate				0.808
National collaboration rate				0.605
Output specialisation (Pratt-publicat.)	-0.310		0.441	0.516

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

Only loadings greater than 0.3 are shown.

First factor includes variables linked to size: number of PhD professors, number of students, number of ISI publications and number of PhD thesis awarded. These variables correlate positively with the age of the universities and negatively with the teaching load measured through the number of students per PhD professor and with PI of ISI publications. The bigger the universities the larger the thematic diversity observed.

Second factor refers to scientific productivity as measured through the number of ISI documents per PhD professor, and to visibility through the number of citations per article and percentage of publications in top journals. It is negatively correlated to the percentage of non-cited articles. A high international orientation, both through international vs. national publications and through international collaboration, contributes positively to explain this factor.

Third factor indicates that the richest regions have highly specialised universities according to the thematic distribution of PhD professors, which correlates with the trend towards international collaboration and to publish in top journals.

Fourth factor refers to the third mission of the universities, as measured through the percentage of documents in collaboration between universities and industry. A positive influence of national collaboration is observed. It correlates positively with specialisation and regional development and negatively with age of institutions and international collaboration.

Figure 4 shows the locations of universities according to their behaviour in the first two components. The two large and older universities of Madrid and Barcelona are located at the right of the figure (UCM and UB). The position of UPF and UAM at the top of the figure is remarkable, and is due to their high productivity and visibility. On the other hand, the position of UC3 and UPM at the bottom of the chart is partly explained by their thematic profile of activity. UC3 is a small university with strong orientation to Social Sciences and Engineering/Mathematics. Its low position in the productivity/visibility factor can be explained by the under-representation of Social Sciences & Humanities and Engineering in the ISI database, together with the lower citation scores of these topics. The position of the Poly-technical university of Madrid (UPM) is explained by its larger size and the small visibility of Engineering in the ISI database, as conference proceedings are not covered.

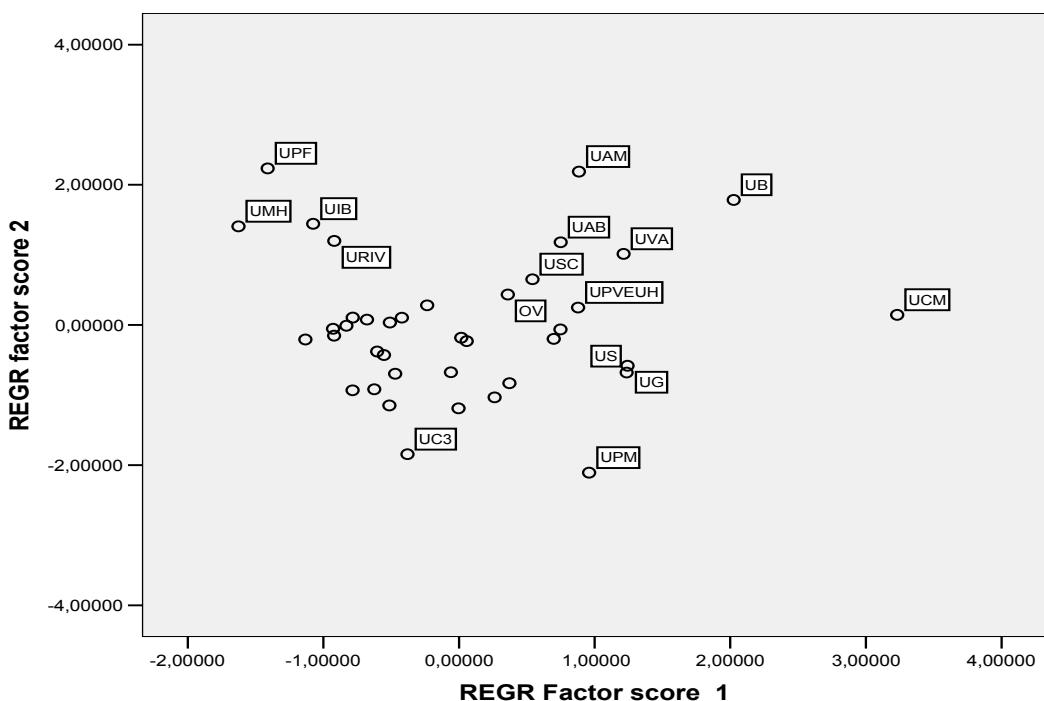


Figure 4. Distribution of universities according to factors 1 and 2

Discussion

Public and private universities

Public universities are more numerous, as well as older and larger than private ones. They are also more often generalist universities. The fact of a university being general or specialised is in part related to its age, since the oldest universities have gradually grown to cover a large diversity of disciplines. Less growth has been described for private universities, since they seem to control their thematic dispersion following profitability criteria and offering competitive speciality degrees (De Miguel et al. 2001).

A particular case is that of the Engineering Technical Schools, which are specialised in spite of their long history. In fact, they were officially established as separate schools in Spain at the beginning of the 19th century (Civil Engineers, Industrial Engineers, and so on), but they were only recently aggregated into Technical universities in 1971. They are traditionally more oriented to teaching and to solving applied problems than to research, following a similar model to that of Germany (Moed, 2006).

It is interesting to note that public universities show a higher research orientation than private ones (higher national and international productivity, higher number of thesis per PhD professor). Again the lower age of private universities can be an explanatory factor, since some of them have not had enough time to develop their PhD programs. But also, teaching predominates over research in private universities because many of them are principally oriented to provide training in areas of professional interest for society. Accordingly, the number of professors per student in private universities is higher than in public ones. In our study we focus on “research productivity”, measured through scientific output, whilst other aspects related to teaching are not considered. Private universities usually obtain “good marks” in teaching related productivity, measured through the rate of graduate students, unemployment rate among graduates or type of job obtained, but these aspects were not analysed in the present study.

In Spain there is no official difference between teaching and research universities, although in 1968 the two “Autonomous universities” in Madrid (UAM) and Barcelona (UAB) were created to enhance quality and research activities, with a higher proportion of professors per student than was the average at the time. Later in the nineties, again searching for quality, two other small public universities with a high rate of professors per student were created in Madrid and Barcelona: UCIII and UPF, respectively. In our data, UAM is quite outstanding considering eight research performance indicators, including high productivity of ISI documents and thesis per professor, publications in top journals and citations per article, while UPF shows high positions in research productivity and visibility indicators. Among the private universities, the Navarra university (UNAV) shows good output values at the level of the best public universities, considering citations per article or thesis per professor, together with a high performance as to teaching results (it has a high number of professors per student). Both UNAV and UAM are universities created more than 30 years ago.

Size, age and specialisation

The analysis of the relationship between structure and performance of universities has been limited to the public ones, due to the low scientific output of many private universities, partly explained by their short history.

A positive correlation between university size –measured through the number of students and professors- and absolute number of ISI publications and thesis is observed. However, ISI productivity does not always increase with university size. Interestingly, the higher productivity is associated with a high international orientation and a high impact. That means that all large universities do not obtain high visibility and productivity, but it is easier for universities with high production to obtain them. This finding is coherent with previous studies, which have observed that the largest Spanish universities show higher relative impact than the university sector of their own country, as described also for other European countries (Moed, 2006). Finally, the positive effect of international collaboration on impact is being observed, in coincidence with previous results (i.e. Narin, Stevens and Whitlow, 1991).

Two different criteria were used in this study to define a university as general or specialised: the distribution by areas of: a) PhD professors (input specialisation), and b) scientific publications (output specialisation). According to the profile of the input specialisation, two groups of specialised universities are identified: technical universities and those specialised in Social Sciences & Humanities (with Pratt index 0.69 and 0.77 respectively). A cluster of generalist universities (PI=0.40) includes all the century-old universities and some more recent ones. Two other clusters are very slightly specialised in Agriculture-Environment and Biomedicine, on the one hand and on Engineering-Social Sciences-Humanities on the other. It is interesting to note that the highest productivity in ISI documents and PhD thesis per professor corresponded to the generalist cluster, data that support the high involvement in research of these old universities.

Interestingly, large and old Spanish universities are the ones best located in the Shanghai (SJTU) and The Times ranking (THES). Only the Universidad de Barcelona (UB), which is the largest Spanish

university according to its ISI publications, is classified among the top 200 of the world. Considering the top 300 universities, the two largest Madrid universities (UCM and UAM) are also included.

However, we should mention that some young specialised universities located in rich areas show signs of good performance, such as high level of international collaboration and publication in top journals (i.e. UPF).

Concerning specialisation, a higher thematic concentration was found according to the input-specialisation than to the output measure. The problem of the latter is that Social Sciences & Humanities are under-represented in the output side, as their publications are not properly covered by ISI databases, due to the importance of books and national journals in these areas. We can see that around 45% of PhD professors and almost 60% of the total students of Spanish universities belong to SSH, vs. only 9% of the scientific publications of Spanish universities in the ISI databases. This is an important limitation of the output-specialisation indicator that could be overcome including publications from other sources besides ISI databases.

Third mission

Concerning the third mission of universities, this study has analysed only one indicator based on publications, which is the collaboration between university and companies as recorded in the address of publications. The strength of the links between these sectors is a partial indicator of knowledge flows between academy and industry, as has been stressed in different studies (Leydesdorff, 2003). Our results show that as national collaboration increases the collaboration with companies also tends to rise, and that it is more frequent in young specialised universities located in rich regions, such as Madrid, Cataluña or Valencia. In fact, it has been described that a strong industrial activity in the same area of the university creates synergy effects in European regions (Zitt, 2003). On the other hand, it is probable that young universities are more aware than old ones of the interest in taking part in joint research with companies. Regional government actions to foster triple helix effects play also an important role promoting university links with local industry. Some of the results are a bit surprising; since several universities located in highly industrial regions, as is the País Vasco, do not show outstanding values of collaboration with industry. However, we have only analysed collaboration in ISI publications, and industry is frequently more visible in national publications (Gómez et al. 2006). Moreover, we consider that it would be interesting to introduce other input indicators, such as the number of joint contracts between industry and university as well as patents, to complement the present study.

The results here shown are part of an on-going research project that aims at providing useful data for the research management of universities. We would like to support policy makers providing data on different aspects of the behaviour of the university system. Which structural features are more positive for research? Do specialised universities attain higher levels of excellence in their areas of expertise than generalist ones? Which should be enhanced, generalist or specialised universities? Can we measure the positive effects of combining teaching and research in the same centre and even in the same persons, as has been frequently defended in the literature (Universities UK, 2003)? In spite of the interest of attaining excellence in research, is it true that national level research can be more useful to local industry and is more involved in the third mission of the university?

Our study provides a general overview of the behaviour and characteristics of Spanish universities showing differences between public/private and generalist/specialised universities. Relationships between structural, input and output indicators are put forward. The higher research orientation of some public traditional universities is observed. We consider that an in-depth analysis of inter-areas differences is essential. Therefore, the performance and characteristics of the different universities within each area will be studied in the next future.

References

- De Miguel, J.; Caïs, J.; Vaquera, E. (2001). Excelencia. Calidad de las universidades españolas. Centro de Investigaciones Sociológicas, Madrid.

- Funding Research Diversity. (2003). Technical Report. Universities UK.
- Gómez, I.; Sancho, R.; Bordons, M.; Fernández, M.T. (2006). La I+D en España a través de publicaciones y patentes. En: Sebastián, J. and Muñoz, E. Eds. *Radiografía de la investigación pública en España*. Biblioteca Nueva. Madrid.
- INE. Instituto Nacional de Estadística. <http://www.ine.es/inebase>
- Leydesdorff, L; Meyer, M. (2003). The Triple Helix of university-industry-government relations. *Scientometrics*, 58 (2), 191-203.
- Moed, H.F.(2006). Bibliometric ranking of world universities. CWTS Report 2006-01. Leiden..
- Narin, F; Stevens, K.; Whitlow, ES. (1991). Scientific cooperation in Europe and the citation of multinational co-authored papers. *Scientometrics*, 21, 313-323.
- SJTU. (2005). Academic Ranking of World Universities–2005. Shanghai Jiao Tong University, Institute for Higher Education. Retrieved 7, 2006 from <http://ed.sjtu.edu.cn/rank/2005/ARWU2005Main.htm>
- THES. (2005). World University Rankings. Who is Number One? The Times Higher Education Supplement. <http://www.thes.co.uk/worldrankings/> (7 Nov. 2006)
- Zitt, M.; Ramanana-Rahary, S.; Bassecoulard, E.; Laville, F. (2003). Potential science-technology spill-overs in regions: an insight on geographic co-location of knowledge activities in the EU. *Scientometrics* 57(2): 295-320.

Revisiting the "Heroic" Age: From Externalism to Internalism in Serial History of Science¹

Yuri Jack Gómez

yjgomez@unal.edu.co

Universidad Nacional de Colombia, Department of Sociology, Social Studies of Science and Technology Unit,
Carrera 30 No.45-03 Ciudad Universitaria, Edificio 205, Of. 230. Bogotá (Colombia)

Abstract

This paper attempts a re-valuation of a set of historical (bibliographic) evidence germane to the so named "heroic age" of bibliometrics, an age commonly dated on twentieth-century first half. This revaluation is conducted under a frame-work somehow neglected by the historiography of science and the historiography of bibliometrics, in particular. Contrasting with much of the literature that references "heroic" works as episodic, unconnected and isolated efforts, this paper examines the "heroic" literature in the context of the debate among historians between those supporting the idea of an history of the unique (*historie événementielle*) and those supporting a *serial history*. From this particular point of view, this paper deals with the "heroic age" of bibliometrics and presented it as a continuous and systematic effort for developing a serial history of science whose major concern was to provide some sort of social explanation for science's development and growth. Its major achievement was, however, methodological in as much as the heroic age might be characterised as a quest for a standard data source for making the history of science.

Keywords

history of bibliometrics; serial history of science; journalization of literature ; standardisation of data sources.

Introduction.

According to Merton's (1977) reputable account, bibliometrics was some sort of method in search of a theory. Since the method was scientifically used for the first time in the advancement of the empirical research programme on the scientific community, bibliometrics, Merton adds, should be considered as a speciality specific research method for the sociology of science. Despite this, it is my contention that bibliometrics, as a field of research, has an 'independent' origin of its own. Merton's account had to do much more with the process of acquisition of professional and cognitive identity for the sociology of science taking shape at Columbia University, than with the actual emergence of bibliometrics. The transition from socio-historical to empirical investigations in the sociology of science followed, at Columbia, a well-established pattern of *boundary work*² in which the utilisation of quantitative techniques was considered a hallmark of scientific status for an intellectual discipline (Camic, C. & Xie, Y., 1994). In this work, an independent history of bibliometrics is attempted by revisiting the so-called heroic period of the field. From this exercise two main conclusions emerge, First, that the history of bibliometrics might be fruitfully linked to contemporary historiographic debates on *serial history*. Second, that as serial history of science, bibliometrics' first great achievement was the definition of a standard source of historical data: the scientific periodical.

The uses of quantification in early history of science.

The institutionalization of the history of science, it is commonly stated, started to take place in the 19th century, when a clear tradition of historical analysis, continuity of problematics, and communication among practitioners can be clearly detected (Thackray & Merton, 1972; Merton, 1977). Unlike science, this history of science dealt with the description of a situation or state of affairs which was unique, and this assumption guided the assessment of significance in history. This kind of history cannot be

¹ Thanks to COLCIENCIAS and the Universidad Nacional for supporting my attendance to the Conference with a travelling grant.

² Boundary-work (Geyerin, 1983) is a notion initially formulated for explaining how scientists construct or maintain the boundaries of their community against threats to its cognitive authority, whether these treats come from within or outside. Boundary-work has found useful policy-relevant applications (Guston 1999, 2001). In this context, derivative notions such as boundary-objects, boundary-organizations, and even co-production have been advanced (Star and Griesemer, 1989; Sheila Jasanoff, 1996; Bowker & Star, 1999).

defined as a mere narrative of certain selected ‘events’ (a great man of science, a scientific discovery) along the time axis. First and foremost it was based on the idea that these events were unique and could not be set out statistically or compared to any antecedent: what there is of the unique and incomparable in an event is what makes it historical³. The historian of the unique, was not merely interested in describing facts but in understanding them. Understanding calls for interpretation, classification and assessment that can only be attained by grasping the structuring relationship of relevance and causation that may account for the sudden irruption of the event into the succession and concatenation of historical facts.

From this standpoint, historical events are not necessarily or simply temporal. Every occurrence is a concurrence, a hierarchy of events and sub-events. Reality occurs to the mind not in unique analytical samples, but in unique contexts, which are eventful. Hence, the framework of an event is both chronological (time-series made of antecedents and contemporaneous) and contextual (relationship-series made of significance and causality among antecedents and contemporaneous) (Meadows, 1944). History is not chronology, but chronology is always the first step of the historian. The chronological and systematic organisation of the evidence (bibliographic evidence most of the time), and its further quantitative examination constituted an ancillary technique serving the purpose of planning historical investigation (Thackray & Merton, 1972).

A typical example of this quantitative treatment of the literature as an ancillary technique is Sarton’s report to the Carnegie Institution in 1923 in which numerical ratios were displayed to illustrate the importance and ‘relative progress of thought’ accomplished at different periods outside classical antiquity and, at the same time, these curves and tallies provided a work-framework displaying the extent of the task in front of the historian.

The application of these techniques for marshalling the literature, although instrumental and useful for the assessment of historical significance was, nevertheless, different from the historical assessment itself. To the historian of the unique the substance of history is not constituted by the quantitative patterns used for planning the work, rather it is the historian’s assessment based on his or her knowledge of the event and its context that constitutes the very substance of history. In fact, as Merton and Thackray underlined, Sarton was never keen to make public this sort of historiography-in-action and he never published as historical contributions what he surely considered simply his quantitative craftsmanship; he did not conceive of it as substantive history.

Perhaps as a result of the particular understanding of historical facts as unique events, and of substantive history as a narrative in which that event’s uniqueness is stated and assessed, an alternative narrative form has been systematically neglected⁴ or downgraded⁵ in the official accounts of the emergence of the history of science namely *serial history*⁶.

In general terms serial history is a history oriented toward the ascertaining of patterns that occur in time and towards the understanding of historical change and historical regularities, in quantitative terms; a completely different understanding all together compared with that of the historians of the scientific-event. Serial historians of science considered it possible, for example, to ‘reduce to geometrical form the activities of the corporate body of anatomical research, and the relative importance from time to time of each country and division of the subject’ (Cole and Eales, 1917: 578); or to resolve ‘the *Zeitgeist* into its component parts by statistical method’ (Hulme, 1923: 30); or that one can establish the quantitative laws ‘governing the production of scientific and technological

³ See, Sorokin (1931); Meadows (1944); Joynt & Rescher (1961); Furet (1971).

⁴ See, Joynt et al. (1961); Aydelotte (1966); Marczewski (1968); Price (1969); Furet (1971); Jensen (1974); Swierenga (1974); Erickson (1975); Fogel (1975); Thackray et al. (1972); Thackray (1978).

⁵ See, Edge (1980)

⁶ According to Pritchard ‘classical’ bibliography on bibliometrics, the first contribution within the genre, a simple counting of chemistry publications, is as old as 1874. More often, though, it is the contribution on the history of comparative anatomy by Cole and Eales (1917) which is commonly acknowledged as the older work within the genre. Noteworthy to mention is also the little bibliographic note of Fredrick Woods (1911) who proposed the term ‘historiometrics’ as the ‘new name for a new science’.

discoveries of mankind' (Weinberg, 1926: 44). If for the historians of the unique discoveries were unique and of individual significance, for serial historians all discoveries were equally valuable as units of measurement. Serial historians were interested in discoveries *only* because they are all *equally* discoveries and further because they are *different* discoveries (Rainoff, 1929: 289-291).

Serial historians and historians of the unique share the same kind of historiographic resources, histories, bibliographies and chronologies used and produced throughout the centuries. However, the intellectual focus of serial historian was not the individual event in its uniqueness; serial history is not *historie événementielle*, but quantitative history. Serial historians were interested in the construction of long-term series (predominantly bibliographical) whose quantitative description and statistical generalisation constituted for them the matters of fact of the history of science. From this stand, the traditional ancillary materials and techniques used in the history of the unique turned out to be the very substance of serial history accounts.

I preferred the expression, 'serial history', instead of 'statistical bibliography or bibliometrics of science' (Thackray, 1978) for several reasons. First, although the expression 'statistical bibliography' might describe an important methodological aspect of 'serial history' it obscures the fact that serial history is more than merely statistics, it is first and foremost, history; and by making 'statistical bibliography' equivalent to 'bibliometrics of science', the idea that bibliometrics is merely a methodological resource gets reinforced. Second, as an institutionalised field of research and as a widely accepted expression, 'bibliometrics' is recognisable only in the late 1960s. Although bibliometrics shares with early contributions to the genre many assumptions and procedures, it seems somehow improper to refer to these early contributions as bibliometrics because there are important differences between these two sets of works which I will address in terms of serial externalism/serial internalism respectively drawing on Shapin's (1982; 1992) and Daston's (2001) organisation of the history of science as a field of studies⁷. Thus, by 'serial history' I embrace both the early contributions of 'statistical bibliography' (externalist serial history) and the late bibliometrics (internalist serial history) in as much as they both share a common narrative of a historical continuity (science) in the form of discontinuity (bibliographic data series): *serial history*.

Indeed, serial (externalists) historians considered that the lists of inventions and publications chronologically arranged and statistically summarised provided an insight into science as an historical process. By contrast with the historians of the unique and its narratives of discontinuous events in the mode of continuity, serial (externalist) history is a narrative of a continuous process (of production) in the form of a discrete or discontinuous (bibliographical) data series. Serial (externalists) historians were not interested in understanding what makes a particular contribution to science unique and significant. Rather their major concern was to identify and characterise trends in the production of 'science' by applying statistical techniques to bibliographic material. And, like sociologists, they were in search of general explanations for the quantitative patterns that the scientific production exhibits by relating this patterns with other data-series from the economic and demographic realm, this is why I have distinguished this early representatives of serial history of science as 'externalists'.

Just like Price, early serial historians did not fix their gaze on that 'specific molecule called George, travelling at a specific velocity and being in a specific place at some given instant'. Rather they attempted to understand science as the totality of its bibliographical 'events', and to describe this understanding in quantitative terms. Just like him, they defined science, to a considerable extent, as the bulk of published science. Just like him, they perform measuring operations upon this production, i.e., upon bibliographies, chronologies biographies and historical accounts as sources of bibliographical data⁸. Just like him, they proposed general models to describe science in quantitative terms based on

⁷ It must be said that none of these reviewers of the history of science as a field of studies, or either of the other reviews on quantitative history used for this work, talk about 'serial history of science'; let alone terms such as 'serial-internalism' or 'serial-externalism' in relation to the history of science.

⁸ Price was not as eclectic as his predecessors in the selection of his data sources; particularly he tried to restrict the use of historical accounts to the minimum. This makes a huge difference.

the historical behaviour of the literature⁹. And like him, theirs were attempts to relate these quantitative representations of scientific production with other series in the economic, demographic, and geopolitical realms (as with Hulme, Rainoff or Gross). The substantial difference was, however, that the models used to organise the data-series made by serial-externalists were taken from these other 'external' realms and not from the internal structure of the data as Price did.

There are subtleties of course. Like the histories of comparative anatomy (Cole and Eales) or the history of nitrogen fixation by plants (Wilson and Fred), which are like abridged compendia of *historie événementielle* illustrated with charts very much like Sarton's unpublished craftsmanship. However, it is the emphasis on long-term trends, the recursive deployment of bibliographic material as measured object (instead of retrieval tool or research plan) and the particular understanding of science as a continuous process of production that makes me put together both bibliometrics and 'statistical bibliographies' as representatives of serial history of science.

The emergence of a Serial (externalist) History of Science and the problem of the standard data source: towards the 'Journalization of the Literature.'

The connection between serial history of science and the sciences of economics and demography is clearly detected in the contributions of Hulme (1923) and Rainoff (1929), and to some extent in the works of Gross (1927), Wilson and Fred (1935). All of them, to different degrees, attempted to provide causal (external) explanations for science development and growth by linking (via graphical superposition of data series) their bibliographic findings with economic historical data series. In the contributions of the former two there are explicit references to previous contributions made by economists. Hulme was somehow 'inspired' by the observed fluctuations reported by Cole and Eales and he found in R. A. Lehfeldt's (1916) 'normal law of progress' a plausible explanation for the fluctuations observed in the bibliographical series¹⁰. Rainoff elaborates on this uniform characteristic in terms of wave-like fluctuations already described by E Slutsky (1927) in his contribution 'on the summation of random causes as the source of cyclic processes'. Furthermore, Rainoff pointed out that the economic literature has established 'the existence of long waves in the economic dynamics of all the main West-European countries', a statement supported in a variety of economic studies ranging from Jevons' (1865) 'analysis of the variation of price and the value of currency since 1782', to Kondratieff's long-waves model in the movement of the general price-index for some European countries (Rainoff, 1929: 301ss, and n.5). Wilson and Fred (1935), commenting on the deviation of their empirical data from a logistic model of growth, considered that this deviation was due to the superposition of 'internal' and 'external' factors governing the development of the field of nitrogen fixation. Intensive agriculture, particularly in the USA, is underlined among the external factors influencing the deviation of the empirical data from the theoretical curve during the decade of the 1920s: 'the effect on the literature curve of nitrogen fixation by leguminous plants is unmistakable; the similarity of the curve with that of 50 representative stocks on the New York exchange for the period from 1923 to 1933 is readily apparent' (Wilson and Fred, 1936: 247-248).

⁹ Surely the models differ: different 'statistical approaches' (Camic and Xie, 1994) were used, different time periods were examined, and different bibliographic subject matters were considered. Whereas Cole and Eales, Hulme and Rainoff, for example, talked about cycles (a topic on which Sorokin (1927) elaborates from a quite different perspective), Willson and Fred (1935) talk about scientific literature as an organism in growth, and Weinberg in terms of a direct relationship between experience and productivity in the life cycle of scientist (1925) and of science exhaustion (1926). On this last point regarding the debate between theories of finalisation in science and the limits of scientific progress among science advocates in the turn of the century Europe see Rescher (1978).

¹⁰ In this contribution, published in the *Journal of the Statistical Society*, Lehfeldt stated that the statistical evidence in relation to progress suggests that this is a typical instance of a stationary condition. If that is so, Lehfeldt concluded, 'progress must consist in a transition from one stationary state to another, and this transition may be expected to show some uniform characteristics'. Hulme understood the fluctuations reported by Cole and Eales as example of this stationary condition of progress in science, and he supported this point of view by adding more evidence gathered in demographic studies: Griffith Taylor's paper 'on the future distribution of the white population' and Raymond Pearl's 'the population question' published in *The Geographical Review*, both in 1922 (Hulme, 1923: 29).

Early contributions to serial history of science were already aware of the need for the standardization of the data. Although Cole and Eales proposed publication as the basic unit of measurement,¹¹ a common characteristic of early contributions of serial historians was the heterogeneity of both the textual sources used in their studies and of the very data contained in those sources. Despite this awareness and the historiographical efforts to achieve completeness and consistency in the construction of historical series, early contributions to the field faced a diversity of problems¹².

For example, the bigger the time span of the series used, as in the works of Cole and Eales (1534-1860) or in Weinberg (3000 BC to 1940), the greater the heterogeneity of the data, and the more the reliance on historians' accounts for its compilation. Weinberg (1926), for example, based his investigation on the 'laws of the evolution of discoveries' entirely in Dramstädter's *Handbuch zur Geschichte der Naturwissenschaften und der Technik* (2nd edition, Berlin, 1908 — used in Merton's dissertation as well). This is a historiographic source, a chronology in fact, containing 12,000 dated discoveries and inventions since ancient times. Although Weinberg was aware of the several sources of bias in this chronology, he considered the source as a useful asset enabling the treatment of the history of mankind in a 'uniform manner' and thus, facilitating his enquiry into the laws of the evolution of mankind¹³. As expected, the time span used by Weinberg started in the year 3000 BC and the heterogeneity of the data is enormous: 'the invention of the axe', 'the cultivation of cereals', 'an inscription in an Egyptian pyramid', a publication, etc. Everything fell into the tally (Weinberg, 1926: 43-44). Although the reliance on historical sources cannot be entirely substantiated for the contribution of Cole and Eales, since authors gave no clue to draw a conclusion on this regard, the assumption that they were using historic-bibliographic materials is not entirely unjustified. When considering the time-span used for their history of comparative anatomy (1534 to 1860) and the very form of their narrative, which to a considerable extent consists of lists of names and works distinguished as 'important' 'prominent' and so on, but without any further considerations of why they were 'prominent', or in which regard they constitute 'important' contributions to the field other than the fact of having been associated with 'great men of science', 'formidable names', 'prominent anatomists' and the like. Cole and Eales, it seems, took at face value the judgments of the historians regarding the 'significance' of men and works, and proceeded to plot them 'in squared paper' and to 'reduce to geometrical form the activities of the corporate body of anatomical research, and the relative importance from time to time of each country and division of the subject' (Cole and Eales, 1917: 578). As a result of these operations of tabulation, plotting and geometrical translation performed on a list of bibliographic entries, they were able to claim an overall increase in the literature produced in a somehow fluctuating mode referred to in terms of 'revivals' and 'declines', whose causes they found in 'external' factors such as the demography of the field, the transformation in the means of circulation (the emergence of the periodical publications), the social organisation of science into scientific academies, and the market constraints of the printing and publishing enterprise.

¹¹ 'Publication is an isolated and definite piece of work, it is permanent, accessible, and may be judged, and in most cases it is not difficult to ascertain when, where and by whom it was done and to plot the results on squared paper' (Cole and Eales, 1917). It is noteworthy that this contribution of 1917 was the second attempt of Cole for writing a quantitative history of comparative anatomy. His first attempt was in 1914 (Cole, 1914) and in that occasion he used the 'museum' as unit of analysis but as he noticed, 'the result was more interesting than conclusive, principally because the number of such museums which could be traced (533) was too small to admit of satisfactory treatment by statistical methods' (Cole and Eales, 1917: 578).

¹² The reader should take into consideration at this point that secondary tools for bibliographic control of periodicals were a relatively late 19th century innovation. According to Vessenyi (1974) the world's oldest abstracting service is the *Chemisches Zentralblatt* which originated in 1830 in Berlin; the *Fortschritte der Physik* covering physics and mathematics was published in 1847; *Physical Abstracts* was published in 1847; the *Jahrbuch über die Fortschritte der Mathematik* ran from 1868 to 1942; *L'Année Biologique* was published in 1897, *Index Medicus* in 1879; *Chemical Abstracts* was published for the first time in 1907 by the American Chemical Society; the *Berichte über die gesamte Biologie* was published in 1920; *Biological Abstracts* started in 1926; *Psychology Abstracts* dates from 1927.

¹³ These laws will lay the foundations of what Weinberg named as 'humanology' a new sort of science concerning the evolution of mankind: 'cette particularité présentait une grande valeur pour mes études personnelles, car j'ai fait, au cours de ces dernières années, quelques tentatives pur découvrir des lois mathématiques de l'évolution de l'humanité et pur prouver ainsi la possibilité d'une science qui servirait à établir ces lois et que je désigne sous le nom d'humanitologie' (Weinberg, 1926: 44).

Regarding the heterogeneity of the textual sources and the prominence of the historical ones among them, Rainoff's (1929) study on the wave-like fluctuations of the scientific production in physics serves as a good example. Rainoff's is a statistical analysis of F. Auerbach's *Geschichtstafeln der Physik* (1910) a catalogue of discoveries, and the 'capital bio-bibliographical encyclopaedia of Poggendorf, with its subsequent continuations'. From the first source Rainoff gather data on discoveries, and dates, from the second, data on discoveries by countries. In addition, as a 'validation test', he used as 'independent sources' for data on discoveries, dates and countries an 'authoritative physical compendium', The *Course of Physics*, by O.D. Khvolson (Vols. I-V), the latest *Geschichte der Physik* (1926) by E. Hoppe, and two books by Mach — *Principien d. Wärmelehre* (1923), and *Principien d. physikalischen Optik* (1921). As a result of this 'validation' Rainoff presented a full-fledged historical series from 1771 to 1900 for physics. As Hulme (1923) (by contrast with Weinberg, Cole and Eales), Rainoff explained fully the constructive procedure of his data series from historical, bibliographic and educational text sources¹⁴.

As early as 1923 though, it was already noticed that the data series required for a serial history of science should be international in scope, and should enable the distinction between original work and other kind of publications. Furthermore, it was recommended that the statistical description of bibliographic data series should be always guided by a specific cognitive competence on the intellectual field under examination. In his contribution of 1923 Hulme prized Cole and Eales' contribution as exemplar of compliance with these three characteristics that a quantitative approach to the history of science must have. To Hulme, both Francis Cole, a professor of zoology at University College of Reading (now University of Reading in England) and Nelly Eales a bachelor in science working as a curator for the museum of the same College fulfil the requirement of cognitive competence. By the same token, their separate data series on German, French, and English production in comparative anatomy accounts for the international scope as required. The definition of the basic 'unit' of analysis as stated by Cole and Eales did not, however, allow them to establish a clear distinction between original and 'educational' publications to use Hulme's expression. In fact, although considered of little impact for the purposes of their study, they acknowledge, nevertheless, that their series contained data on publications of an 'encyclopaedic and textbook character' (Cole and Eales, 1917: 595).

Among those who further pursued a more restricted definition of 'publication' as unit of analysis there is a more clearer consciousness of the journalization of the literature by which I meant the gradual taking-over (normalization) of periodical literature as standard source for the serial history of science. Gross' (1927) study on the impact of the first inter-imperialist war (the Great War) on fundamental science is a good early example in which the standard characteristics of the scientific journal were deployed in historical investigation. Gross used as primary standard source the *Journal of the American Chemical Society* upon which he performed a citation analysis in order to ascertain by these means the impact of war on basic research¹⁵.

Hulme's (1923) approach to the question of the growth of modern civilisation is yet more significant for adopting a self-conscious bibliographic view point¹⁶. He considered that the 'fabric of modern civilisation is essentially artificial — resting upon a combination of the forces of Bibliography i.e. the science of the organisation of recorded knowledge, Education, and Research' and that bibliographic material 'furnishes us with the best mirror of the human mind'. Moreover, he added that 'whereas the data of the statistician are known to be progressively defective, as his survey extends from the beginning of the nineteenth century backwards, the records of human knowledge have on the whole been satisfactorily preserved from a much more remote period' (Hulme, 1923: 9). Thus, in his first lecture he proposed a 'bibliographic method' grounded in the idea that the examination of the changes

¹⁴ Hulme, 1923: 33-38; Rainoff, 1929:292-297.

¹⁵ For the entire volume of 1926 the *Journal of the American Chemical Society* published 459 works in pure chemistry containing 4857 citations to previous work. These references were to 247 different journals including the American journal source. From the increase in the citation rates to the American journal Gross concluded that for the case of pure chemistry the Great War exert a positive influence in the USA and constituted a powerful deterrent in the case of the combatants countries.

¹⁶ Bibliographic control was the particular 'field' in which Hulme claimed to have a specific cognitive competence.

in the intellectual attitude of mankind at different periods is best undertaken through its ‘literary’ (bibliographic) counterpart, as it is reflected, for example, in the evolution of bibliographical classification systems. He warns that although useful to some extent, the analysis of classification systems for libraries does not necessarily reflect the state of knowledge at a given time. A book classification, he claimed, is not a classification of current knowledge. Thus from the absence of a literary class no valid conclusion can be drawn as to the state of knowledge in respect of that class; the organisation of knowledge portrayed by literary classes and sub-classes does not always coincide with that of the subject matter since the latter ‘may be adequately represented in published works and yet find no counterpart in the book classification’ (Hulme, 1923: 10¹⁷). To Hulme the introduction of sub-categories within subject-matter classification systems was a useful indicator of what had happened in the frontiers of knowledge, but these modifications take place with a considerable delay. The application of the ‘bibliographical method’ for the study of the growth of civilisation then is only useful for historical investigations. Hulme pointed out that what is really important in relation to the progress of civilisation is the process of specialisation within the literary classification, not the summing up of bibliographical entries under each class or sub-class¹⁸. From the application of statistics to the phenomena of the past (to classification systems of books) ‘little in the way of novelty can be expected: confirming the conclusions of existing authorities, while adding something of precision to certainty’ (Hulme, 1923: 11).

This is why Hulme, in his second lecture, undertook a quantitative analysis of the *International Catalogue of Scientific Literature* (a bibliographical control tool for periodical literature) as providing a very firsthand view of the research front for the present. In this regard, Hulme’s contribution was an important step towards the journalization of standard sources for the serial history of science. In fact his use of the *International Catalogue* allowed him a systematic and internally consistent construction of the scientific production as a measured object. Going farther than Cole and Eales in defining ‘publication’ as a unit of analysis, Hulme, in practice, restricted this definition to the realm of journal publications, and he understood scientific growth in terms of the statistical analysis of the changing patterns of growth of periodical literature (journalization) under different headings and subheadings used by the *International Catalogue* which were more attuned with the actual specialisation of the sciences than the traditional classification system used in libraries. Taking advantage of relatively recent technical developments of bibliographical control over the periodical literature of his time, he was able to transform a standardized bibliographical control tool, a catalogue of periodical publications, into a data series for writing serial history of science¹⁹.

It should be noted, however, that to Hulme, the growth of periodical literature was the result of external (extra scientific) forces. As Rainoff did a few years after him (1929: 301ss), Hulme based his explanations on comparisons between production cycles in the sciences and other demographic and economic variables (Hulme, 1923: 37-38)²⁰.

¹⁷ Hulme illustrates the point: “Thus the appearance of the first monograph is of itself of bibliographical rather than historical importance. When, however, the subject of a monograph shows signs of division, this is a symptom of growth...A thousand textbooks on the Einstein Theory attest the width of the popular appeal, but when its literature splits into the countless aspects of its philosophical and experimental standpoints, we know that philosophy and science are readjusting their attitude towards the new doctrine and that the Einstein theory is taking a permanent place in the stock of recorded knowledge.” (*ibid*)

¹⁸ Hulme illustrates the delay between knowledge organisation on the one hand the actual state of knowledge at a given time commenting Dr. and Mrs. Charles Singer’s Hand-list of Scientific MSS — a ‘bibliographic’ control tool for manuscripts of the pre-printing age. On the other hand, he illustrates the growing specialization in literary classes and sub-classes using his own Tabular Surveys of the Division of Literature in Architecture and the Textile Industries, covering the 16th, 17th and 18th Centuries, and compiled from the Patent Office Library Subject Lists.

¹⁹ ‘What is wanted for our immediate purpose is a broad survey of the entire field of science with an analysis ...sufficient to show the relative growth of the principal subclasses in each section: but so far as I am aware, no such attempt has yet been made ... Bibliographical statistics, employed with the requisite qualification, are without question able to reveal the shape and period of such movements. *The International Catalogue of Scientific Literature* complies sufficiently with the two requirements of scope and standard of its entries’ (Hulme, 1923: 33)

²⁰ He assumed ‘on a priori grounds that increased activity in the literary output of a science can invariably be associated with pre-existent causes ...I do not, of course, suggest that the statistician of science should commence his studies by reading political speeches or by following the movements of the rubber market. His proper course is to analyse the statistics of science and tabulate the results in such a way as to indicate at what points in the line the pressure of scientific investigation is

Significance in Sociohistorical Investigations versus Productivity and the Decline of Serial (Externalist) History of Science.

I can distinguish at least two major reasons for an almost complete obliteration of serial externalism as historical genre in the history of science vary depending on whether one deals with the canonical (among sociologists of science and current bibliometrists) bibliography of the history of science or with the historiography of serial history. In the first case the major reason for downgrading serial externalism as an historical genre has to do with the problem of the assessment of historical significance in the history of science and its scientific use in sociological (socio-historical oriented) investigations. In the second case, there are at least three interrelated reasons that might explain the obliteration of serial history of science by quantitative historians. The first one has to do with the relative hegemony that economic and demographic history has had for quite a long time as the quantitative genre *par excellence* and the concomitant resistance of a great deal of historians towards quantification in history. Another reason for the obliteration of serial (externalist) history of science has to do with the lack of standardization of the data series used by the serial externalists (as showed above) in comparison with those used by economics and demography historians who took advantage of the organization of archives for historical purposes achieved in the 19th century (Furet, 1971). And finally, serial historians lacked the kind of mathematic and statistic foundation for the making of models of the kind provided by economics and demography for quantitative historians²¹.

Regarding the first source of oblivion, Sorokin and Merton's methodological contribution of 1935 offer a significant insight. This contribution is particularly interesting because it set an important standard in socio-historical investigations (historical sociology) as to methods and problems concerned. In doing so, they downgraded the serial (externalist) history of science as seriously limited. The problem that historical sociology faced at the time was that of the scientific use of historical data, and specifically in this contribution, the problem was how to deal scientifically with the assessment of historic significance. Historians of science (historians of the unique) were constantly compelled to make evaluative judgements in their treatment of intellectual history. This evaluative procedure is attested by historians' use of descriptive phrases such as, 'scientific genius', 'epoch-making discovery', 'the most original contribution', 'scientific advance (or decline)', used to express and summarise historians' judgements regarding the significance of an historical event in science. If such systematic procedure on which the assessment of historical significance rests can be regarded as scientific, Sorokin and Merton reasoned, "there can scarcely be any objection to, or necessarily any more subjectivism in, the systematic utilisation of these (historical) estimates as a basis for the organization of quantitative indices which would recapitulate the course of movement of a given culture or of a given social process" (Sorokin and Merton, 1935: 516).

By contrast to the historians of the unique whose interest was the assessment of what makes a scientific event unique; the historical sociologists were interested in the construction of a 'scale of translation' by means of which the 'event', distinguished as unique by the historian, can be rendered comparable across centuries or civilizations for the purposes of scientific (socio-historical) investigation²².

The difference, then, between event-history and the historical sociology resided neither in subject-matter nor in method, but in the objectives of the research and the consequent perspective that is taken in looking at the past. As Sorokin explained, the line of distinction between history and sociology is a rather subtle matter concerning the differentiation between history and sociology in terms of their aims as 'particularising' and 'generalising' sciences, respectively. Both the sociologist and the historian can take an interest in precisely the same series of phenomena: say the course of Arabian intellectual development. Both may bring essentially the same apparatus to bear on the study of the facts: say

being applied or relaxed. When, however, he comes to explain the causes governing the trend of scientific investigation, he will, I suggest, be led to admit that the course of science is profoundly influenced by agencies external to it, and that the initiative, as was found in my former lecture, proceeds from departments of human activity which are not at present recognized at Burlington House' (Hulme, 1923:32-33, 38-39).

²¹ I deal with the last two reasons in my second contribution for the Conference.

²² For a review of different methods of scaling using prosopographic and historical sources see Woods (1911).

Sarton's *Introduction*, or his report of 1923. But for the sociologist, the events that the history of the unique deal with constitute 'case studies' of a more general phenomenon that takes place in the 'interstitial space' of the special sociologies (Sorokin, 1931)²³. History serves solely as empirical evidence for a sociological mind seeking the formulation of general rules governing the relationships between science and a variety of other social phenomena such as, religious affiliation, for example (Sorokin, 1931; Merton, 1970[1938]).

This is why Sorokin and Merton considered that previous contributions attempting a similar aim²⁴—precisely those I have identified as serial (externalist) history of science—had failed. To the eyes of the sociologists, these studies failed as history in as much as they were unable to state the historical significance of scientific events. And they failed as science, because they were unable to provide an explanation of science as social phenomenon; they failed to explain what is the uniqueness of science (across centuries and among cultures) throughout historical 'case studies'; and by being unable to provide this explanation, they failed to set science apart from other social phenomena such as the church, the state, or the market²⁵. History deals solely with unique events and processes, but this uniqueness of science, for the sociologists, is not an historical problem anymore, for history deals only with events in time, not with general social structures. What makes science unique is the characteristic social structure in operation whether we look at it in seventeenth-century England or in twentieth-century Big Science.

Sorokin and Merton noticed that serial externalists use of quantitative indices assumed each datum—scientific discovery, scientific paper, scientific journal, patent—mentioned by chronologists, bibliographers or historians as having equal (historical) significance and, accordingly, assigned equivalent values to these items, they translate the unique into a unit of measurement. At this point the problem of index-construction arises: should each item be assigned an equal numerical value, or a value varying according to its significance? Should items be counted or should they be valued. A scale based on simple counting is entirely inappropriate in as much as it obscures differences in scientific achievement (between civilisations, historical periods, scientific discoveries, scientific publications and scientists). The 'heuristic fiction' of the equal significance of individual accomplishments offers a premium for the mere quantity of output and completely disregards qualitative differences and to this extent it is less adequate than a 'scale of translation' that assigns unequal values proportional to historical significance. It is not the relationship that a bibliographic entry holds with other entries as 'indifferent unit' that matters for the sociologists, neither how they were classified as acceptable units within a bibliographic list (that's the business of the historian of the unique). The whole point of the 'translation' is to guide socio-historical investigations by preserving and making visible the historical significance as stated by historians for the purposes of comparison. It is the equivalence between events in terms of historical significance (as historically equivalent 'case studies') established by means of the scale, that renders its cross-temporal comparison meaningful in sociological terms. The utility of Sorokin's and Merton's 'scales of translation' lies in how they guide the selection of comparable cases required for a scientific (sociological) study. For example, in his proposal of a scale from one to three, "a value of 3 was assigned to those individuals who are described as having 'had a very important influence,' or as being 'one of the greatest...' or as having 'done very important work in

²³ To be sure, the sociology of science as a sociological speciality is not quite present yet in Sorokin's 'map of the discipline' sketched in 1931, however, there 'must be' one, *master dixit*.

²⁴ At this point Sorokin and Merton referenced the studies of Cole & Eales (1917), Weinberg (1926), Rainoff (1929), Ogburn & GilFillan (1933). More than forty years later Merton (1977) added Hulme's lectures on statistical bibliography and the growth of civilization. It is noteworthy that in this lapse of time this early literature is introduced without criticism. Whereas in 1935 the imputed work on scales that this early contributions were supposedly proposing was considered as inadequate, in 1977 the whole pack is presented both as methodological antecedents in the emergence of the sociology of science which 'failed to converge into an intellectual tradition'. Moreover, forty years hindsight allows Merton to explain this failure as a result, on the one hand, of the resistance to this approach on the side of the 'established scholars in the then discernible field of the history of science' and the absence of an institutional and cognitive identity for the still to come sociology of science. On the other hand, the failure is also attributed to the absence of specialized journals, the small numbers of papers devoted to the subject and the consequent dispersion of the subject throughout a scarce overlapping readership.

²⁵ An explanation that, as we all know, was based on that complex of norms and values whose formulation was the result of the socio-historical investigations.

the field,' etc." Thus those who are ranked as three in the scale are all comparable cases, those in the rank of two are equally comparable (Sorokin and Merton, 1935: 518)²⁶. But these scales were not, properly speaking, a metric, they simply enable analysts to account for the selection of significant events for sociological investigation.

The methodological tension between significance and productivity has been carried on within the sociology of science. After all, as Sorokin and Merton admit towards the conclusion of their methodological paper, the plausibility of a scale of translation in which every item is given equal significance indicates something nevertheless; it is a 'rough index of activity'. As such, it was used in Merton's doctoral dissertation (Merton, 1938: Chap. 3) but there, again, what made the use of 'statistical bibliography' suitable was the nature of the inquiry. After having defined in the first place scientific elite in 17th-century England using a prosopography of eminent scientists²⁷, in chapters two and three of his doctoral work, he was interested in a further characterisation of this elite by studying changes in its focus of intellectual interests; statistical bibliography served this purpose. Merton, however, did not use scientific productivity for assessing the eminence of the elite or the historical significance of its contributions. Since his aim was just 'to ascertain the relative degrees of scientific activity reflected in output', there was no need for establishing a precise one-to-one correspondence between individual discoveries and the 'units' of the tabulation (bibliographic entries taken from Prof. Darmstädter's popular handbook). And although Merton talked about 'cycles' and 'fluctuations' the important thing about them wasn't the size of these waves or its periodicity, but that they simply reflected shifts in the intellectual focus of the scientific elite. Hence Merton's aim was not to make a mathematical generalization out of these trends. Rather the explanation of these fluctuations was linked to 'some extra-scientific elements that strongly influenced, if they did not determine, the centring of scientific attention upon certain fields of investigation' (Merton, 1970/1938: 54). These extra-scientific elements (Merton's externalism) came to be found, on the one hand, in the sphere of values and interest: the hypothesis of the relationship between Puritanism and the emergence of science in chapter six of Merton's dissertation. And, on the other hand, they were detected in economic and demographic pressures.

This very same conflict between productivity and significance was dealt with thirty years later in the well-known controversy between Crane (1965) and the Cole brothers (1967) regarding the measurement of scientific achievement²⁸. Recognised by then as a necessary though limited measurement, scientific output constituted the base line for the empirical research on the reward system of science²⁹. However, the major problem then was to find a way to measure scientific significance in sociological rather than historical terms. As Gilbert pointed out, the historical investigations advanced by pre-bibliometric sociology of science were seriously compromised by systematic dependence on historical accounts and the subjectivity of historians' assessments. This situation helped little with the differentiation of the sociology of science as a scientific speciality at Columbia University³⁰, and it is in this context that Merton (1977) dealt with Price's and Garfield's

²⁶ In a second scale Sorokin and Merton proposed a weight of 15 'points' to be assigned "to each individual who was designated as, say, 'one of the greatest physicians of all times'. The individual who was held to be 'one of the greatest scientists of the Middle Ages' was deemed worthy of a weight of ten points: 'one of the greatest of Islam' was given seven points; 'one of the greatest of his time' was given a weight varying from four to seven, depending upon the general level of the intellectual contributions of the period; [...] and lastly, those who were simply mentioned were weighted with one point" (Sorokin and Merton, 1935: 518)

²⁷ Using the *Dictionary of National Biographies*, a bio-bibliographical source. For further distinctions between prosopography of elites and of masses see, Stone (1971).

²⁸ Bear in mind that by then, the first generation of sociologists of science was striving to set the appropriated 'scientific' standard expected at Columbia which clearly distinguished between the old-fashioned socio-historical research and the new empirical sociology (See, Camin and Xie, 1994).

²⁹ "If we are to discover the social conditions most favourable for the advance of science, we must be able to measure scientific output." ((Cole, 2000): 287).

³⁰ As Camic and Xie pointed out, by mid 1960s the days of socio-historical investigation at Columbia were already over, and the age of empirico-quantitative research in the social sciences was 'hot' stuff. For this reason the use of citation as a measurement of quality in the methodological paper of the Cole brothers (1967) and the possibility of making this measurement upon large samples brought by the introduction of the SCI where functional for the institutionalization of the sociology of science at Columbia.

contributions as imported techniques contributing to the cognitive identity of the empirical programme of the sociology of science, *master dixit*.

The Historiographic Problem of the Standard Sources: and the fall into historical oblivion of serial externalists history of science.

Concerning the second source of oblivion, the development of serial history in general and the resistance that for some time this historical genre generated among historians is something yet unexplored in relation to early contributions of serial externalist historians of science. It is not this paper aim to render an account of this debate. However, it is useful to note that this resistance towards quantification in history was grounded in the perception that quantitative historians were unduly reducing the field of historical research to economics or demography exclusively. Economic historians, the early champions of serial history, attempted to turn history into a kind of retrospective econometrics (scientific history of economics). On the basis of modern national accounting they attempted to fill in all the columns of an input-output table for past centuries. To economic historians total and systematic quantification were indispensable both for the elimination of arbitrariness in selecting historical data and for the use of mathematical models to process it. According to this view, genuine quantitative history would be the result of a two-fold reduction of history: first, the reduction of its scope to the field to economics alone, and secondly the reduction of its descriptive and interpretative system to the one worked out by the paradigm of scientificity in the social sciences that economics represented (Furet, 1971; Marczewski, 1968). The same holds for demography and demographic history. Thus, a truly scientific history was conceived as that which can be dealt with in quantitative terms, and one able to connect present and past via an homogeneous data series (Derosières, 1998; Furet, 1971). Of course, there have to be data for the past just as for the present; or at least it has to be possible to work them out with a sufficient degree of accuracy, or to reconstruct or extrapolate them. This circumstance sets the limits to the complete quantification of economic or demographic history. Complete quantification, even if possible at all before the 19th century, could not go back beyond the introduction of the statistical or proto-statistical recording of data, which coincides with the centralisation of the great European monarchies. Some factors thus assisted anti-quantitative historians in their claim that history did not begin with the English political arithmetic or the German *Statistik* (Derosières, 1998; Furet, 1971).

However, the reason why I mention this debate has to do precisely with the problem of the sources for the construction of standardized data series and the mathematical modelling of this series; both crucial requirements highlighted by economic history. These two issues will help us to understand another source of oblivion of serial (externalist) history of science. The early contributions of serial historians of science were somehow in the middle of this debate concerning quantification in history. Serial history of science was downgraded by historians of the scientific event and neglected by quantitative (economic and demographic) historians. Neglected by quantitative historians, because the early contributions of serial history of science did not meet the standards that historical evidence should have in the eyes of early twentieth-century economic and demographic historians. On the other hand, historians of the unique rejected it because serial history represented a fundamentally different understanding of the historical facts that was completely blind to the event and its uniqueness³¹.

On these grounds, the development of serial history of science could be seen as an increasing effort to construct standardized data series, on the one hand, and to develop quantitative models able to account for the issue of ‘significance’ based on an analysis of the ‘internal’ structure of the data.

Concluding Remarks: From externalism to internalism in serial history of science.

Serial externalist historians laid down the foundations of a new form of reading discontinuous elements as constituting a continuous entity. In the particular case of science, serial historians were the first to propose the reading of a discontinuous list of bibliographic entries as a continuous thing, as

³¹ Several times Price (1978), and Merton (1977) have touch upon the ‘cold’ reception among historians of Price’s 1951 first contribution to the 6th International Congress of the History of Science (Amsterdam) in which he proposed some measurements of the development of science.

science. The crucial translation of a singular and unique (eventual) thing that a bibliographic entry might represent, into a unit of measurement; of a bibliographic list, into a quantitative data series, was the fundamental contribution of serial historians. They constructed scientific ‘literature’ as a magnitude called science (Price, 1965). In this regard, continuous technical innovation in bibliographic control of periodical literature (from substantive periodical to any form of secondary literature) have both contributed a great deal to the consolidation of serial (internalist) history in times of Price and Garfiel and the rest of the band.

Serial historians from all times have devoted themselves to defining quantitative patterns of scientific growth, productivity and consumption once they were in possession of this understanding of science as a magnitude. In this regard, the “heroic age” of bibliometrics does not differ from the “classic age” inaugurated by Derek Price and Eugine Garfield; both were serial history of science. The difference between internalist and externalist serial history of science rest on the one hand, in the explanation provided for these patterns, and on the other, in the kind of models used for representing them. Whereas externalist history understood scientific growth as the result of changes in the realm of economy, politics, and culture, internalist history understood scientific growth as resulting from science’s demography and patterns of consumption among scientists. Whereas externalist historians found their models in economic history, internalist historians found theirs in the internal structure of the data statistically defined. In these two respects the very classic of Dereck Price *Big Science Little Science* can be better appreciated not as the first bibliometric work properly speaking, but as the culmination of long term effort of serial externalist historians. The significance of Price for the serial history of science, of course, does not stops short within the confines of a serial externalist history of science. Price deserves also the credit for being the first “internalist” among serial historians of science in as much as he integrated under a single historical model of growth (logistic model), Lotka’s and Bradford’s previous results. This allowed him to attempt a different kind of explanation for scientific growth, an explanation that focus on the internal structure and dynamics of the literature rather than on external socio-economic processes. Garfield’s efforts were from the very begging oriented in the same vain. Papers' standards for our XI Conference does not leave room for presenting this second strain of serial history of science in some detail. However, these second period is familiar for bibliometrics practitioners and historians of science who has explore it in many important and well-known works.

References

- Aydelotte, W. O. (1966). Quantification in History. *The American Historical Review*, 71, 803- 825.
- Bowker, G. & Star, S. (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge MA: The MIT Press.
- Camic, C. & Xie, Y. (1994). The Statistical Turn in American Sociology of Science: Columbia University, 1890 to 1915. *American Sociological Review*, 59, 773-805.
- Cole, F. J. & Eales, N. B. (1917). The History of Comparative Anatomy: A Statistical Analysis of the Literature. *Science Progress*, 11, 578-596.
- Cole, J. R. (2000). A Short History of the Use of Citations as a Measure of the Impact of Scientific and Scholarly Work. In B.Cronin & H. Atkins (Eds.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 281-296). Medford N.J.: Information Today.
- Cole, S. & Cole, J. R. (1967). Scientific Output and Recognition: A Study in the Operation of the Reward System in Science. *American Sociological Review*, 32, 377-390.
- Crane, D. (1965). Scientists at Major and Minor Universities: A Study of Productivity and Recognition. *American Sociological Review*, 30, 699-714.
- Daston, L. (2001). History of Science. *History of Science*, 6842-6848.
- Desrosières, A. (1998). *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, Mass.: Harvard University Press.
- Edge, D. (1980). Why I Am Not a Co-Citationist. In *Essays of an Information Scientist* (1977-1978) (pp. 240- 246). Philadelphia: ISI Press.
- Erickson, C. (1975). Quantitative History. *The American Historical Review*, 80, 351-365.
- Fogel, R. W. (1975). The Limits of Quantitative Methods in History. *The American Historical Review*, 80, 329- 350.
- Furet, F. (1971). Quantitative History. *Daedalus*, 100, 151-167.

- Gieryn, Thomas F. (1983) Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review*, 48, 781-795.
- Gross, P. L. K. (1927). Fundamental Science and War. *Science*, 66, 640-645.
- Guston, David H. (1999). Stabilizing the Boundary between US Politics and Science: The Role of the Office of Technology Transfer as a Boundary Organization. *Social Studies of Science*, 29, 87-111.
- Guston, David H. (2001) Boundary Organizations in Environmental Policy and Science: An Introduction. *Science, Technology, & Human Values*, 26, 399-408.
- Hulme, E. W. (1923). *Statistical Bibliography in Relation to the Growth of Modern Civilization*. London: Butler & Tanner.
- Jasanoff, Sheila (1996) Beyond Epistemology: Relativism and Engagement in the Politics of Science. *Social Studies of Science* 26, 393-418.
- Joynt, C. B. & Rescher, N. (1961). The Problem of Uniqueness in History. *History and Theory*, 1, 150-162.
- Lehfeldt, R. A. (1916). The Normal Law of Progress. *Journal of the Royal Statistical Society*, 79, 329-332.
- Marczewski, J. (1968). Quantitative History. *Journal of Contemporary History*, 3, 179-191.
- Meadows, P. (1944). The Scientific Use of Historical Data. *Philosophy of Science*, 11, 53-58.
- Merton, R. K. (1970[1938]) Science, Technology and Society in Seventeenth Century England. New York, etc.: Harper & Row
- Merton, R. K. (1977). The Sociology of Science: An Episodic Memoir. In R.K.Merton & J. Gaston (Eds.), *The Sociology of Science in Europe* (pp. 3-141). Carbondale-London: Southern Illinois University Press Feffer and Simons.
- Ogburn, W. F. & GilFillan, S. C. (1933). The Influence of Discovery and Invention. In *Recent Social Trends in the United States* (pp. 122-166). New York, London.
- Ossowska, M. & Ossowski, S. (1936). The Science of Science. *Organon*, 1, 1-11.
- Price, D. J. d. S. (1964). The Science of Science. In M.Goldsmith & A. Mackay (Eds.), *The Science of Science. Society in the Technological Age* (pp. 195-208). London: Souvenir Press.Rainoff, T. J. (1929). Wave-like Fluctuations of Creative Productivity in the Development of West-European Physics in the Eighteenth and Nineteenth Centuries. *Isis*, 12, 287-319.
- Price, D. J. d. S. (1965). *Little Science, Big Science*. New York: Columbia University Press.
- Price, J. M. (1969). Recent Quantitative Work in History: A Survey of the Main Trends. *History and Theory*, 9, 1-13.
- Pritchard, A. (1981) *Bibliometrics: A Bibliography and Index*. London: ALLM Books.
- Rainoff, T. J. (1929). Wave-like Fluctuations of Creative Productivity in the Development of West-European Physics in the Eighteenth and Nineteenth Centuries. *Isis*, 12, 287-319.
- Rescher, N. (1978). *Scientific Progress. A Philosophical Essay on the Economics of Research in Natural Science*. Oxford: Basil Blackwell.
- Shapin, S. (1982). History of Science and its Sociological Reconstructions. *History of Science*, 157- 211.
- Shapin, S. (1992). Discipline and Bounding: The History and Sociology of Science as Seen Through the Externalism-Internalism Debate. *History of Science*, 30, 333-369.
- Sorokin, P. A. (1927). A Survey of the Cyclical Conceptions of Social and Historical Process. *Social Forces*, 6, 28-40.
- Sorokin, P. A. (1931). Sociology as a Sience. *Social Forces*, 10, 21-27.
- Sorokin, P. A. & Merton, R. K. (1935). The Course of Arabian Intellectual Development, 700-1300 A.D. A Study in Method. *Isis*, 22, 516-524.
- Star, S. L. and Griesemer, J. R. (1989). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science* 19, 387-420.
- Swierenga, R. P. (1974). Computers and American History: The Impact of the 'New' Generation. *Journal of the American History*, 60, 1045-1070.
- Thackray, A. (1978). Measurement in the Historiography of Science. In Y.Elkana, J. Lederberg, R. K. Merton, A. Thackray, & H. A. Zuckerman (Eds.), *Toward a Metric of Science: The Advent of Science Indicators* (pp. 11-30). New York etc: John Wiley & Sons.
- Thackray, A. & Merton, R. K. (1972). On Discipline Building: The Paradoxes of George Sarton. *Isis*, 63, 473-495.
- Vessenyi, P. E. (1974). *An Introduction to Periodical Bibliography*. Ann Arbor (Mich): Pierian Press
- Weinberg, B. (1925). Sur les lois D'évolution de la Pensée Humanine. *Revue Générale des Sciences*, 36, 565-569.
- Weinberg, B. (1926). Les lois D'évolution des découvertes de L'humanité. *Revue Générale des Sciences*, 37, 43-47.

- Wilson, P. W. & Fred, E. B. (1935). The Growth Curve of a Scientific Literature. *The Scientific Monthly*, 41, 240-250.
- Woods, F. A. (1911). Historiometry as an Exact Science. *Science*, 33, 568-574.

The paper with pages 360-364 has been retracted from the conference shortly before the preparation of the final version of these proceeding. Therefore pages 361-364 in this volume are missing.

Self-citation Networks as Traces of Scientific Careers^{1,2}

Iina Hellsten*, Renaud Lambiotte**, Andrea Scharnhorst* and Marcel Ausloos**

**iina.hellsten@vks.knaw.nl*

The Virtual Knowledge Studio for the Humanities and Social Sciences at the Royal Netherlands Academy of Arts and Sciences, VKS-KNAW, Cruquiusweg 31, 1019 AT Amsterdam (The Netherlands)

***renaud.lambiotte@ulg.ac.be*

SUPRATECS, Université de Liège, B5 Sart-Tilman, 4000 Liège (Belgium)

Keywords

self-citation; network; field mobility; co-authorship; keywords.

Extended Abstract

This paper introduces a new approach to detecting scientists' field mobility by focusing on an author's self-citation network, and the co-authorships and keywords in self-citing articles. Contrary to much previous literature on self-citations, we will show that author's self-citation patterns reveal important information on the development and emergence of new research topics over time. More specifically, we will discuss self-citations as a means to detect scientists' field mobility. We introduce a network based definition of field mobility, using the Optimal Percolation Method (Lambiotte & Ausloos, 2005; 2006). The results of the study can be extended to self-citation networks of groups of authors and, generally also for other types of networks.

In much of the literature in citation analysis, author's self-citations are excluded as 'noise' or they are treated as a bias for the analysis (e.g. MacRoberts & MacRoberts, 1988; Leydesdorff & Amsterdamska, 1990; Persson & Beckermann, 1995; van Raan, 2006). This approach is often linked to the use of citation analysis for science policy purposes, i.e. as a way to measure the impact of journals, authors or whole university departments (e.g., Moed, 2005).

Recently, research on citation, co-citation and co-authorship networks has gained interest also in information sciences, in particular, mapping knowledge domains (Chen, 2003; Börner et al., 2003; Boyak et al., 2005), and in statistical physics (Newman, 2001; Barabási et al., 2002; Redner, 2005; Börner et al., 2005). Scientific development has been visualized as sequence of temporally evolving graphs (Burger & Budjosó, 1985). The accumulation of published articles enables also drawing evolutionary tree-like structures of referencing over time. One famous example is the idea of a historiograph proposed by Garfield (Garfield, 1973; Garfield et al., 2003; Garfield, 2004).

Field mobility, or field migration (Vlachy, 1981), is defined as scientists moving into new research topics. Field mobility can be measured by identifying different research topics (fields), allocating the activity of scientists in these fields, and following the activity of scientists over time to mark the transitions. Field mobility has been investigated already since the 1980s (Le Pair, 1980, Van Houten et al., 1983, Hargens, 1986). Field mobility has been discussed as the driving force for the exploration of new territories in the 'landscape' of science (Urban, 1982; Scharnhorst, 2001). More specifically, field mobility has been modeled as an exchange mechanism between research fields leading to a co-

¹. This paper is an outcome of the Critical Events in Evolving Networks (CREEN) project, funded by the EU under its 6th Framework, NEST-2003-Path-1, 012684. We are grateful for Loet Leydesdorff, Paul Wouters and Sally Wyatt for their comments on earlier versions of this paper.

² The full text of the paper is accepted for Scientometrics (I. Hellsten, R. Lambiotte, A. Scharnhorst, M. Ausloos (2007): Self-citations, co-authorships and keywords: A new method for detecting scientists' field mobility? *Scientometrics* (forthcoming)). A pilot study was published as I. Hellsten, R. Lambiotte, A. Scharnhorst, M. Ausloos (2006): A journey through the landscape of physics and beyond - the self-citation patterns of Werner Ebeling. In: *Irreversible Prozesse und Selbstorganisation*. Ed. by T. Pöschel, H. Malchow, and L. Schimansky-Geier, Logos Verlag Berlin, pp. 375-384.

evolution or coupled growth of scientific specialties (Ebeling & Scharnhorst, 1986; Bruckner et al., 1990). So far, however, there have not been systematic, empirical studies on the role of self-citations for detecting field mobility in scientometrics.

Changing patterns of scientific activity have been also discussed in the context of inter-disciplinarity. Attempts to measure inter-disciplinarity rely on citation and publication patterns (see e.g. Rinia et al., 2002; Morillo et al., 2003). However, some studies also follow certain authors through their publication records (Urata, 1990; Pierce, 1999). Some studies use interviews and surveys to trace academic careers but this approach is restricted to rather small case studies (van Houten et al., 1983; Wagner-Doebler & Berg, 1993). Career moves of scientists are also a topic of science historical or sociological research (see, e.g. for an earlier research Gilbert, 1977). Currently, there are no automated techniques for quantitatively measuring scientists' field mobility. In this paper, we will present such a technique by focusing on the evolving self-citation networks of an author. This way we trace academic development of a scientist with hindsight via her/his use of citations to her/his own work.

Our theoretical focus builds more upon science studies tradition than science policy approach within scientometrics (Wouters, 1999). The paper contributes to science and technology studies by presenting a new quantitative measure for scientists' field mobility. Methodologically our study takes part in the recent fusion between social sciences and complex network theory (e.g. Watts, 2004). We will use physics methods from complex network theory to make visible scientists' field mobility using their self-citations. As an additional level in our analysis we use co-authorships and the keywords in the ISI database of these self-citing articles. In particular, we rely on a new model which links collaboration patterns with the diffusion of ideas between research fields and the mobility of scientists (Lambiotte & Ausloos, 2006).

As our aim is to present theoretically and methodologically new way of approaching field mobility via self-citations and not to give a comprehensive review or analysis on self-citations, we have initially focused on the self-citation network of one scientist: Professor Werner Ebeling. We use his publication record to introduce our method, but we believe that the proposed tool can be applied to other individual bibliographies or the bibliographies of groups of scientists.

References

- Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. & Vicsek, T. (2002). Evolution of the social network of scientific collaborations, *Physica A*, 311(3-4), 590-614.
- Bruckner, E., Ebeling, W. & Scharnhorst, A. (1990). The Application of evolution models in scientometrics, *Scientometrics*, 18 (1-2), 21-41.
- Burger, M. & Budjosó, E. (1985). Oscillating chemical reactions as an example of the development of science. In Field, R. & Burger, M. (Eds.), *Oscillations and traveling waves in chemical systems* (pp. 565-604). New York: Wiley.
- Börner, K., Chen, C.M. & Boyack, K.W. (2003), Visualizing knowledge domains, *Annual review of information science and technology*, 37, 179-255.
- Börner, K., Dall'Asta, L., Ke, W. & Vespignani, V. (2005), Studying the emerging global brain: analyzing and visualizing the impact of co-authorship teams, *Complexity*, 10(4), 57-67.
- Boyack, K.W., Klavans, R. & Börner, K (2005), Mapping the backbone of science, *Scientometrics*, 64(3), 351-374.
- Chen, C. M. (2003), *Mapping scientific frontiers: The quest for knowledge visualization*. Berlin et al.: Springer
- Ebeling, W. & Scharnhorst, A. (1986), Selforganization models for field mobility of physicists, *Czechoslovak Journal of Physics* B36, 43-46.
- Garfield, E. (1973), Historiographs, librarianship, and the history of science. In: Rawski, C.H. (Ed), *Toward a theory of librarianship: papers in honor of Jesse Hauk Shera* (p. 380-402). Metuchen, N.J.: Sacrecrow press. Reprinted in: Garfield, E. (1974-1976), *Essays of an Information Scientist*, Vol. 2, pp. 136-150.
- Garfield, E., Pudovkin, A.I. & Istornin, V.S. (2003), Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology*, 54(5), 400-412.
- Garfield, E. (2004), Historiographic mapping of knowledge domains literature, *Journal of Information Science*, 30(2), 119-145.
- Gilbert, G. N. (1977), Competition, differentiation and careers in science, *Social Science Information*, 16(1), 103-123.

- Hargens, L.L. (1986), Migration patterns of U.S. Ph.D.s among disciplines and specialties, *Scientometrics*, 9(3-4): 145-164.
- Houten, J., van Vuren van, H.G., Le Pair, C. & Dijkhuis, G. (1983), Migration of physicists to other academic disciplines: situation in the Netherlands, *Scientometrics*, 5(4):257-267.
- Lambiotte, R. & Ausloos, M. (2005), Uncovering collective listening habits and music genres in bipartite networks, *Physical Review E*, 72(6), 066107.
- Lambiotte, R. & Ausloos M. (2006), On the genre-fication of music: a percolation approach, *The European Physical Journal B*, 50(1-2), 183-189.
- Le Pair, C. (1980), Switching between academic disciplines in universities in the Netherlands, *Scientometrics*, 2(3), 177-191.
- Leydesdorff, L. & Amsterdamska, O. (1990), Dimensions of citation analysis, *Science, Technology and Human Values*, 15(3), 305-335.
- MacRoberts, M. & MacRoberts, B. (1988), Problems of citation analysis: A critical review, *Journal of the American Society for Information Science*, 40(5), 342 – 349.
- Morillo, F., Bordons, M. & Gomez, I. (2003), Interdisciplinarity in science: a tentative typology of disciplines and research areas, *Journal of the American Society for Information Science and Technology*, 54(13),1237-1249.
- Newman, M.E.J. (2001), The structure of scientific collaboration networks, *Proceedings of the National Academy of Sciences*, 98(2), 404-409.
- Persson, O. & Beckmann, M. (1995), Locating the network of interacting authors in scientific specialties, *Scientometrics*, 33(3), 351-366.
- Pierce, S.J. (1999), Boundary crossing in research literatures as a means of interdisciplinary information transfer, *Journal of the American Society for Information Science*, 50(2), 271-279.
- Raan, A. F. J. van (2006), Performance-related differences of bibliometric statistical properties of research groups: cumulative advantages and hierarchically layered networks, *Journal of the American Society for Information Science and Technology*, 57(14), 1919-1935.
- Redner, S. (2005), Citation statistics from 110 years of physical review, *Physics Today*, 58(6),49.
- Rinia, E.J., van Leeuwen, T.N., Bruins, E.E.W., van Vuren, H.G., van Raan, A.F.J. (2002), Measuring knowledge transfer between fields of science, *Scientometrics*, 54(3): 347-362.
- Scharnhorst, A. (2001), Constructing knowledge landscapes within the framework of geometrically oriented evolutionary theories. In: Matthies, M., Malchow, H. & Kriz, J. (Eds) *Integrative Systems Approaches to Natural and Social Dynamics* (pp. 505-515. Berlin et al.: Springer.
- Urata, H. (1990), Information flows among academic disciplines in Japan, *Scientometrics*, 18(3-4), 309-319.
- Urban, D. (1982), Mobility and the growth of science, *Social Studies of Science*, 12(3), 409-433.
- Vlachy, J. (1981), Mobility in physics - a bibliography of occupational, geographic and field mobility of physicists. *Czechoslovak Journal of Physics*, B31(6), 669-674
- Wagner-Doebler, R., Berg J. (1993), Mathematische Logik von 1847 bis zur Gegenwart. Berlin, New York: Walter de Gruyter.
- Watts, D. (2004), The “new” science of networks. *Annu. Rev. Sociol.*, 30, 243-70.
- Wouters, P. (1999), *The Citation Culture*. University of Amsterdam, Faculty of Science. Unpublished PhD Thesis.

Pre 1990 French Doctoral Dissertations in Philosophy: A Bibliometric Profile of a Canonical Discipline

Jean-Pierre V. M. Hérubel

jpvmh@Purdue.edu
HSSE Library, Purdue University, West Lafayette, In. (USA)

Abstract

Established scholars and graduate students researching French doctoral dissertations prior to the last ten years often find that there are deferent doctorates for the same discipline appearing in the grey literature. This exploratory bibliometric study attempts to map doctorates granted in one canonical discipline; i.e. philosophy. Since philosophy transcends *both the social sciences and the humanities* in French academia, doctorates in philosophy were examined for their doctoral type and institutional affiliation. Major and intellectually significant patterns of geographical and institutional dispersion are revealed.

Keywords

bibliometric; doctorates; humanities; philosophy; geographical dispersion

Introduction

Researchers interested in pursuing research literature in a discipline often resort to gleaning necessary and pertinent research appearing in doctoral dissertations. Often they are looking into issues of subject, methodological approaches, theoretical perspectives, or gleaning the often appended bibliographies of consulted scholarship by the dissertation writer. Dissertation advisors will naturally point a doctoral student in the direction of approved dissertations in a respective field of endeavor, reducing the student's efforts in establishing a research project of their own. As doctoral education and dissertations represent advanced work in disciplines, they often represent the most current research in a given field. Conforming to the protocols of disciplinary formation and acculturation, doctoral dissertations reflect the practices accepted by the scholarly professions in which those dissertations are written and defended.

Doctoral education was first established during the European medieval period when universities were first founded (de Ridder-Symoens, 1992; Rothblatt & Wittrock, 1993). Research as now construed follows a less professional program as evolved in medieval universities. Doctoral education as we know it today became formalized in the 19th century in Europe and has evolved into its present formulation. Indeed, the doctorate is often the highest expression of formalized education in the world. Although prone to piecemeal changes and purpose, the doctorate still maintains its prominence as the highest degree within higher education (Hesseling, 1986). Research-oriented, the doctorate, especially as capstone experience for advanced studies and research in disciplines, mirrors the nature of disciplines in the sciences, humanities, social sciences, and technologies (Kouptsov, 1994; Schweitzer, 1965; Bowen, 1992). As guarantor of acceptable training and acculturation for disciplinarians, the doctoral experience and dissertation constitute the salient intellectual characteristics of those respective disciplines.

Doctoral Structure to and After 1984

For English-speaking scholars in the humanities or the social sciences, researching French doctoral dissertations presents a bewildering situation (Rutledge, 1994). Prior to governmental reforms of 1984 that effectively changed the doctoral system from three possible dissertation types to one, researchers were confronted with three distinct and separate degree titles (Wanner, 1975; Assefa, 1988). The *doctorat d'université*, *doctorat de 3ième cycle*, and the *doctorat d'état* represented different qualities and purposes for the student as well as the professor. Each followed different protocols for admission, time in course of studies, and type and purpose of dissertation (Debeauvais, 1986):

-*Doctorat d'Etat*--Established in 1810, representing the highest research contribution in all fields; 10-20 years of preparation ending in massive *chef d'oeuvre*; necessary for full professorship; most prestigious doctorate. Prior to 1970, required one primary and one secondary dissertation; essentially, two dissertations.

-*Doctorat d'Université*--Established in 1897 in all fields, for both French & foreign students [primary candidate population]; most diverse in quality and duration of studies; did not confer any professional status in France; ambiguous and least prestigious; dissertation range & strength is only guarantor of quality.

-*Doctorat de 3ième cycle*--Established in 1954 for sciences & in 1958 for humanities, social sciences; three year long research technique oriented, with dissertation in very narrow subject within a discipline; indication of pursuing research techniques.; qualifies for beginning research or university post. The *doctorat de 3ième cycle* after 1974 (per a change in regulation governing this doctorate) was referred to as *3ième cycle*; dissertations registered prior to 1974 and finished after 1974 were referred to as *doctorat de 3ième cycle ancien régime*.

- *Doctorat Nouveau Régime*--Established in 1984 for all disciplines; 3-5 years in duration; equivalent to and modeled on the American Ph.D.

- *Habilitation à diriger les recherches*--Although not strictly identified as a doctorate per se, it was established in 1984; no time constraint; open to holders of the *doctorat d'Etat* or *doctorat nouveau régime*; higher order representing major critical accomplishment in field; by dissertation, or collection of articles, etc; required for full professorial status.

It should be noted that these dissertations reflect research under sustained direction from a director responsible for the student's progress and initiation into the research ethos of the discipline in question. The cultural environment in which French doctoral studies evolved is one that is tied to governmental decrees and administration. Under French national ministries of education, universities offered doctoral degree studies according to the needs and requirements set forth by faculties and professorial interests. Overtime, especially after World War Two French higher education visionaries were keenly interested in expanding doctoral opportunities to students who would assume positions in research primarily and teaching secondarily. The growing economy demanded researchers in science, inaugurating the *doctorat de 3ième cycle* in 1954. The numbers of such doctorates grew to accommodate the necessities of all sectors of French life. In 1958, the humanities and the social sciences received the right to train students for the *doctorat de 3ième cycle*, bringing normalization to doctoral education for this particular diploma. Unlike the *doctorat de 3ième cycle*, the *doctorat d'état* not only represented the pinnacle of serious research achievement, it truly represented the highest levels of research leading to genuine original contributions to respective disciplines.

Among the pantheon of doctoral studies, the *doctorat de l'université*, after careful deliberation, was established in 1897 in all fields, for both French and foreign students. Wishing to attract and expand the foreign student population, French officials supported a doctoral degree that was accessible to those students who were not educated in France. Although far from uniform and diverse in quality and duration of studies, this degree represented a level of advanced graduate achievement. Under French law such degree holders could not exercise professional largess in French professions. Less prestigious, but still indicative of sustained research activity and training, the *doctorat de l'université* represented a unique and vital doctoral degree to both French students who desired a doctorate and those foreign students who wished to pursue studies in France.

Reformation of doctoral studies assumed a critical mass when in the pivotal year of 1984 French higher education officials put in place a newly sanctioned doctoral regime. Firstly, the old tripartite regime was completing and irrevocably abolished. A new doctoral degree was set in its place--the *doctorat nouveau régime*. This degree effectively assumed the purpose of the previous doctoral degree structure and practice, privileging uniformity. Without devaluation of the previous doctorates, the *doctorat nouveau régime* would continue the best of practice represented by the previous regimes while aligning French doctoral education with the more Anglo-Saxon counterparts. However, a caveat was administratively set in place by the appearance of a higher-order diploma analogous to the German degree of *habilitation*. The French version harkened to the glorious past represented by the

august *doctorat d'état*. Without qualification, those holders of the *doctorat nouveau régime* (currently known as the *doctorat*), could enroll for the *habilitation à diriger les recherches* (Mousnier, 1965; Duroselle, 1967; Duroselle, 1970). Only with this doctorate could a researcher assume a permanent position as professor and direct doctoral students (Slone, 1990).

The above typology of degrees offers the non-French researcher a better understanding of degree structure and what that structure represents intellectually. The trifurcated doctoral structure can be better understood when examined through individual disciplines and their respective specializations. As centralized as French academic culture is, various universities granted doctoral degrees according to their respective strengths. To gain a clearer sense of this an introductory bibliometric mapping of doctoral degrees for specific disciplines reveals patterns specific to different universities. For the researcher wishing to consult French doctoral research, an appreciation for the types of doctorates encountered would be instrumental in understanding the unique structure of this grey literature.

Purpose of Study and French Academic Philosophy

As an academic discipline philosophy has a long and illustrious heritage in French culture, society, and especially in French academia. Dissertations have been defended during the medieval period until today, making philosophy a stellar discipline within the French academic landscape. Since the 19th c. philosophy has enjoyed longevity within the French secondary and higher education systems (Fabiani, 1988). Among French academic disciplines, philosophy stands with those disciplines considered canonical, i.e. French literature, mathematics, Greek and Latin. Philosophy represents a discipline that exhibits strong cultural currency, if not intellectual reputation for scholarly rigor and accomplishment (Bourdieu, 1988). As academic disciplines arise from processes of intellectual and social acculturation, it is critical for the international researcher to appreciate, if not understand the significance accorded philosophy in French academia (Chimisso, 2000; Chimisso, 2005). Within French academic philosophy lie the various sub-disciplines of logic, philosophy of science, philosophy of mind, epistemology, as well as history of philosophy and aesthetics, etc. An additional characteristic is that it is split along an axis demarcating the canon from the non-canonical. Dissertations can be written on acceptable subjects, i.e. Plato and Aristotle, or on a historical problem in philosophy. A particular problem attendant to the ancient Greco-Roman period, or medieval period, or Kant, or Descartes can be broached. Rarely, can contemporary topics, reflecting current interests be pursued, unless under the careful administration of an advisor. The canon is strict in its prescriptions and orientation while multidisciplinary and more unorthodox subjects and approaches can be researched as well, but are considered less indicative of philosophical acculturation and professionalization.

This study represents an attempt to offer the international researcher interested in pursuing French doctoral dissertations an appreciation of the doctoral culture of French dissertation research. The complexity of a many-faceted doctoral system for each discipline in the humanities and social sciences may pose confusion in the mind for the international researcher not acquainted with French doctoral types and titles, especially for the years prior to 1984 when such types were effectively abolished. Their continued appearance in various reference works and published scholarship indicates a need for clarification of such nomenclature. Through a bibliometric approach focusing on a single discipline, an appreciation of French doctoral degree nomenclature and general characteristics provides salient information concerning philosophy which can be applied to other disciplines (Hérubel, 1991, Buchanan and Hérubel, 1993; Goedeken, & Hérubel, 1992; Kuyper-Rushing, 1999; Bernstein, 2002). For these reasons philosophy, as a canonical discipline with the epistemological and institutional power to transcend both the humanities and the social sciences, was chosen as the subject discipline for this study.

Methodology and Sample

To effectively map French doctoral dissertations, dissertations were retrieved from DocTheses, an online database. The database covers all recognized programs in France with doctoral granting authority and that culminate in doctoral degree awards. Dissertations in philosophy were chosen as the subject discipline for bibliometric analysis. The Docthèse rubric for *philosophy* was used to retrieve pertinent dissertations from among possible dissertations dealing with philosophical subjects to

illustrate the evolution of doctoral dissertations vis-à-vis different types of doctoral degrees for the years 1971 to 1990. These years represent a critical period before and after 1984 when doctoral research underwent a seismic shift in dissertation culture. Tabulation of data for year, discipline, granting institution, and type of degree were noted. Each category was analyzed for salient characteristics pertinent to discipline and type of doctoral degree. Where relevant, this data was examined as it reflected additional permutation, especially subject of degree.

Since this study is an attempt to offer some clarification of the three doctoral degree structure in France prior to the reforms of 1984, it is instructive to demonstrate some of the disciplinary as well as subject emphasis as they reflect characteristics critical to researchers not familiar with retrospective French dissertations prior to 1984. It is not unreasonable to consider that different disciplines would have different dissertation profiles in terms of type of degree granted. French universities enforce privileged certain disciplines over others as a normative administrative condition of higher education. Without this additional characteristic French doctoral education can not be easily appreciated. Consequently discussion of disciplinary permutations is necessary to a clearer picture of French doctoral culture.

Findings and Discussion

Upon examination, the data revealed a number of characteristics animating philosophy dissertations produced in France. A total of 2,491 dissertations were tabulated for 1971-1990. Throughout the sample years, production of philosophy dissertations gains to a peak of 164 in 1980, only to begin a slow and systematic decline to 94 dissertations in 1990 (Table 1)

Table 1. Dissertations produced per year

Year	No.	Year	No.
1971	98	1981	157
1972	95	1982	159
1973	51	1983	165
1974	111	1984	146
1975	121	1985	126
1976	128	1986	116
1977	100	1987	166
1978	107	1988	134
1979	147	1989	106
1980	164	1990	94

Another phenomenon is the type of dissertation produced over time (Table 2). Clearly the *doctorat de 3ième cycle* regimes constitute the largest dissertation grouping among doctoral degrees in philosophy. When both regimes are added together they form a numerically dominant 64.4% of the sample. However, they do not represent the highest intellectual achievements, nor the rigor demanded of the *doctorat d'état*, and represent a more systematized doctoral approach to research than the *doctorat d'université*. The emerging *doctorat nouveau régime*, soon to supplant all three previous doctoral regimes, shows substantial growth, while the others sustain general decline. The only caveat is the strength of the *doctorat d'état*, since it still carried previous perceived intellectual strength and cultural capital in French academia and could still be defended even into the late 1990s.

In general, as the *doctorat de 3ième cycle* and *doctorat nouveau régime* gained acceptance and strength, declines of the older established doctorates accelerates. Eventually, the *doctorat nouveau régime* will supplant the entire system. Interestingly, the *habilitation à diriger les recherches* does not yet appear.

Table 2. Dissertation types per year

Year	D'Etat	D3ancien	D3	DU	DN
1971	25	60	0	13	0
1972	19	73	0	3	0
1973	25	23	0	3	0
1974	24	82	0	5	0
1975	22	95	0	4	0
1976	32	92	0	4	0
1977	27	65	4	4	0
1978	18	45	41	3	0
1979	28	45	67	7	0
1980	33	22	100	9	0
1981	23	11	117	6	0
1982	29	11	118	1	0
1983	20	11	128	6	0
1984	28	7	108	3	0
1985	30	5	78	4	9
1986	28	0	63	1	24
1987	47	0	79	2	38
1988	46	0	33	1	54
1989	23	0	17	0	66
1990	28	0	3	0	63
<i>Total</i>	555	647	956	79	254
% of total	22.3%	26%	38.4%	3.2%	10.2%

The Doctorat D'Etat

Dissertation types reveal important information concerning a discipline's research culture; some disciplines, i.e. information science and communication, as well as less established disciplines, may not have produced *doctorat d'état* to the same extent as canonical disciplines. An even more interesting phenomenon is the case of the *doctorat d'état* and which institutions produced them. In this study, certain institutions and geographical dispersion are responsible for the majority of *doctorat d'état* produced in France for 1971-1990 (Table 3).

Table 3. Institutions producing *doctorat d'état*

Institution	No.	Institution	No.
<i>Aix-Marseilles I</i>	9	<i>Paris I</i>	173
<i>Amiens</i>	3	<i>Paris III</i>	3
<i>Besançon</i>	1	<i>Paris IV</i>	116
<i>Bordeaux III</i>	1	<i>Paris V</i>	1
<i>Caen</i>	2	<i>Paris VII</i>	3
<i>Claremand-Ferrand II</i>	4	<i>Paris VIII</i>	38
<i>Dijon</i>	11	<i>Paris X</i>	17
<i>EHESS</i>	3	<i>Perpignan</i>	1
<i>Grenoble II</i>	12	<i>Poitiers</i>	7
<i>Lille III</i>	9	<i>Reims</i>	2
<i>Lyon II</i>	1	<i>Rennes I</i>	1
<i>Lyon III</i>	21	<i>Rouen</i>	3
<i>Nice</i>	1	<i>Strasbourg II</i>	10
<i>Nancy II</i>	6	<i>Toulouse II</i>	18
<i>Nantes</i>	3	<i>Tours</i>	4

Because the *doctorat d'état* represents and occupies such a powerful presence in French academia, institutions producing it exercise greater advantage within disciplinary culture and evolution (Comité Français des Sciences Historiques, 1980; Thibault, 1972). Not only is full professorial status not possible without it, but the research agendas and future direction of disciplinary activity could be measured by successful completion of the *doctorats d'état* (Association Histoire au present, 2006). Here the evidence indicates the centralizing power of Parisian institutions and geographical clustering of *doctorats d'état*. Total for all Parisian Universities producing *doctorats d'état* is 409, or 73.7% of *doctorats d'état*. Provincial and other institutions produced 26.3% of *doctorats d'état* dissertations with Lyon III producing 21, Toulouse II 18, Grenoble II 12, and Strasbourg II 10 *doctorats d'état* while Lille III produced 9 and Dijon 8 respectively. Parisian institutional dominance, especially Paris I, Paris IV, Paris X, and Paris VIII universities, form a true canonical position within French academic philosophy (Joly, 1996; Godechot et Nicolas Mariot, 2003). Again it must be remembered that individual advisors or special research strengths appear at geographical venues and do not necessarily indicate differences in quality.

The Other Doctorates

Unlike the *doctorat d'état*, the *doctorat de 3ième cycle ancient régime*, *doctorat de 3ième cycle*, *doctorat d'université*, and *doctorat nouveau régime* represent different conditions as well as specific purposes. As mentioned before, these doctorates achieve different objectives, all culminating with vetted dissertations for directed research successfully undertaken. Although they may represent different regimes and different research agendas and protocols, each is a contribution to knowledge as understood by doctoral granting faculties. For this reason, they can be useful to international researchers requiring knowledge, techniques, or perspectives not otherwise attainable. If one examines those institutions responsible for these doctorates, an interesting topography emerges. When types of doctorates are tied to institutions, an intellectual geography of doctoral degrees becomes apparent. Numerically dominant, the *doctorat de 3ième cycle ancient régime*, *doctorat de 3ième cycle* reveal an institutional landscape that includes provincial universities not otherwise apparent (Table 4). Among the most prominent universities is Lyon III, Aix-Marseille I, and Toulouse II, while Tours, long known as a center for Renaissance Studies occupies a strong locus of philosophy among provincial universities. Poitiers, with its strong concentrations in Medieval Studies offers additional proof of provincial strengths. Strasbourg II forms a very strong center for philosophical research, especially as philosophy of religion is concerned. Among the non-university institutions, EHESS in Paris is an additional center for philosophy of science and social sciences orientation. However, Parisian universities, especially Paris, I, Paris IV, and Paris X comprise 66.1% of *doctorat 3ième cycle* doctorates.

The least numerically prominent, the *doctorat d'université* offers a unique perspective on research culture, especially relevant for philosophy. Often these dissertations represent unique approaches to philosophical topics, and often reveal unusual and often non-canonical research subjects. Combining philosophical rigor with interest in art or literature; or, eclectic philosophical perspectives, make this set of dissertations particularly interesting. Those institutions producing these dissertations reveal an equally interesting locus of activity (Table 5). The *Université de Paris VIII* occupies a unique position within the constellation of the Parisian university campuses (Soulié, 1998). It is not surprising to find that Paris VIII is the major producer of this doctorate, as many of the subjects entertained in the Department of Philosophy rarely conformed to the strict canonical subjects researched at other institutions, especially Paris I, Paris IV (Soulié, 1995; Soulié, 1994). Surprisingly, Paris X did not produce a single dissertation for this degree, since as Paris VIII, it too would entertain innovative and creative non-canonical subjects for philosophical research.

Table 4. Institutions producing *doctorat de 3ième cycle*

Institution	Total	Institution	Total
<i>Paris I</i>	600	<i>Nancy II</i>	9
<i>Paris X</i>	250	<i>Caen</i>	8
<i>Paris IV</i>	210	<i>Nice</i>	8
<i>Lyon III</i>	73	<i>Besançon</i>	7
<i>Aix-Marseille I</i>	66	<i>Paris III</i>	7
<i>Paris VIII</i>	64	<i>Paris V</i>	7
<i>Strasbourg II</i>	46	<i>Rouen</i>	7
<i>Toulouse II</i>	44	<i>Bordeaux III</i>	6
<i>Tours</i>	37	<i>Claremand-Ferrand II</i>	3
<i>Poitiers</i>	29	<i>Paris VII</i>	3
<i>Lille III</i>	28	<i>Amiens</i>	2
<i>Dijon</i>	24	<i>Nantes</i>	2
<i>EHESS</i>	20	<i>Claremand-Ferrand I</i>	1
<i>Grenoble II</i>	19	<i>Rennes I</i>	1
<i>Montpellier III</i>	12	<i>EPHE</i>	1
<i>Lyon II</i>	10	<i>Toulouse III</i>	1

Table 5. Institutions producing *doctorat d'université*

Institutions	Total	Institutions	Total
<i>Paris VIII</i>	22	<i>Dijon</i>	3
<i>Paris IV</i>	15	<i>Lyon III</i>	3
<i>Paris I</i>	9	<i>Poitiers</i>	2
<i>Strasbourg II</i>	7	<i>Lille III</i>	1
<i>Grenoble II</i>	6	<i>Reims</i>	1
<i>Toulouse II</i>	5	<i>Tours</i>	1
<i>Bordeaux III</i>	4		

Since 1984, the *doctorat nouveaux régime* has become the sole doctorate available to advanced graduate students. Consequently, its appearance and rapid acceptance is only natural. Those institutions granting it reveal similar profiles as seen for the previous older regime doctorates (Table 6).

Table 6. Institutions producing *doctorat nouveau régime*

Institutions	Total	Institutions	Total
<i>Paris I</i>	94	<i>Grenoble II</i>	3
<i>Paris IV</i>	41	<i>Paris XII</i>	3
<i>Paris X</i>	16	<i>Nancy II</i>	2
<i>EHESS</i>	14	<i>Nantes</i>	2
<i>Paris VIII</i>	14	<i>Rennes I</i>	2
<i>Lyon III</i>	13	<i>Amiens</i>	1
<i>Strasbourg II</i>	10	<i>Besançon</i>	1
<i>Nice</i>	7	<i>Bordeaux II</i>	1
<i>Poitiers</i>	6	<i>Montpellier I</i>	1
<i>Aix-Marseille I</i>	4	<i>Paris V</i>	1
<i>Lille III</i>	4	<i>Reims</i>	1
<i>Tours</i>	4	<i>Rouen</i>	1
<i>Caen</i>	3	<i>Toulouse II</i>	1
<i>Dijon</i>	3	<i>Strasbourg I</i>	1

Again Parisian institutions dominate the sample, indicating their longstanding emphasis in academic philosophy. Provincial institutions new to granting doctorates in philosophy, i.e. Amiens or Reims registered their first *doctorat nouveau régime* in 1989 and 1990 respectively. The most salient finding

is the predominance of Parisian universities, paralleling similar findings for the other doctorates, even as institutional dispersion is evident among provincial universities.

Parisian Dominance in Academic Philosophy

This limited study has revealed the most salient characteristic animating French academic philosophy. The dominance of Parisian universities exercises a powerful intellectual force within the intellectual geography of French academic philosophy. For the international researcher, this is critical as he/she may need to concentrate their information searching within this Parisian center of activity. Although Paris universities exert influence, other institutions represent a necessary intellectual counter-balance to Parisian dominance. Centers of strength and advanced pedagogy and research have been well established in such universities as Lyon III, Poitiers, Toulouse II, or Tours, where centers have emerged in certain periods, i.e. medieval philosophy, etc. It is crucial to understand that Parisian numerical dominance does not preclude provincial excellence; specialists in all philosophy sub-disciplines may be found throughout the French higher education landscape (Musselin, 2003). To appreciate this phenomenon it is interesting to examine the four dominant Parisian universities responsible—*Universités de Paris I, Paris IV, Paris X, and Paris VIII*. Each grew out of the turbulent 1960s reformist movements, producing distinctive campuses with distinctive objectives and disciplinary orientations. Both Paris I and Paris IV emerged from the massive Sorbonne, with Paris I maintaining social sciences, i.e. geography, political science, economics, etc. and the latter campus responsible for foreign languages and literatures, French literature, and classical studies. Since their inception, more social sciences and humanities disciplines have joined both campuses. Paris X was built in the Parisian suburb of Nanterre to accommodate larger student cohorts attending higher education in Paris; it offered similar program as Paris I and Paris IV. Paris VIII was born out of the immediate chaos of the student rebellion of 1968, quickly emerging as a truly experimental, highly innovative, if not pedagogically radical institution, celebrating multidisciplinary initiatives.

The data revealed that these four universities individually and collectively form a critical mass of philosophical doctoral research. Each exhibits its own intellectual orientation to philosophy. Paris I and Paris IV support a more established and traditional curriculum and research agenda, preparing students for entry in various academic venues. They represent the replication of professional disciplinary culture and pedagogical and research protocols, whereas Paris VIII remains exceedingly experimental and radical in its approach to philosophy. Unorthodox perspectives, however innovative are complemented by Marxist as well as other politicized orientations. Often subjects transcend the traditionally understood boundaries of what constitutes academic philosophy. Paris X remains a strong center for all aspects of philosophy, especially phenomenology and contemporary Continental philosophy. When the data was triaged for each institution, the following doctoral production configurations emerged attesting to the powerful centers each institution represents (Table 7).

Table 7. Parisian doctoral dominance

Institutions	D'Etat	D3ancien	D3	DU	DN	% of Total
<i>Université de Paris I</i>	174	219	381	9	94	35.2
<i>Université de Paris IV</i>	116	97	125	15	41	15.8
<i>Université de Paris X</i>	77	126	138	0	16	14.3
<i>Université de Paris VIII</i>	38	26	38	22	15	5.6

Both Paris I and Paris IV tend to produce canonical subjects, i.e. dissertations devoted to the ancient Greek philosophers, or Aristotle, Plato, and especially Kant. They are more historically grounded, often following closely wrought textual analysis of argumentation. Among other major figures, Leibniz, Hegel, or Descartes appear. Subjects appearing in Paris X dissertations reflect similar patterns of research attendant of Paris I and Paris IV; however, historical work appears less. Unlike its sister institutions, Paris VIII produced dissertations heavily linked to political or social concerns and philosophy. Aesthetics, where bridging art production or literary activity, were evident. Those dissertations emphasizing an historical approach were minor in comparison with subjects melding

various perspectives. Contemporary subjects, Marxist orientation and explication are complemented by traditional subjects as well as some canonical emphasis. When taken together, these Parisian universities constitute 70.9% of total doctoral production among those institutions offering doctoral education in philosophy. Their clear dominance of *doctorat d'état* as well as the other doctoral types cannot be under estimated nor exaggerated.

Conclusion

Researchers pursuing subjects in philosophy may find that they need to consult a dissertation in France. The bewildering nature of dissertations designated with pre-1984 titular regimes pose an interesting problem to the non-initiated. This introductory study attempts to offer exposure to the corpus of French doctoral dissertations as well as cursory examination of philosophy degrees and their institutional affiliations. To this end, a bibliometric examination of a single discipline offers an open window to the topography of French doctoral regimes in philosophy. Data indicated that the *doctorat de 3ième cycle* as a separate doctorate as opposed to the *doctorat d'état* or *doctorat d'université*, constituted the prevailing doctorate granted. Although not the highest or most rigorous, it represents the first professionally recognized doctoral qualification as well as systematic research accomplishment. The *doctorat d'état* remains the most singular symbol of research attainment and contribution in philosophy until its abolition in 1984. Idiosyncratic in orientation, but still significant for research purposes, the *doctorat d'université* offers another doctoral venue for research consultation.

When institutional affiliation was examined Parisian universities constituted a veritable research nexus of philosophical doctoral activity. Here, the power of the *doctorat d'état* dominates the scholarly landscape in such a fashion that French academic philosophy is certainly highly centralized. Barring specialties or a particular doctoral advisor, the Universités of Paris I, Paris IV, Paris X, and Paris VIII emerge as concentrated centers of doctoral activity in philosophy. For international researchers, the knowledge that certain institutions are major producers of doctoral dissertations and research can prove invaluable when considering preliminary research endeavors. From the first reference attempt to the possible *de visu* consultation, knowledge of French dissertation culture can prove critical to a successful research objective.

Although preliminary, this study offers possibilities for future research. It would be interesting to know which institutions produce dissertations in specific sub-disciplines. How does the type of doctorate affect the subject undertaken or what is the subject profile of doctoral types for a selection of years? More research could reveal relevant characteristics pertaining to doctoral research culture as manifested in French doctoral training and education. As revealed in the present study, the salient features of French doctoral research is that there are different regimes and types governing doctoral research and achievement, something that existed in French academia prior to 1984. Researchers may encounter dissertations with older titles into the 1990s indicating an evolutionary disappearance of those older doctoral regimes. This study only focused on a single discipline, but, it is an indicator of the entire pre-1984 doctoral degree structure attendant at one time for all academic doctoral programs in all disciplinary fields. For this critical reason, it is important to understand the nature of French doctoral dissertation research and culture.

References

- Assefa, A. M. (1988). *France: A Study of the Educational System of France and a Guide to the Academic Placement of Students in Educational Institutions of the United States*. New York: World Education Services & A.A.C.R.A.O., pp. 73-78.
- Association Histoire au présent, *Débuter dans la recherche historique*. Retrieved July25, 2006 from <http://www.bhistoire.com/e05p.htm>.
- Bernstein, J. H. (2002). First Recipients of Anthropological Doctorates in the United States, 1891–1930. *American Anthropologist*, 104, 551-564.
- Bowen, W. G. (1992). *In pursuit of the PhD*. Princeton, N.J.: Princeton University Press.
- Bourdieu, P. (1988). *Homo Academicus*, trans. by Peter Collier. Stanford: Stanford University Press.
- Buchanan, A. L. & Hérubel, J.-P.V.M. (1993). Comparing Materials Used in Philosophy and Political Science Dissertations: A Technical Note. *Behavioral & Social Science Librarian*, 12, 63-70.

- Chimisso, C. (2000). The Mind and the Faculties: The Controversy Over 'Primitive Mentality' and the Struggle for Disciplinary Space at the Inter-war Sorbonne. *History of the Human Sciences*, 13, 47–68.
- Chimisso, C. (2005). Constructing Narratives and Reading Texts: Approaches to History and Power Struggles Between Philosophy and Emergent Disciplines in Inter-war France. *History of the Human Sciences*, 18, 83 - 107.
- Comité français des sciences historiques. (1980). *La Recherche historique en France depuis 1965*. Paris: Éditions du Centre national de la recherche scientifique.
- Debeauvais, "M. (1986). Doctoral Theses in France: A Case of *reformitis*. *European Journal of Education*, 21, 375-384.
- Duroselle, J.-B. (1967). Thèses d'histoire contemporaine, faut-il bouleverse le système ? *Revue d'Histoire Moderne et Contemporaine*, 14, 173-180
- Duroselle, J.-B. (1970). Le nouveau doctorat" *Revue d'Histoire Moderne et Contemporaine*
- Fabiani, J.L. (1988). *Les philosophes de la République*. Paris: Editions de Minuit.
- Godechot, O. et Mariot, N. (2003). Les deux formes du capital sociale Structure relationnelle des jurys de thèse et recrutement en science politique. *Document de travail du GRIOT*, 17, 1-43.
- Goedeken, E. A. & Hérubel, J.-P. V.M. (1992). Dissertations in Military History, 1973-1988: A Survey and Analysis. *Journal of Military History*, 56, 651-657.
- Hérubel, J.-P. V. M. (1991). Philosophy Dissertation Bibliographies and Citations in Serials Evaluation. *The Serials Librarian*, 20, 65-73;
- Hesseling, P. G. M. (1986). *Frontiers of Learning: The Ph.D. Octopus*. Dordrecht: Foris Publications.
- Joly, G. (1996). Les thèses de géographie en France," *Inergéo-bulletin*, 124, 87-110
- Joly, G. (1997). Une base de données sur les thèses de géographie soutenues en France. *Cybergeo* 18 (1997), Retrieved July 25, 2006 from <http://www.cybergeo.presse.fr/ttsavoir/joly.htm>
- Kouptsov, O. compiler & edited by Barrows, L. C. (1994). *The Doctorate in the Europe Regio*. Bucharest: CEPES.
- Kuyper-Rushing, L. (1999). Identifying Uniform Core Journal Titles for Music Libraries: A Dissertation Citation Study. *College & Research Libraries*, 60, 153-63.
- Mousnier, R. (1965). Note sur la thèse principale d'histoire pour le doctorat des lettres. *Revue Historique*, 234, 123-127
- Musselin, C. (2003). *The Long March of French Universities*. NewYork: Routledge Falmer.
- de Ridder-Symoens, H., ed. (1992). *Universities in the Middle Ages*. Cambridge: Cambridge University Press.
- Rothblatt, S. & Wittrock, eds. (1993). *The European and American university since 1800: Historical and Sociological Essays*. Cambridge: Cambridge University Press.
- Rutledge, J. (1994). *Collection Management*. 19, 43-67.
- Schweitzer, G. K. (1965). *The Doctorate; A Handbook*. Springfield, Illinois: C.C. Thomas.
- Slone, G. T. (1990). The Ephemeral Doctorate. *Contemporary French Civilization*, 14, 85-88.
- Soulié, C. (1994). *La fabrique des philosophes ou des usages sociaux de l'U.F.R. de philosophie de Paris 1* (thèse du doctorat: Ecole des Hautes Etudes en Sciences Sociales).
- Soulié, C. (1995). Anatomie du goût philosophique." *Actes de la Recherches en Sciences Sociales*, 109, 3-28.
- Soulié, C. Le destin d'une institution d'avant-garde: histoire du département de philosophie de Paris VIII. Retrieved July 25, 2006 from <http://www.univ-Paris8.fr/sociologie/fichiers/soulie1998a.html>
- Thibault, A. (1972). L'Analyse des espaces régionaux en France depuis le début du siècle. *Annales de Géographie*, 81, 129-170.
- Wanner, R. E. (1975). France: A Study of the Educational System of France and a Guide to the Academic Placement of Students in Educational Institutions of the United States. New York: World Education Services & AACRAO, pp. 116-123;

Local Government Web Sites in Finland: A Geographic and Webometric Analysis

Kim Holmberg* and Mike Thelwall**

**kim.holmberg@abo.fi*

Information Studies, Åbo Akademi University, Tavastgatan 13, 20500 Åbo (Finland)

***m.thelwall@wlv.ac.uk*

School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB (UK)

Abstract

It has been shown that information collected from and about links between web pages and web sites can reflect real world phenomena and relationships between the organizations they represent. Yet, government linking has not been extensively studied from a webometric point of view. The aim of this study was to increase the knowledge of governmental interlinking and to shed some light on the possible real world phenomena it may indicate. We show that interlinking between local government bodies in Finland follows a strong geographic, or rather a geopolitical pattern and that governmental interlinking is mostly motivated by official cooperation that geographic adjacency has made possible.

Keywords

webometrics; hyperlinks; link creation motivations; local government; geography

Introduction

The web is an enormous source of information, both of a visible kind in the form of the content of web pages, and also of a more hidden kind, for example through the connections that hyperlinks create between different web sites and the organizations they represent. The research field of webometrics tries, among other things, to create new knowledge from this hidden information and to understand what kinds of real world phenomena it may represent.

Webometrics is still a young field of research and that is why a majority of webometric studies, after Almind and Ingwersen (1997) set the foundations for webometrics (and also gave the research field its name), have focused on developing and improving methods for data collection, processing and interpreting the results. These include studies about the use and coverage of commercial search engines for informetric purposes (Bar-Ilan, 1999; Ingwersen, 1998; Lawrence & Giles, 1999; Rousseau, 1997; Snyder & Rosenbaum, 1999; Thelwall, 2001a), interlinking between academic web sites on both national and international level (Thelwall, 2002a; Thelwall & Smith, 2002), networks and small worlds on the web (Björneborn, 2004; Björneborn & Ingwersen, 2004) and the use of links as indicators of business performance and business competitive positions (Vaughan, 2004; Vaughan & Wu, 2004; Vaughan & You, 2005). The results have collectively increased knowledge of how search engines function and how data gathered from the web should be processed and interpreted, how links and networks they create between web sites can be used to improve web information retrieval and how links can be used to indicate real world phenomena both in academia and business.

Although academic and, to a lesser extent, commercial web sites have been the subject of many webometric studies, much less has been written about government web sites. In this paper we seek to redress this imbalance through a case study of local government web sites in the region of Finland Proper (Varsinais-Suomi in Finnish) in the southwestern part of Finland. From the perspective of local government, we believe that geography is particularly important and hence we focus on this issue.

Literature review

Geography Some previous research has investigated geographic factors affecting linking. In a study of academic interlinking in the UK Thelwall (2002a) showed that universities close to each other tend to

interlink more than distant ones. The exact coordinates of the locations of the universities were used to calculate the distances between them for this research. Later, Tang and Thelwall (2003) found only limited evidence of geographic trends in interlinking within the web sites of three different U.S. academic disciplines.

International interlinking has also been occasionally studied, giving an alternative perspective on geographic factors. A study of the links between universities in UK, Australian and New Zealand gave an interesting finding: that New Zealand was rather isolated on the web (Smith & Thelwall, 2002). A larger-scale study of international interlinking covered the whole Asia-Pacific region through its universities and mapped the connections in network diagrams (Thelwall & Smith, 2002). This showed the central roles of both Australia and Japan, indicating that geographic proximity was less significant than geopolitical role in this context.

Motivations To be able to interpret link counts or network diagrams created based on links, it is essential to understand the underlying link creation motivations. Kim (2000) investigated authors' motivations for creating links in e-journals and found that some motivations were similar to those previously discovered for citing in printed scientific journals. Those that were new were related to accessibility of electronic sources. In a classification of academic hyperlinking, five types of creation motivations were identified: ownership, social, general, navigational and gratuitous (Thelwall, 2003). In a more extensive study Wilkinson, Harries, Thelwall and Price (2003) classified a random collection of 414 links between UK universities. Although 90% of the links were related to scholarly activities, only 1% of the links were equivalent to citations in scientific journals. This suggests that interlinking between academic web sites is evidence of informal rather than formal scientific communication. In a case study of interlinking between Israeli academic institutions Bar-Ilan (2004; 2005) proposed a multi-faceted classification of link types, in which she included many different aspects of the links and the source and target pages, i.e. source page types, target page types, creators, intention of linking and the relationship between the link area and the target. Bar-Ilan's categorizations give a very complex picture and cover many different aspects of academic link creation. Chu (2005) crystallized her findings about reasons for hyperlinking into one sentence: "One links to a site for what it is about", meaning that links are made to point to other sites that are relevant for the outlinking site. From the above results we can conclude that academic interlinking is based on informal communication with parties tending to be close to each other. This suggests that geographical distance is a component in informal scientific communication.

Government links E-Government, in terms of information technology use, has been studied from several aspects. Moon (2002) studied how e-government has evolved among municipalities and discovered that the size of the municipalities and cities (by population) and age of the web sites were positively associated with the adoption of e-government. Petricek, Escher, Cox and Margetts (2006) developed a methodology to quantitatively evaluate structural characteristics of e-government as indicators for web site navigability and 'nodality'. Their methodology gave meaningful evaluations of government web sites in a cheap and efficient way using freely available web link data. Government links alone have not been studied extensively before with webometric methods, but there is some research about university-government links and university-industry-government linking. Stuart and Thelwall (2005) investigated university-to-government links as indicators of collaboration between UK universities and government. Although a correlation between university's research productivity and number of outlinks was found, there was no evidence of a causative connection between university's research productivity and the number of university-to-government links. In a later study Stuart and Thelwall (2006) examined whether web URL citations could be used as 'weak benchmarking indicators' to investigate collaboration between universities, industry and government. As a part of this study directional graphs based on the URL citings showed clustering between the district councils (LAU 1¹) and their respective county councils (NUTS-3¹) and also between

¹ The Nomenclature of Territorial Units for Statistics (NUTS) is a uniform scale that makes it possible to compare regional statistics in European Union (Statistical Regions of Europe, 2006).

neighboring local government bodies. The study showed that county councils tend to link more often to other county councils and equally district councils tend to link more often to other district councils under the same county council. There were also geographic patterns in government to university URL citations. Although URL citations correlate statistically with some real world phenomena, a closer analysis of the URL citations indicated that these were more often created to point to information sources than to reflect cooperation or other relationships between the studied organizations, which was also the case in Stuart and Thelwall (2005).

Background

The regional and local administration levels in Finland are divided into six provinces, 20 regions and 431 municipalities. The provinces are on level 2 on the *Nomenclature of Territorial Units for Statistics* (*Nomenclature des Unités Territoriales Statistiques - NUTS*) scale (Statistical Regions of Europe, 2006). The regions are on NUTS-3 level and municipalities are on NUTS-5 level. The more unofficial regions called functional regions are placed on NUTS-4 level. The NUTS-4 level is equal to the newer LAU 1 level (*Local Administrative Units*) and NUTS-5 equals LAU 2.

The region of Finland Proper is located in the southwestern part of Finland and is slightly smaller than Wales, Israel or New Jersey. About half of this area is land and the rest is sea. In January 2006 the population was 456,000 people. The city of Turku is the capital of the region and has over 170,000 residents. The region of Finland Proper has five functional regions (Figure 1).

- Loimaa is known for agricultural and provisions production.
- Salo is a leader in high technology industry and contains some of mobile phone manufacturer Nokia's main plants.
- Turunmaa (Åboland in Swedish) attracts tourism for its unique archipelago.
- Vakka-Suomi hosts Finland's car factory and main metal industry.
- Turku, the capital of Finland Proper, is an old university city with high quality research and manufacturing, e.g. in biotechnology.



Figure 1. Functional regions in the region of Finland Proper (Varsinais-Suomi)

Research Questions

From a broad perspective, the goal of this research is to contribute to understanding government-related web linking, both to help develop effective webometric techniques and to build understanding of the phenomenon on an international scale. The aim of this case study is to map possible geographic trends in interlinking between municipalities in Finland Proper and to study what motivates the interlinking between municipalities. Can interlinking between municipalities be used to trace some other existing trends or phenomena? The following research questions drive this investigation.

1. Does local government web site interlinking in Finland Proper follow geographic lines?
2. Why do local government web sites in Finland Proper interlink?

Methods

Data collection There are three possible methods for collecting link data. The first is to visit every web site and manually collect the information needed about the content and links. This can be very strenuous and only suitable to study a very small set of web sites. It is not practical to collect the link data manually from the 54 web sites in this study. The second method is to use commercial search engines to count the occurrences of links and pages. It has been shown that no search engine covers the whole web, and that coverage of major search engines may be as low as 16% (Lawrence & Giles, 1999). Even if later studies have found much higher coverage rates (Gulli and Signorini, 2005) it is still unclear exactly how well commercial search engines cover the web. It has also been shown that search engines have coverage biased towards the U.S. (Vaughan & Thelwall, 2004). This is a problem because if we used a search engine then we couldn't be sure if all the pages and links would be covered. The third and final method is to use a web crawler that will automatically and independently follow every link within a given web site and collect information about the links and the content on the pages. A crawler also has its disadvantages. The crawler may not be able to index as many types of non-html pages and dynamically created pages as a commercial search engine might be able to. We still used this approach because it gave us greater control of the results and hence was a better solution than using commercial search engines (Thelwall, Vaughan & Björneborn, 2005).

Data modeling A problem with using link data is that a single person can create several links in a directory or even on a single web page, which would skew the data collected about the links. To improve the quality of link data, the 'Alternative Document Model' (ADM) concept has been developed (Thelwall, 2002b; Thelwall & Wilkinson, 2003). When using ADMs, duplicate links are combined at different hierarchical levels of a web site's file structure. This minimizes the effect of links that are created repeatedly because of a web design decision. For instance, in this study the municipality of Lieto had a link on almost every page of the site to the web based map service hosted on the web site of Turku. Using the site level of ADM removes such duplicate occurrences of links, leaving only one for the whole site. Counting links at the site ADM level in the above example would give only one outlink from Lieto to Turku. In this study we are only interested in the existence or the lack of a link between sites and therefore use the site level ADM.

Visualization The data about interlinking links were collected with a special information science web crawler in July 2006 and the site ADM level of counting links was applied on the data with another software distributed freely in the same package as the crawler (Thelwall, 2001b, 2002c). The link data was converted to a squared binary 54x54 site-by-site matrix with BibExcel (Persson, 2006). In this matrix, 1 indicates the existence of a link and 0 indicates the lack of a link. The outlinks can be read from the rows and the inlinks can be read from the columns in the matrix, one row and column per municipality. The matrix was imported to Ucinet (Borgatti, Everett & Freeman, 2002) for analysis and visualized as a two-dimensional network map (Figure 2) with Pajek (Batagelj & Mrvar, 2003). In the map nodes represent municipal websites and the lines represent hyperlinks between them. Arrows indicate the direction of the hyperlinks. The locations of the nodes and the distances between the nodes were calculated with the commonly used Kamada-Kawai spring-based algorithm, (Kamada & Kawai, 1989), via Pajek (Figure 3). Kamada-Kawai positions nodes close to other nodes that they are linked to by pulling them closer as if the links were springs. A network map was chosen rather than a proximity-based mapping technique, such as Multi-Dimensional Scaling, because there are few

enough links for the network diagram to be clear and hence the network diagram is easier to interpret because it directly reflects the raw data (i.e. 1s and 0s of the matrix).

Geographic testing To study possible geographic trends in interlinking between municipalities we created a simple binary matrix based on shared borders of the municipalities. Another possibility would have been to use distance between cities and municipalities. This approach would have been problematic because cities seldom lie in the geographic center of the municipality. Also the very large archipelago has definitively a big influence on the geographic distances between municipalities. To create a binary matrix based on shared borders gives a simple way to compare interlinking with geography. Strictly speaking our matrix is geopolitical rather than geographic. In the matrix a neighbor or a shared border is indicated with a 1 and the lack of a shared border is indicated with a 0. In a geographic matrix based on neighbors, the relationship, or the neighborhood, is always reciprocal. So if Nagu is a neighbor of Pargas, then Pargas has to be a neighbor of Nagu. The matrix is therefore symmetric, in contrast to the interlinking matrix, which does not have to be symmetric. We used QAP (described below) to test for similarity between the geographic and link matrices.

Classification To answer the second research question we took a random sample of links between the municipalities and visited both the source pages and the target pages to determine the underlying motivations to create links. We took a random sample of up to 20 interlinking outlinks from each of the 54 municipalities. This would have given a total of 1080 links, but because most of the municipalities did not have 20 outlinks to other municipalities, only 496 links were found. Some of the links were located on a single source page, giving 337 source pages to start from. We used the combined content and intention of the sources, the links and the targets as a basis for our categorization. As previous studies have shown, there is no uniform model for categorizing links, so we developed our own categorization designed for links between local government bodies and reflecting the duties and the responsibilities of municipalities. Once the categories were completed we grouped the categories based on a mutual intention or purpose of the content of the categories.

Results

Geography When counting links with the site level ADM, there were 315 links between municipal web sites. When converted to a site-by-site matrix with BibExcel, there were 315 links or ties present in the matrix, which is 11.0 percent of all possible ties. This is also the density of the matrix. For a binary squared matrix this also measures the probability that any given tie between two random municipalities is present. Figure 2 reveals the central role that Turku plays in the area and that Turku is the “capital” of the region, even on the web. Turku has the most outlinks to (53) and receives the most inlinks from (47) other municipalities.

The municipalities are located quite clearly in clusters according to their functional regions, with the exception of the functional region of Turku and the functional region of Vakka-Suomi. The border between these two functional regions is like an invisible barrier, which is also the case in the real world. The municipalities in the functional regions of Turunmaa Loimaa both create quite strongly connected clusters. The municipalities in the functional region of Salo also create a clear cluster, but this cluster is not as well connected. It is almost only the links from and to Salo (which is the biggest city in the functional region) that keep the functional region connected to the rest of the municipal web space in the region of Finland Proper. Figure 3 below shows the interlinking between municipalities drawn with Pajek and using the Kamada-Kawai algorithm. The functional regions are drawn by hand, showing clearly that the regions do not overlap.

When comparing the matrix of neighboring municipalities with the matrix of interlinking between municipalities in the same region, we found that the matrices matched each other to 87.2% accuracy. In other words, there is a probability of 87.2% that for any value in any given cell in one of the matrices, the same value would be found in the corresponding cell from the second matrix. This match was tested with a dyadic matrix correlation test called the Quadratic Assignment Procedure or QAP (Krackhardt, 1992). QAP matrix correlation calculates the probability of getting the same match by

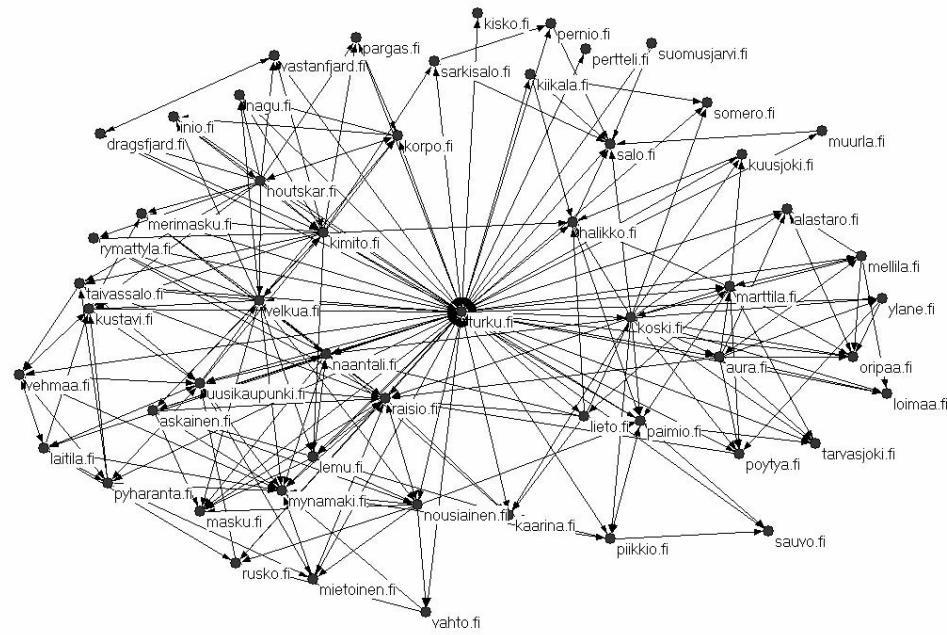


Figure 2. Interlinking municipalities in the region of Finland Proper (Varsinais-Suomi)

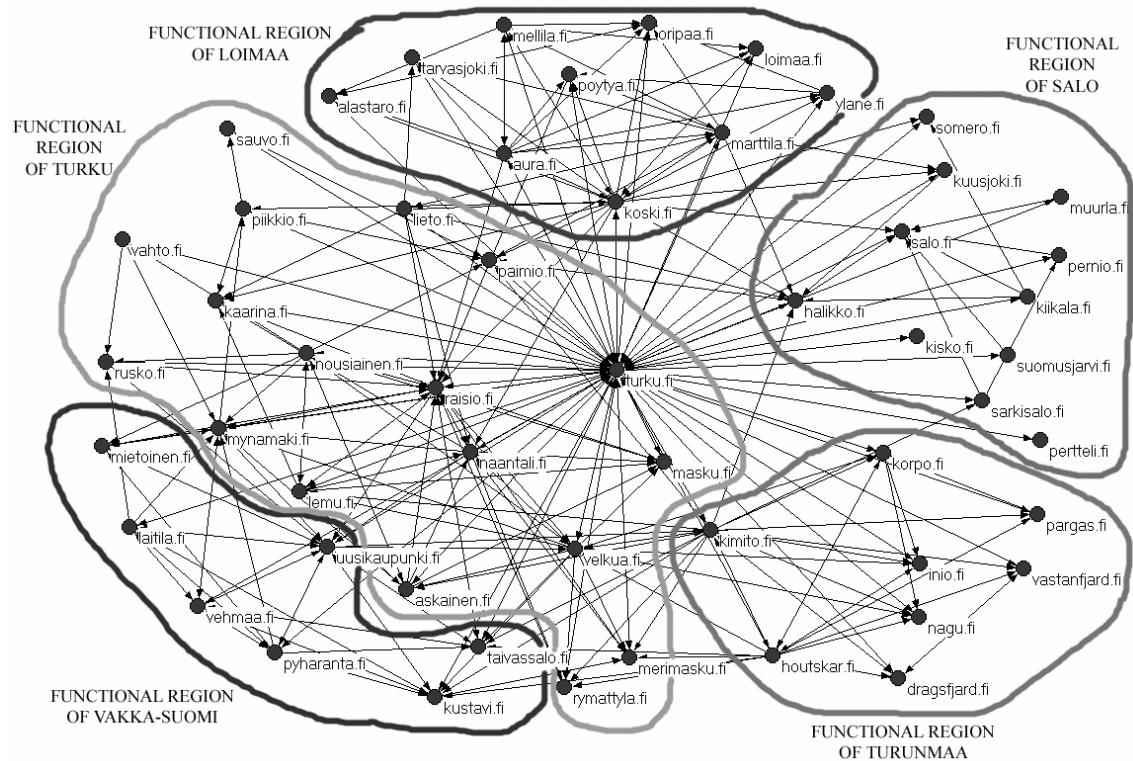


Figure 3. Interlinking and the functional regions in the region of Finland Proper (Varsinais-Suomi)

accident, compared to random permutations of the rows and columns in the matrices. QAP indicated that a match of 80.4% between the matrices would occur if there was no underlying similarity. This high figure is due to both matrices having mainly zeros. The difference between 87.2% and 80.4% is big enough that the chance of getting a match of 87.2% or over between the matrices is approximately $p=0.000$. This is strong statistical evidence that geographic proximity plays an important role in creating links between the studied municipalities, answering the first research question.

Link creation motivations A total of 337 links between municipalities were visited and categorized, via their source and target pages, with the results shown in Table 1. In fact, both the source and target pages had the same topic in a majority of cases, something that has not been true for academic source and target pages (Bar-Ilan, 2004, 2005).

Table 1. Categorization of the motivations to link

Number of links	Type of motivation for linking	Percent of total	Official Cooperation	Geographic motivation
59	Map service, hosted by Turku municipality	17.51%		17.51%
43	Rescue services in the region, hosted on Turku municipality's website	12.76%	12.76%	12.76%
27	Link list, to closest municipalities (about 2-8 municipalities)	8.01%		8.01%
22	Libraries, agreement of shared resources	6.53%	6.53%	6.53%
22	Tourist information, attractions, link list to the closest municipalities	6.53%	6.53%	6.53%
21	Education, agreement about placing pupils	6.23%	6.23%	6.23%
18	Link list	5.34%		
17	Health care in the region	5.04%	5.04%	5.04%
12	Education, some cooperation and link list to close municipalities	3.56%	3.56%	3.56%
11	Consumer information, external information source	3.26%	3.26%	3.26%
11	Database of Turku Library	3.26%	3.26%	3.26%
9	Culture, links to closest municipalities	2.67%		2.67%
9	Business information services, area development, external information	2.67%	2.67%	2.67%
9	Nature and environmental protection	2.67%	2.67%	2.67%
9	Recreation and hobbies in close municipalities	2.67%		2.67%
8	Businesses, organizations and associations in close municipalities	2.37%		2.37%
8	Social service information	2.37%	2.37%	2.37%
6	Agriculture, cooperation, external information	1.78%	1.78%	1.78%
5	Waste management and recycling	1.48%	1.48%	1.48%
4	Sports in close municipalities	1.19%	1.19%	1.19%
3	Living, links to close municipalities	0.89%	0.89%	0.89%
2	Local newspapers	0.59%		0.59%
1	Traffic, local buses in Turku	0.30%	0.30%	0.30%
1	Unemployment, links to close municipalities	0.30%	0.30%	0.30%
337	Total	100.00%	60.84%	94.67%

Two trends were discovered: links were often motivated by official cooperation between municipalities and links were often motivated by the geographic closeness of the municipalities. These two groups overlapped, as can be seen from Table 1 above. The group of links motivated by official cooperation includes links to services that the municipalities had bought from other municipalities or that they had agreed to join together for. For instance, some municipalities do not have Swedish-speaking schools even if they have Swedish-speaking residents. These have an agreement with another adjacent municipality to buy a Swedish-speaking education for their residents. A total of almost 61% of the interlinking is motivated by similar official cooperation between the municipalities. Examples of links motivated by geographic proximity are local rescue services and tourist attractions. A total of almost 95% of the links are motivated for various reasons by the fact that the municipalities are close to each other. In answer to the second research question, we can conclude that the results strongly indicate that links are primarily created to reflect official cooperation and that geographic closeness is a factor in the vast majority of cases.

Discussion and conclusions

The aim of this study was to investigate if interlinking between municipalities in the region of Finland Proper would correlate with the geography of the region and also why these links were created. We found that interlinking correlates quite strongly with geography. The most interesting finding, however, is that the interlinking tends to be motivated by official cooperation that geographic closeness has

made possible, in contrast to Stuart and Thelwall's (2006) UK study. Although Stuart and Thelwall discovered linking to reflect some real world phenomena, motivations for creating links and URL citations did not support a cause-and-effect relationship. A possible reason for this difference may be the fact that Stuart and Thelwall studied larger local authorities (NUTS-3 and LAU 1) while in our study we concentrated on municipalities (LAU 2). Some of the municipalities in our study were very small (only 245 residents in one municipality) and hence more dependent on cooperation with other municipalities. This cooperation also seems to be reflected on the web. As a result, we can conclude that it is possible that interlinking can be used to map official cooperation and networks based on cooperation between municipalities and cities in Finland.

Our case study has produced unusually clear-cut results for a webometric analysis, both in terms of identified geographic factors and the link creation motivations. Although two weaknesses of the classification study are the use of a single classifier and the need to infer link creation motivations rather than directly asking the link creators why the links were created, its geographic results are at least corroborated by the QAP test. Nevertheless, there is a general limitation: it is not clear to what extent the results can be generalized. Presumably, similar results would be found for any other area of Finland but perhaps there are few other countries that have as extensive a system of local government (Finland has the highest UN democracy index) and as extensive use of the web (Finland is one of the most online countries in the world). This hints that Finnish online local government may even be a best case for the world, rather than a typical example. If this were to be true, as future research will hopefully reveal, then the results would be useful as an extreme case of what is possible.

References

- Almind, T. C. & Ingwersen, P. (1997). Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4), 404-426.
- Bar-Ilan, J. (1999). Search engine results over time - A case study on search engine stability. *Cybermetrics*, 2/3(1), paper 1. Retrieved February 28, 2007 from <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.pdf>.
- Bar-Ilan, J. (2004). A microscopic link analysis of academic institutions within a country - the case of Israel. *Scientometrics*, 59(3), 391-403.
- Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management*, 41(4), 973-986.
- Batagelj, V. & Mrvar, A. (2003). *Pajek - Analysis and visualization of large networks*. In: *Graph Drawing Software*, M. M. Jünger, P., Editor. Berlin: Springer, 77-103.
- Björneborn, L., *Small-world link structures across an academic Web space: A library and information science approach*. PhD Thesis. Royal School of Library and Information Science, Copenhagen, Denmark, 2004. Retrieved February 28, 2007 from <http://vip.db.dk/lb/phd/phd-thesis.pdf>.
- Björneborn, L. & Ingwersen, P. (2004). Towards a basic framework for webometrics. *Journal of American Society for Information Science and Technology*, 55(14), 1216-1227.
- Borgatti, S. P., Everett, M. G. & Freeman, L. C. (2002). *Ucinet for Windows: Software for Social Network Analysis*. Harvard, MA: Analytic Technologies.
- Chu, H. (2005). Taxonomy of inlinked Web entities: what does it imply for Webometric research? *Library & Information Science Research*, 27(1), 8-27.
- Gulli, A. & Signorini, A. (2005). *The indexable web is more than 11.5 billion pages*. Proceedings of the 14th international World Wide Web conference (WWW2005). [CD Version] Retrieved February 28, 2007 from http://www.di.unipi.it/~gulli/papers/f692_gulli_signorini.pdf.
- Ingwersen, P. (1998). The calculation of Web impact factors. *Journal of Documentation*, 54(2), 236-243.
- Kamada, T. & Kawai, S. (1989). An Algorithm for Drawing General undirected Graphs. *Information Processing Letters*, 31(1), 7-15.
- Kim, H. J. (2000). Motivations for hyperlinking in scholarly electronic articles: a quantitative study. *Journal of American Society for Information Science*, 51(10), 887-899.
- Krackhardt, D. (1992). A caveat on the use of the Quadratic Assignment Procedure. *Journal of Quantitative Anthropology*, 3, 279-296.
- Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the web. *Nature*, 400, 107-109.
- Moon, M. Jae (2002). The Evolution of E-Government among Municipalities: Rhetoric or Reality? *Public Administration Review*, 62(4), 424-433.
- Persson, O. (2006) Bibexcel - a toolbox for bibliometrists (Version 2006-05-11). [Computer software]. Retrieved May 11, 2006 from <http://www.umu.se/inforsk/Bibexcel/>.

- Petricek, V., Escher, T., Cox, I. J. & Margetts, H. (2006). *The Web structure of e-government - developing a methodology for quantitative evaluation. Proceedings of the 15th international conference of World Wide Web*. Edinbugh, Scotland: ACM Press. Retrieved February 28, 2007 from <http://www2006.org/programme/files/pdf/1041.pdf>.
- Rousseau, R. (1997). Sitations, an exploratory study. *Cybermetrics*, 1(1), paper 1.
- Smith, A. & Thelwall, M. (2002). Web impact factors for Australasian universities. *Scientometrics*, 54(3), 363-380.
- Snyder, H. & Rosenbaum, H. (1999). Can search engines be used for web-link analysis? A critical review. *Journal of Documentation*, 55(5), 577-592.
- Statistical Regions of Europe, 2006. *Nomenclature of territorial units of statistics*. Retrieved November 22, 2006 from http://ec.europa.eu/comm/eurostat/ramon/nuts/home_regions_en.html
- Stuart, D. & Thelwall, M. (2005). *What can university-to-government web links reveal about university-government collaborations? Proceedings of the 10th International Conference of the International Society of Scientometrics and Informetrics*. Stockholm, Sweden: Karolinska University press. Vol. 1, 188-192.
- Stuart, D. & Thelwall, M. (2006). Investigating triple helix relationships using URL citations: a case study of the UK West Midlands automobile industry. *Research Evaluation*, 15(2), 97-106.
- Tang, R. & Thelwall, M. (2003). U.S. academic departmental Web-site interlinking in the United States Disciplinary differences. *Library & Information Science Research*, 25, 437-458.
- Thelwall, M. (2001a). The responsiveness of search engine indexes. *Cybermetrics*, 5(1), paper 1.
- Thelwall, M. (2001b). A web crawler design for data mining. *Journal of Information Science*, 27(5), 347-359.
- Thelwall, M. (2002a). Evidence for the existence of geographic trends in university Web site interlinking. *Journal of Documentation*, 58(5), 563-574.
- Thelwall, M. (2002b). Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university web sites. *Journal of American Society for Information Science and Technology*, 53(12), 995-1005.
- Thelwall, M. (2002c). Methodologies for crawler based web surveys. *Internet Research: Electronic networking and applications*, 12(2), 124-138.
- Thelwall, M. & Smith, A. (2002). A study of the interlinking between Asia-Pacific University Web sites. *Scientometrics*, 55(3), 335-348.
- Thelwall, M. (2003). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(3). Retrieved February 28, 2007 from <http://informationr.net/ir/8-3/paper151.html>.
- Thelwall, M. & Wilkinson, D. (2003). Three target document range metrics for University Web sites. *Journal of American Society for Information Science and Technology*, 54(6), 490-497.
- Vaughan, L. & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management*, 40(4), 693-707.
- Thelwall, M., Vaughan, L. & Björneborn, L. (2005). *Webometrics*. In: *Annual Review of Information Science and Technology*, B. Cronin, Editor. Medford, NJ: Information Today Inc. Vol. 39, 81-135.
- Vaughan, L. (2004). Exploring website features for business information. *Scientometrics*, 61(3), 467-477.
- Vaughan, L. & Wu, G. (2004). Links to commercial web sites as a source of business information. *Scientometrics*, 60(3), 487-496.
- Vaughan, L. & You, J. (2005). *Mapping business competitive positions using web co-link analysis. Proceedings of the 10th International Conference of the International Society of Scientometrics and Informetrics*. Stockholm, Sweden: Karolinska University press. Vol. 2, 534-543.
- Wilkinson, D., Harries, G., Thelwall, M. & Price, L. (2003). Motivations for academic web site interlinking: evidence for the Web as a novel source of information on scholarly communication. *Journal of Information Science*, 29(1), 49-56.

Visualizing the Topic Space of the United States Supreme Court¹

Peter A. Hook

pahook@indiana.edu

Indiana University, School of Law, 211 South Indiana Ave., Bloomington, IN 47405 (USA)

Abstract

This article describes the creation of several domain maps based on the topic space of opinions issued by the United States Supreme Court. Topics assigned by West Publishing were harvested off of the Westlaw database and visualized using Principal Components Analysis (PCA), Multidimensional Scaling (MDS), and graph visualization software (Pajek). Peculiar topic adjacencies were noted and attributed to the unique nature of cases argued at the level of the United States Supreme Court. The work is contextualized throughout by the author's desire to create a rigorous base map on which to layer additional data for teaching purposes.

Keywords

domain maps; visualizations; United States Supreme Court; Law; political science; westlaw

Introduction

Background and Purpose

Scientometrics and bibliometrics owe a debt of gratitude to the legal research publishing industry in the United States. Frank Shepard's legal citator (Ogden, 1993) was part of the inspiration for Eugene Garfield's *Science Citation Index* and subsequent products (Garfield, 1955 & 1979). This in turn was part of the inspiration for Page and Brin's PageRank algorithm—the foundation for Google (Hopkins, 2005; Battelle, 2005). Now, the tools of scientometrics may assist the legal research publishing industry to more optimally organize its materials. Legal information is itself exciting because it is one of the largest and most atomistically indexed bodies of information.

This research seeks to identify the topical adjacencies of subjects addressed in legal cases by the United States Supreme Court based on the co-occurrence of top level topics assigned by West Publishing (Thomson/West, 2006). It is in furtherance of the author's goal of creating a rigorous substrate map on which to layer over sixty years of Supreme Court topic data to be used for teaching purposes. In addition, the research is related to a growing body of work detailing and analyzing the network structure of legal opinions and their citation linkages (Chandler, 2005; Cross & Smith, In Press; Cross, Smith & Tomarchio, In Press; Fowler et. al., In Press; Smith, In Press), judicial and legislative co-voting networks (Fowler, 2006; Epstein et. al., 2005; Johnson et. al., 2005; Poole, 2005; Porter et. al., 2005; Sirovich, 2003; Brazill, 2002; Grofman, 2002; Martin & Quinn, 2002; Spaeth & Altfeld, 1985; Schubert, 1962 & 1963; Thurstone & Degan, 1951; Pritchett, 1941), and the move in legal academia toward quantitative empirical scholarship (George, 2006).

Maps of inherently non-spatial data that use a spatial substrate on which to layer additional information are common in information science (Hook & Börner, 2005). These maps employ the distance-similarity metaphor by which the viewer infers that items more proximate in space are more related than items further apart (Montello et al., 2003; Skupin and Fabrikant, 2003). The benefit of a substrate map is that it provides a common background from which changes may be readily perceived and is thus useful for pedagogy and illustrating changes over time.

Spatial layouts of inherently non-spatial data may be created in several ways. The first way is by the opinion of experts as to which topics are most similar and by laying out those topics by intuitive warrant or heuristics (*See* Bernal, 1939; Ellingham, 1948). The second way is by algorithmic comparisons of similarity and automated layouts using objective measures such as citation linkages or the co-occurrence of terms (Börner, Chen, & Boyack, 2002). Finally, a third method is a fusion

¹ Full color images, the text of this paper, and additional appendixes are available at: <http://ella.slis.indiana.edu/~pahook/index.html>.

approach which combines elements of each of the first two methods. For the most part, this paper employs multivariate statistical techniques that fall into the second category. These techniques are principal component analysis (“PCA”) and multidimensional scaling (“MDS”). However, elements of the fusion approach were used when the author placed data elements into higher level categories based on his training in and experience of the United States legal system before employing the multivariate statistical techniques.

Methods, Materials, Procedures, and Equipment Used

Data Summary

The dataset used for this research consists of bibliographic information about all United States Supreme Court cases that have been issued West topics by West Publishing from the 1944 Term through the end of the 2004 Term (October 1944 through July 2005). The author harvested the data as an academic end user from the Westlaw database. The data contains information about 7,948 unique Supreme Court cases to which 19,789 topic assignments have been made. Of the 405 top level topics in the West taxonomy, 290 appear in opinions issued by the Supreme Court for this time period. All but one (“Reference”), co-occur with other topics resulting in 22,345 edges between cases sharing a similar topic. There are 3743 unique topic pairings.

About the Data

For over a hundred years, West Publishing has identified unique statements of law within court cases (Surrency, 1990). Human editors working at West assign these unique and legally controlling statements topic identifiers from its taxonomy of the law known as the West Topic and Key Number System (Doyle, 1992; Snyder, 1999; Thomson/West, 2006). Before the advent of online full-text searching, the West Topic and Key Number System was one of the only ways to research cases on a given issue. Now, the Topic and Key Number System is used primarily to augment free text searching and to convince a researcher that he or she has found all of the appropriate cases on a particular topic. The Westlaw Database, owned by Thomson/West Publishing, provides online access to United States Supreme Court opinions, numerous other cases, and additional legal material. It is a proprietary subscription database that includes both the actual language of court opinions plus editorial enhancements provided by West such as topic assignments from the West Topic and Key Number System.

Data Harvesting

The data was harvested off of the Westlaw database during March through April, 2004. As of March 18, 2004, there were 405 top level topics in the West Topic and Key Number System. A search as to each of the 405 topics was conducted by hand using the conventional end user interface. A typical search statement was: TO("2 Abatement and Revival"). The TO in this case means topic and the scope of the database at the time included all Supreme Court opinions from the 1944 term to date. The resultant list of cases for each of the 405 topic searches were placed into a spreadsheet along with the topic that caused the case to be returned by the database. Topic assignments were aggregated such that each case was listed with all of its topic assignments and did not appear more than once. Subsequent Supreme Court cases and their topic assignments were added later. Annually, West makes changes to its taxonomy. In order for the dataset to include cases after the original March through April 2004 harvesting period, the author had to account for these changes. On several occasions, new topics were converted to their previous equivalents to bring the dataset current through the end of July 2005.

Additional Human Coding of the Data

Additionally, the author employed his legal training and knowledge of how concepts are taught in law school to make additional subject matter assignments to the 405 West topics: (1) Doctrinal – relevant to a specific subject taught in law school. (Constitutional Law, Administrative Law), (2) Factual – with unique factual circumstances relating to the topic but whose doctrinal elements are drawn from other topics (Aviation Law, Automobile Law), and (3) Procedural – capable of arising in almost any factual or doctrinal situation (Federal Courts, Federal Civil Procedure). For the doctrinal and

procedural topics, the author also assigned categories to the topics based on in what course they are most likely to be covered in law school.

Data Manipulation and Visualization

The data was imported to the R statistical computing environment. Before applying the multivariate statistical visualization techniques, the data had to be put into matrix form. The data comprises a sparse matrix of 3743 unique topic pairings out of a theoretically possible 83,521 (289 x 289). The range of topic co-occurrence counts is 1 to 896 (with Constitutional Law and Federal Courts (896) being the most commonly co-occurring topics and Constitutional Law and Criminal Law (468) being the second most common). The mean topic co-occurrence count was only 5.97 and the median and mode were both 1. Both PCA and MDS were performed on the data. PCA was performed using Singular Value Decomposition (SVD). The resultant plots were useful to characterize the major dimensions in the variation in the data of topic co-occurrence. (See generally Paolillo and Wright, 2006). Additionally, the dataset was visualized in its network form using the network visualization and analysis tool, Pajek (Batagelj & Mrvar, 1998). In the parlance of network science, the nodes represented West Topics and the edges represented the co-occurrence of those topics in Supreme Court cases.

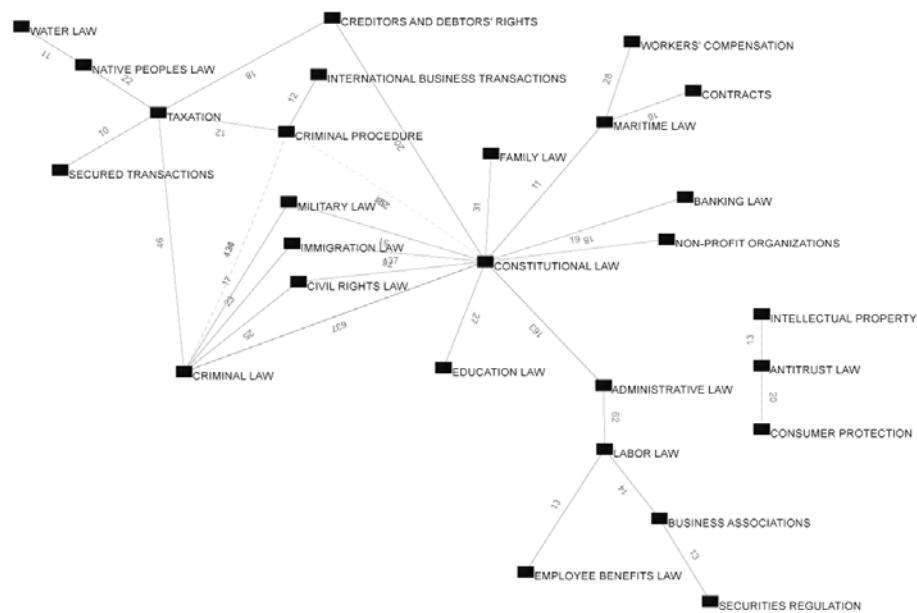


Figure 1. West Topic Space of the United States Supreme Court—Network Layout

Findings, Discussion and Conclusions

Network Graph Approach

Initial network based attempts to create a domain map of the topic space of Supreme Court cases using the spring force layout algorithm in Pajek proved unsatisfying. The procedural and factual topics, which may co-occur with just about any doctrinal topic, pulled everything to the center of the graph.ⁱⁱ In order to derive any insight using this approach, the author had to visualize just the doctrinal topics. Furthermore, to obtain readable visualizations, all of the co-occurrences were aggregated up from the West Topic level to the law school subject level (the course offered in law school most likely to teach that particular topic). The graph was then subjected to another double treatment. First, the most tenuous (least numerous) co-occurrences between subjects were discarded. This was a bit subjective and was again informed by the author's familiarity with legal topics. It was necessary because almost every subject co-occurred with Constitutional Law and a few other similarly ubiquitous topics. Second, amongst the remaining subjects, the graph was thresholded at 10 or more case co-occurrences between the subjects. This resulted in network visualization pictured in Figure 1.

Apparent from the visualization were several counterintuitive adjacencies that reflect the unique jurisdiction of the United States Supreme Court. Maritime cases invoke federal jurisdiction. Furthermore, to the extent that maritime cases involve contract disputes or workers' compensation claims, these issues are heard by the Federal Courts. Outside of the context of maritime law, contracts and workers' compensation cases are state court issues not typically heard by the Federal Courts. Thus, the resultant base map reflects an inherent bias in the dataset. No expert in the law would intuitively co-locate Maritime Law, Workers' Compensation, and Contracts outside the unique context of cases being heard in the Supreme Court.

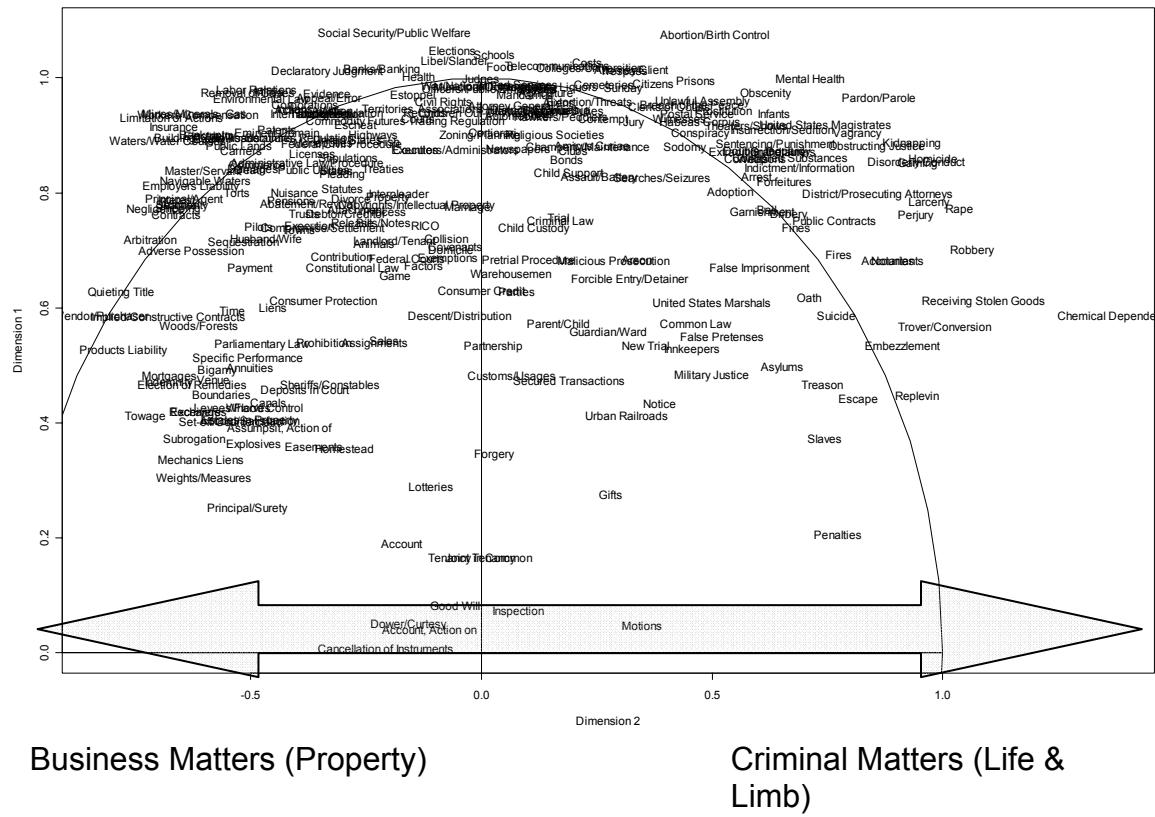


Figure 2. PCA (2nd and 1st Dimensions)

PCA Approach

A plot of the amount of the variance contained in each of the singular values reveals that the first twenty-five dimensions account for almost 4/5ths of the variance. On the whole, the dimensionality plots do not reveal easily identifiable continuums. However, the plot of the 1st and 2nd principal components reveal a readily identifiable continuum between criminal matters on one end (Receiving Stolen Goods, Rape, Robbery, Larceny, Homicide, etc.), and business matters on the other (Quieting Title, Constructive Contracts, Mortgages, etc.). This division between matters of life and limb and those of property corresponds with the popular perception of the justice system as being composed largely of two parts—criminal and non-criminal matters. See Figures 2. This same continuum may also be seen in a non-PCA layout of the topic relationships of one particular Supreme Court term (2004). Cases as nodes are linked to the topics they contain which are also portrayed as nodes. The spatial layout was generated by hand employing the heuristic charge to minimize edge crossings. See Figure 3.

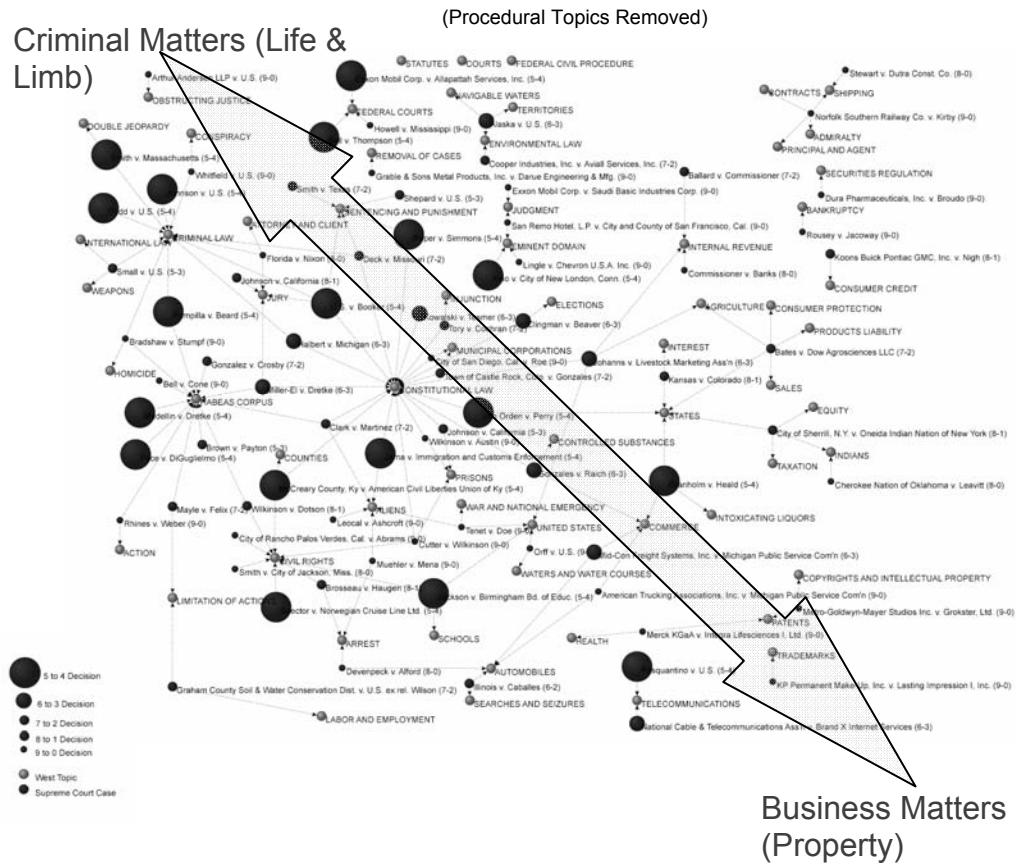


Figure 3. 2004 Supreme Court Term West Topic Space

The layout of topics of the first two principal components revealed topic adjacencies that are contrary to traditional categorizations. For instance, the topic Bigamy, which is a crime, appears on the Business Matters end of the previously identified continuum. This at first appears to be an error. However, further research reveals that the topic Bigamy appears only once in the entire dataset. It occurs in the context of a divorce case in which alimony and the division of marital property were hotly contested. In fact, the alleged bigamy (one spouse got a divorce and remarried in a different state and these actions were not recognized by the original state) was the means to the end of acquiring more marital assets in the divorce proceeding. Thus, the appearance of the topic Bigamy at the Business/Property side of the continuum makes sense even though it is contrary to how a law student would encounter the topic. See Figure 4.

As explained above, the author assigned the West Topics one of three additional labels—doctrinal, procedural and factual. The impetus for this was the fact that when the topic space was rendered using only a node link approach coupled with a spring force algorithm, many of the procedural and factual topics, which may co-occur with just about any doctrinal topic, pull everything to the center of the graph. Fortunately, this appears to be less of a problem using the PCA approach. Interestingly, the layout of the fifth and sixth principal components bears out a separation of the doctrinal (red) and procedural topics (blue) with the factual topics (green) clustered in between the two. See Figure 5.

Furthermore, the layout of the seventh and eighth principal components reveals a continuum between family obligations (Bigamy, Marriage, Estates in Property, Divorce, Homestead, Husband & Wife, etc.) and business obligations (Subrogation, Mechanics Liens, Principal/Surety, Assignments, Payment, etc.). See Figure 6. There were no additional continuums in the remaining first twenty-five principle components that the author could identify.

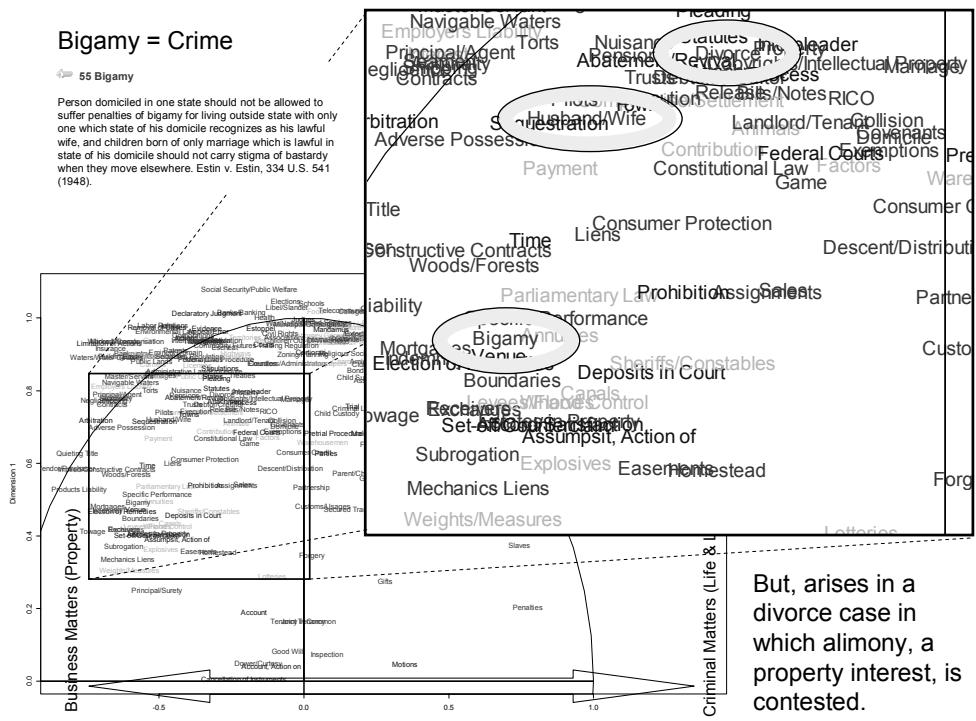


Figure 4. Placement of the Topic Bigamy

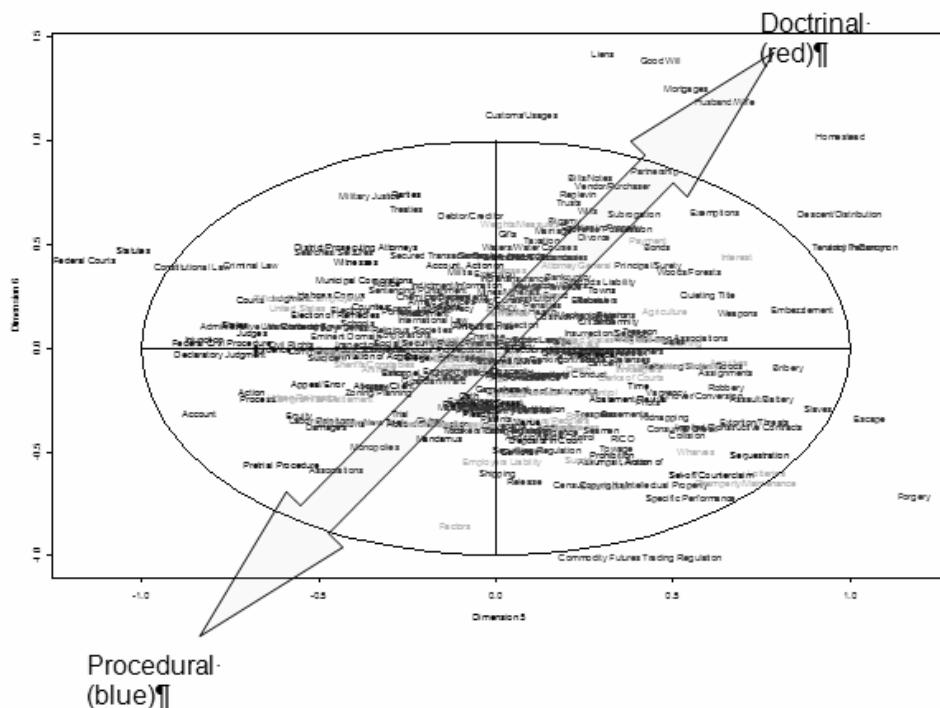


Figure 5. PCA (5th and 6th Dimensions)

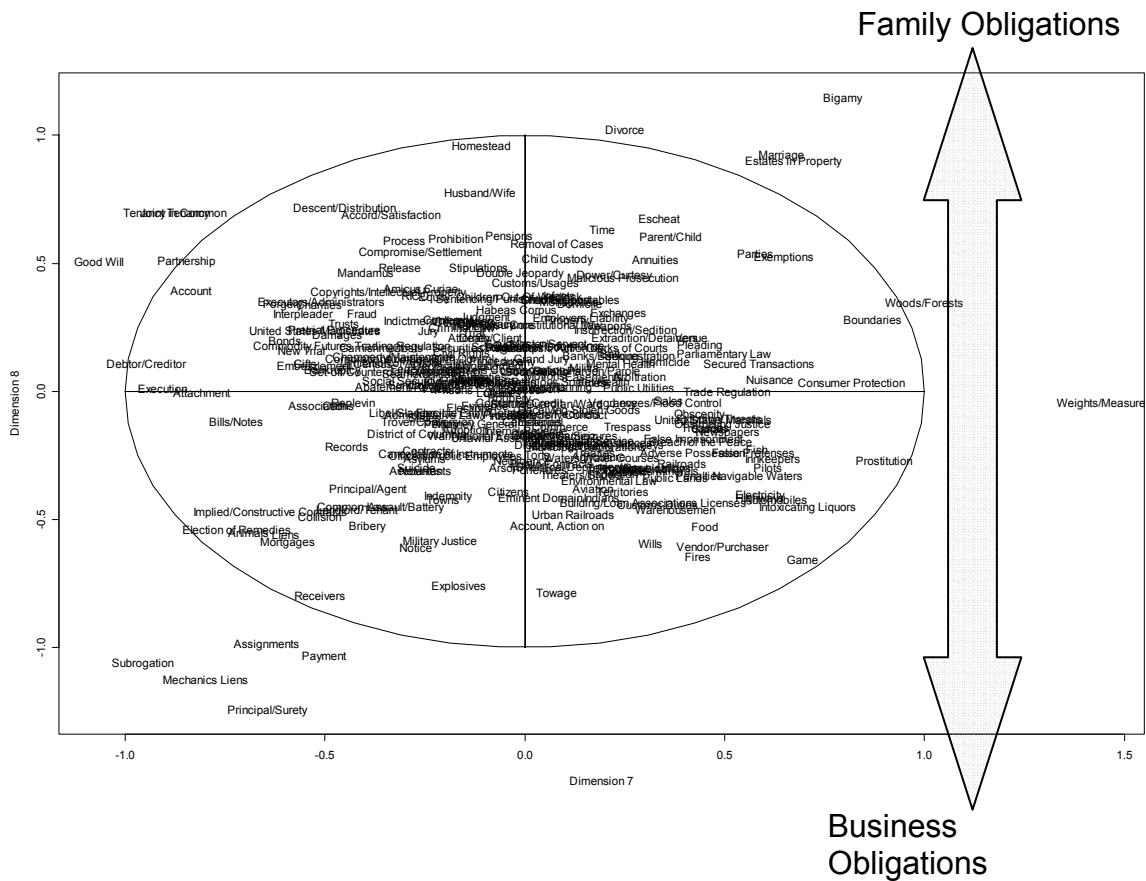


Figure 6. PCA (7th and 8th Dimensions)

PCA Just Doctrinal

As the author's ultimate goal is to create a substrate map based on the layout of topics for use in legal education, he decided to conduct PCA with just the doctrinal topics. To this end, he aggregated the co-occurrence counts up from the West Topic level to the law school subject categories he assigned the doctrinal topics. These additional category assignments are based on where the topic would most likely be encountered in a law school course. The plots of the various dimensions produced adjacencies that were for the most part expected. See Figure 7. The upper left quadrant of the plot nicely co-locates various aspects of the law that deal with land in general (Public Land Law, Water Law, Environmental Law, Oil & Gas Law, Property Law, Native Peoples Law, and Taxation.) The later two course topics most likely cluster in this region because the underlying events that resulted in the cases to which these topics are assigned most likely involved land. Presumably, most of the taxation cases that reach the Supreme Court involve property taxes. Similarly, it is presumed from the PCA analysis that cases assigned the topics involving native peoples most likely also deal with land controversies.

It is similarly consistent with expectations that Criminal Law and Criminal Procedure appear adjacent to one another. These subjects, along with Constitutional Law, are inextricably linked. It is also appropriate that Estate Planning and Wills and Trust are proximate to one another. Taxation would be another subject that the author would expect to cluster closely with Estate Planning and Wills and Trusts. However, as mentioned previously, it is adjacent to topics involving land. It is similarly appropriate that Torts and Products Liability are in the same quadrant. The latter is a subset of the former. Likewise, it is satisfying to one's intuition that Juvenile Justice and Education Law are proximate as both involve children.

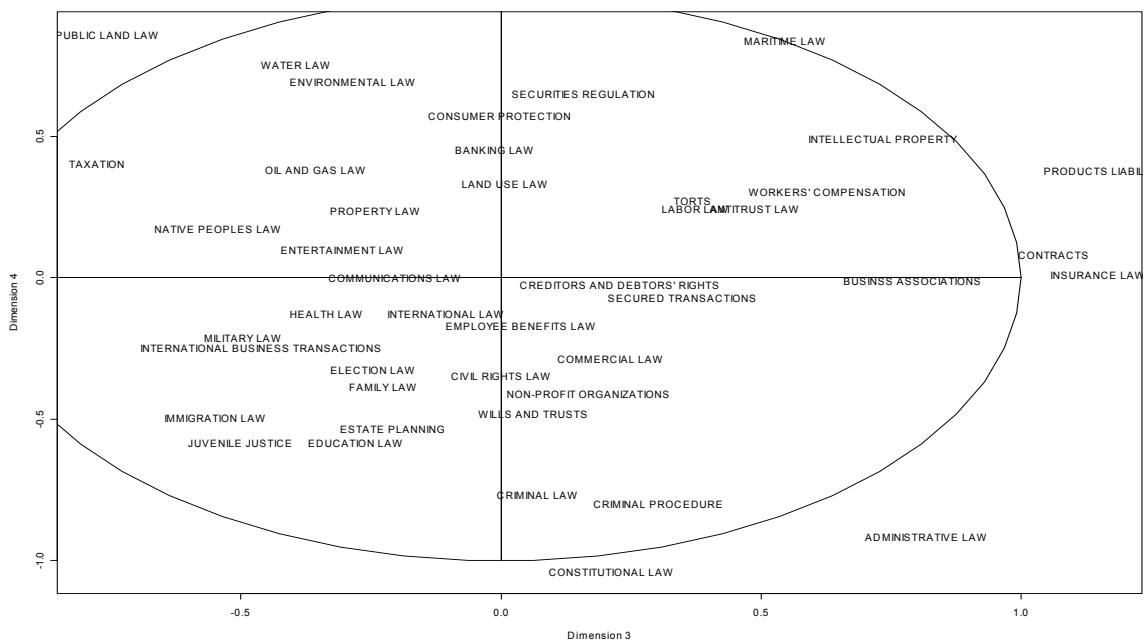


Figure 7. PCA, Doctrinal Subjects Aggregated to Law School Course Offerings

MDS

As the dataset could easily be converted from a matrix of counts, to a matrix of distances, multidimensional scaling (“MDS”) was utilized and to compare the results with those obtained from PCA. Co-occurrence counts were subtracted from the maximum number of co-occurrences (896) to get a measure of the distance between topics. The results were not as satisfactory as those produced by PCA. While some topic adjacencies made sense ((Searches and Seizures, Sentencing and Punishment, Habeas Corpus) and (Admiralty, Shipping, Seaman) most topics were clustered in an unreadable blob at the center of the plot. Similarly, the just doctrinal plots were also less intuitively satisfactory to one trained in the law as were those produced by PCA. See Figure 8.

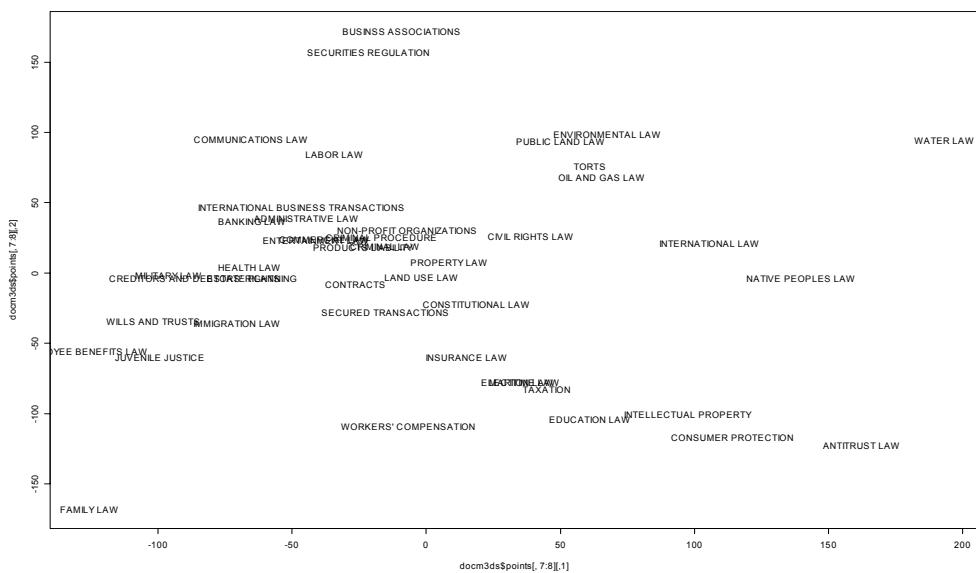


Figure 8. MDS Just Doctrinal (7th & 8th Dimensions)

Conclusion

The multivariate statistical techniques employed in this paper contributed some to the author's ultimate goal of a rigorous topic space substrate map of Supreme Court topics. The techniques reaffirmed the author's intuitive understanding that the primary divide is between business and criminal matters. While the techniques were helpful in suggesting layouts of the super groupings, it will take a fusion approach to produce a substrate map that resonates with how legal topics are conceptualized, bundled, and taught in law school.

Part of the problem is that the approach might be faulty in part. As demonstrated by the 'Bigamy' example, the co-occurrence of topics in cases suggests relationships between topics that though logical, are not how a law student is presented the material. Additionally, part of the problem might also be the need to do the same analysis at an increased level of granularity. While there are only 405 top level West topics, there are about 100,000 sub-topic assignments. The same techniques involving the relationships of all of these sub-topic assignments might produce more intuitively satisfying maps. Also, the fact that the data is only from Supreme Court cases skews the results. Topic co-occurrence data is needed from the two additional levels of Federal Courts and all of the state case material as well. This will require a partnership with the actual database owner to obtain backend access.

References

- Batagelj, V., & Mrvar, A. (1998). Pajek - Program for Large Network Analysis. *Connections*, 21(2), 47-57.
- Battelle, J. (2005). The search : how Google and its rivals rewrote the rules of business and transformed our culture. New York: Portfolio.
- Bernal, J. D. (1939). *The Social Function of Science* London: Routledge & Kegan Ltd.
- Börner, K., Chen, C., & Boyack, K. W. (2002). Visualizing Knowledge Domains. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 37, pp. 179-255). Medford, New Jersey: Information Today.
- Brazill, T. J., & Grofman, B. (2002). Factor analysis versus multi-dimensional scaling: binary choice roll-call voting and the US Supreme Court. *Social Networks*, 24, 201-229.
- Chandler, S. J. (2005). *The Network Structure of Supreme Court Jurisprudence*. Paper presented at the 2005 International Mathematica Symposium. from <http://ssrn.com/abstract=742065>
- Cross, F. B., & Smith, T. A. (In Press). The Reagan Revolution in the Network of Law. from <http://ssrn.com/abstract=909217>
- Cross, F. B., Smith, T. A., & Tomarchio, A. (In press). Determinants of Cohesion in the Supreme Court's Network of Precedents. from <http://ssrn.com/abstract=924110>
- Doyle, J. (1992). Westlaw and the American Digest Classification Scheme. *Law Library Journal*, 84, 229-257.
- Ellingham, H. J. T. (1948, 21 June – 2 July). *Divisions of Natural Science and Technology*. Paper presented at the Royal Society Scientific Information Conference, London: The Royal Society, Burlington House.
- Epstein, L., Martin, A. D., Segal, J. A., & Westerland, C. (2005). The Judicial Common Space. Northwestern University. from <https://www.law.northwestern.edu/faculty/conferences/research/Epstein.pdf>.
- Fowler, J. H. (2006). Connecting the Congress: A Study of Cosponsorship Networks. *Political Analysis*, 14, 456 - 487.
- Fowler, J. H., Johnson, T. R., Spriggs, J. F. I., Jeon, S., & Wahlbeck, P. J. (In Press). Network Analysis and the Law: Measuring the Legal Importance of Supreme Court Precedents. *Political Analysis*
- Garfield, E. (1955). Citation Indexes for Science. *Science*, 122(3159), 108-111.
- Garfield, E. (1979). Citation indexing - its theory and application in science, technology, and humanities. New York: Wiley.
- George, T. E. (2006). An Empirical Study of Empirical Legal Scholarship: The Top Schools. *Indiana Law Journal*, 81, 141-160.
- Grofman, B., & Brazill, T. J. (2002). Identifying the median justice on the Supreme Court through multidimensional scaling: Analysis of "natural courts" 1953-1991. *Public Choice*, 112, 55-79.
- Hopkins, K. (2005). Most Highly Cited. *The Scientist*, 19(20), 22-27.
- Hook, P. A., & Börner, K. (2005). Educational Knowledge Domain Visualizations: Tools to Navigate, Understand, and Internalize the Structure of Scholarly Knowledge and Expertise. In A. Spink & C. Cole (Eds.), *New Directions in Cognitive Information Retrieval* (pp. 187-208). London: Springer.
- Johnson, J. C., Borgatti, S. P., & Romney, K. (2005). Analysis Of Voting Patterns In U.S. Supreme Court Decisions. Paper presented at the Sunbelt XXV, International Sunbelt Social Network Conference. (abstract available at: <http://www.socsci.uci.edu/~ssnconf/conf/SunbeltXXVProgram.pdf>).

- Martin, A. D., & Quinn, K. M. (2002). Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-999. *Political Analysis*, 10(2), 134-153.
- Montello, D. R., Fabrikant, S. I., Ruocco, M., & Middleton, R. S. (2003). Testing the First Law of Cognitive Geography on Point-Display Spatializations. In W. Kuhn, M. F. Worboys & S. Timpf (Eds.), *Spatial Information Theory: Foundations of Geographic Information Science: Proceedings of the Conference on Spatial Information Theory (COSIT '03): Lecture Notes in Computer Science*, 2825 (pp. 316-331). Berlin: Springer Verlag.
- Ogden, P. (1993). "Mastering the Lawless Science of Our Law"; A Story of Legal Citation Indexes. *Law Library Journal*, 85(1), 1-48.
- Paolillo, J. C., & Wright, E. (2006). Social Network Analysis on the Semantic Web: Techniques and Challenges for Visualizing FOAF. In V. Geroimenko & C. Chen (Eds.), *Visualizing the Semantic Web: XML-Based Internet and Information Visualization* (2nd ed., pp. 229-241). London: Springer-Verlag.
- Poole, K. T. (2005). *Spatial Models of Parliamentary Voting*. Cambridge: Cambridge University Press.
- Porter, M. A., Mucha, P. J., Newman, M. E. J., & Warmbrand, C. M. (2005). A Network Analysis of Committees in the U.S. House of Representatives. *PNAS*, 102(20), 7057-7062.
- Pritchett, C. H. (1941). Divisions of Opinion Among Justices of the U.S. Supreme Court, 1939-1941. *The American Political Science Review*, 35(5), 890-898.
- Skupin, A., & Fabrikant, S. I. (2003). Spatialization Methods: A Cartographic Research Agenda for Non-geographic Information Visualization. *Cartography and Geographic Information Science*, 30(2), 95-115.
- Schubert, G. (1962). The 1960 Term of the Supreme Court: A Psychological Analysis. *The American Political Science Review*, 56(1), 90-107.
- Schubert, G. (1963). Judicial Attitudes and Voting Behavior: The 1961 Term of the United States Supreme Court. *Law and Contemporary Problems*, 28, 100-142.
- Sirovich, L. (2003). A pattern analysis of the second Rehnquist U.S. Supreme Court. *PNAS*, 100(13), 7432-7437.
- Small, H., & Griffith, B. C. (1974). The Structure of Scientific Literatures I: Identifying and Graphing Specialties. *Science Studies*, 4(1), 17-40.
- Smith, T. A. (In Press). The Web of Law. from <http://ssrn.com/abstract=642863>
- Snyder, F. (1999). The West Digest System: The Ninth Circuit and the Montana Supreme Court. *Montana Law Review*, 60, 541-597.
- Spaeth, H. J., & Altfeld, M. F. (1985). Influence Relationships within the Supreme Court: A Comparison of the Warren and Burger Courts. *The Western Political Quarterly*, 38(1), 70-83.
- Surrency, E. C. (1990). *A history of American law publishing*. New York: Oceana Publications.
- Thomson/West. (2006). *West's Analysis of American Law* (2006 ed.). St. Paul, MN: Thomson/West.
- Thurstone, L. L., & Degan, J. W. (1951). A Factorial Study of the Supreme Court. *PNAS*, 37(9), 628-635.

ⁱⁱ One reviewer noted the similarity of the problem encountered by Small and Griffith. In the reviewer's own words, this problem was "the effect of methods papers on document co-citation clustering/mapping (these must be removed before a structure can be found -- see any number of papers by Small on this)." (See Small & Griffith, 1974). I wish to thank both unknown reviewers for their comments and feedback.

Determinants for Young Researcher Careers in Germany. Comparative Evaluation of Postdoctoral Programmes

Stefan Hornbostel and Susan Böhmer

hornbostel@forschungsinfo.de, boehmer@forschungsinfo.de
IFQ, Godesberger Allee 90, 53115 Bonn (Germany)

Abstract

Research policy discussions in Germany increasingly focus on the situation of young researchers. Major criticisms include the rigid career paths (Habilitation) and the late independence of young postdoctoral researchers. Some funding organisations consequently offer special programmes/grants for young researchers to enable them to gain early research independence by financing their own research group (generally doctoral students) and so open up the path to professorial appointment without needing a Habilitation. This survey will interview successful and unsuccessful applicants to these funding programmes in order, firstly, to find out how funding affects career patterns, publication activity and academic resonance (citations) and, secondly, to identify factors that can positively or negatively influence a career in research. The second step compares the survey groups (funded young research group leaders from four programmes) to draw conclusions on the effects of the various selection methods or entry requirements. Given the discussion on further structuring the postdoctoral phase, we will attempt to find signs of where and in what fields additional qualification opportunities would make sense and how the actual early independence of young researchers could be achieved.

Keywords:

postdoctoral researchers; postdoctoral programme; careerpath; publication/ citation analysis

Introduction

International comparisons reveal a special feature of the German higher education system: after completing their doctorate (Promotion), a large proportion of young researchers go on to gain a postdoctoral qualification (called the Habilitation) which counts as a prerequisite for appointment to a professorship. The Habilitation includes the requirement to produce a monograph. This system has been in flux now for several years. On the one hand, the introduction of so-called junior professorships created the status of non-tenured professors, while, on the other, the funding of young research group leaders (including positions for doctoral students) opened up an alternative route to professorial appointment. This funding aims to give young research group leaders the earliest possible opportunity to demonstrate through independent research (and corresponding publications) that they have the qualifications needed for professorial appointment, even without a Habilitation. The objective of the funding programmes described here is to select excellent young researchers and to qualify them for a professorship. While it was traditionally only universities that would decide on who to appoint as a professor, responsibility for assessing the competencies, achievements and performance of young research group leaders is moving strongly towards the funding bodies themselves.

Background

One of the most well-known young research group leader programmes in Germany is the Emmy Noether Programme run by the German Research Foundation (DFG). This programme has funded some 380 young researchers since its introduction in 1999. The other major research funding organisations in Germany operate similar funding instruments: Max Planck Society, Volkswagen Foundation, and Helmholtz Association. The programme goals are practically identical, namely to fund excellent researchers, to lower the age on first professorship, to facilitate early independent research, to encourage international networking, and to raise the proportion of women by improving the compatibility of career and family. Since the beginning of 2006, the Institute for Research Information and Quality Assurance (IFQ) has been carrying out an evaluation study on these funding programmes. The first findings will become available in January 2007.

Purpose of the study

Our study aims to check whether and how the programme goals are being achieved and to describe in detail the career paths or typical career patterns that are taken by young researchers. Specifically, this involves:

- Checking the prognostic validity of the peer review system as a means of selecting young research group leaders. This involves carrying out a citation and publication analysis (Basic SCI, SSCI, own publication lists) to examine the publication behaviour of young research group leaders before, during and after the funding term. These data are collected for both rejected and accepted applicants. Information on the assessment of the selection method, socio-demographic variables and on the career patterns are collected via online surveys carried out in both groups. The results of a comparable survey done in Sweden (Melin and Danell, 2006) tell us that the publication behaviour of applicants to such programmes hardly differs, regardless of whether they are funded or rejected. The programmes reveal a high degree of self-selectivity.
Indeed, much seems to indicate that in their assessment reviewers are strongly influenced by the JIF (Journal Impact Factor), although the citations scored do not justify this.
- Online surveys and qualitative interviews with a subpopulation serve to reconstruct the career paths of applicants. The objective is to determine how publication activity, international networking, the funding itself, institutional integration (university / non-university), young research group, and age and sex influence the career pattern (regressive analysis).
- Besides comparing approved and rejected young researchers, our survey design also enables us to examine differences between the programmes offered by the various funding organisations. We expect a reconstruction of the career patterns and the concrete working conditions during the funding to shed light on what aspects of the programmes, the funding conditions and the work situation affect research output. Comparison of the various survey groups will additionally allow us to draw conclusions on the effects of the various selection methods.
- Since these funding programmes are open to all disciplines, we will also compare the effects of the funding between various disciplines (arts, humanities and social sciences versus life sciences and natural sciences).
- Based on the findings of this survey we intend to deduce recommendations for an advancement of the Postdoctoral Programmes themselves.

Methods

To analyse the field of investigation in the greatest possible detail, our study will use the advantages provided by various methods. Our starting point is a current online survey (total survey) which will collect data on educational and career biography, on funding process, on working conditions, etc. The second step will involve problem-centred interviews with a selection of Emmy Noether-Junior Research Group Leaders. Suitable interview partners will be selected on the basis of the data produced by the quantitative survey. A further qualitative survey will interview people from the institutional setting of the funded researchers, through which we expect to obtain information on how the young group is integrated at the universities and on the status of the young researchers. The quantitative and qualitative surveys will be complemented by the results obtained from the analysis of works published by rejected and approved applicants to the Emmy Noether Programme. By means of this methodological triangulation, we aim to be able to study how career biography patterns, specific working conditions, funding decision itself, and the output behaviour of the survey participants and interviewees interact.

Table 1 shows that we will already be able to present comprehensive analyses of the quantitative data and bibliometrical results for the random sample of Emmy Noether Junior Research Group Leaders at the ISSI Conference in Madrid 2007. In June 2007, we will additionally be able to present the first results from the interviews and initial figures from surveys carried out among the control groups. Parallel to the data analyses, we will also characterise the selection methods used by the various funding organisations on the basis of documentary analyses and interviews held with experts. These will also be included in the initial analyses.

Table 1. Survey Groups and Time Periods

Survey Groups		Quantitative Survey	Qualitative Survey	Citation Analyses
<i>Emmy Noether Junior Research Group Leaders (DFG)</i>	<i>Approved (N=378)</i>	Oct-Dec 2006	Feb-Jun 2007	Sep 2006-Jun 2007
	<i>Rejected (N=350)</i>	Jan-Mar 2007		Sep 2006-Jun 2007
	<i>Institutional Setting</i>		Apr-Jul 2007	
<i>Junior Research Group Leaders (Volkswagen Foundation)</i>	<i>Approved (N=67)</i>	Feb-May 2007		Feb-Jul 2007
	<i>Rejected (N=129)</i>	Feb-May 2007		Feb-Jul 2007
<i>Young Investigator Group Leaders (Helmholtz Association)</i>	<i>Approved (N=52)</i>	Apr-Jun 2007		Jul-Dec 2007
	<i>Rejected (N=89)</i>	Apr-Jun 2007		Jul-Dec 2007
<i>Junior Research Group Leaders (Max Planck Society)</i>	<i>Currently funded (N=53)</i>	Apr-Jun 2007		Jul-Dec 2007

Findings

Our analyses will address the following questions:

- Validity of the selection method: How good are peers at predicting future research success?
- Influence on the career pattern: What are the key determinants for successful appointment to a professorship?
- To what extent do rejected (non-funded) young researchers look for and find alternatives or positions abroad?
- What influences research output? How is career development influenced by personal qualifications and achievements, funding programme, sex, institutional setting, subject, subfield, etc?
- Structured funding: What aspects should and can structured funding cover during the postdoctoral phase?

We will do this by analysing information on the career pattern prior to and after the funding or application, publication data (plus other output indicators, such as amount of external funding raised, patent registrations, spin-offs) by the funded researchers before, during and after the funding as well as by rejected researchers before and after application. By combining the three different research methods we expect to obtain more detailed insight into how selection processes, career paths/working conditions and publication behaviour of the interviewed/surveyed persons interact with each other.

Moreover, we will be able to differentiate the relatively vague term of a young researcher's "success" in slightly greater detail 1) by mapping the directness or heterogeneity of career patterns, 2) by determining the time and circumstances of the first appointment, 3) by examining alternative career paths, e.g. in non-university facilities or in business, industry or administration, 4) by analysing the compatibility of career and family as a "success factor", and 5) by using publication and citation analyses to "measure" the productivity and visibility of researchers. In connection with this, we hope on the basis of our findings to contribute valuably to the discussion on success indicators in research.

Discussion/Conclusion

Our current study focuses on the career patterns and success indicators of young researchers who are taking an alternative path to their aspired appointment to a professorship. The study endeavours to contribute substantially to describing the situation of young researchers in Germany and their position in the international field. Direct comparisons with studies carried out in other countries (e.g. MELIN & DANELL 2006; JANSON et al. 2006; LANGFELDT & BROFOSS 2005) serve, not least, to help identify and analyse special national features. The combination of three research methods (quantitative surveys, qualitative interviews, publication/citation analyses) aims to enhance the evidential value of the results and to help explain contexts that were previously difficult to identify and clarify.

References

- Büchtemann, Christoph F. (2001): Deutsche Nachwuchswissenschaftler in den USA. Perspektiven der Hochschul und Wissenschaftspolitik. Bonn: BMBF. Retrieved October 31, 2006 from: <http://www.bmbf.de/pub/talent.pdf>.
- Commission of the European Communities (2003): Researchers in the European Research Area. One Profession, Multiple Careers. Retrieved October 31, 2006 from: http://www.iua.ie/core_activities/research/pdf/careercommunication.pdf.
- Enders, Jürgen & Bornmann, Lutz (2001): *Karriere mit Doktortitel? Ausbildung, Berufsverlauf und Berufserfolg von Promovierten*. Frankfurt a.M./New York: Campus.
- Enders, Jürgen & Mugabushaka, Alexis-Michel (2004): Wissenschaft und Karriere. Erfahrungen und Werdegänge ehemaliger Stipendiaten der DFG. Bonn: Deutsche Forschungsgemeinschaft. Retrieved October 31, 2006 from: http://www.dfg.de/dfg_im_profil/Zahlen_und_fakten/statistisches_berichtswesen/stip2004/download/dfgstip_ber_04.pdf.
- Janson, Kerstin, Schomburg, Harald & Teichler, Ulrich (2006): Wissenschaftliche Wege zur Professor oder ins Abseits? Strukturinformationen zu Arbeitsmarkt und Beschäftigung an Hochschulen in Deutschland und den USA. Kassel: INCHER. Retrieved October 31, 2006 from: http://www.gain-network.org/file_depot/010000000/100000000/16468/folder/44145/INCHER+Studie+zum+wissenschaftlichen+Arbeitsmarkt.pdf.
- Langfeldt, Liv & Brofoss, Karl Erik (2005): Evaluation of the European Young Investigator Awards Scheme. Oslo, NIFU STEP Working Paper 10/2005. Retrieved October 31, 2006 from: http://www.nifustep.no/content/download/15200/88306/file/Arbeidsnotat_10-2005.pdf.
- LeMoulour, Isbelle, Lenecke, Kerstin & Schomburg, Harald (ed.) (2005): Human Resources in Research & Development. Monitoring System on Career Path and Mobility Flows. Retrieved October 31, 2006 from: http://www.sister.nu/pdf/MOMO_REPORT11.pdf.
- Melin, Göran & Danell, Rickard (2006): Effects of Funding Young, Promising Scientists. Retrieved October 31, 2006 from: http://www.fteval.at/papers06/success_1.htm.
- Melin, Göran (2005): The dark side of mobility: negative experiences of doing a postdoc period abroad. *Research Evaluation*, 14, 3, Dec. 2005, pages 229-237.
- Melin, Göran (2004): Postdoc abroad: inherited scientific contacts or establishment of new networks?. *Research Evaluation*, 13, 2, Aug. 2004, pages 95-102.
- Mugabushaka, Alexis-Michel, Rahlf, Thomas & Gündler, Jürgen (2006): Antragsaktivität und -erfolg von Juniorprofessoren bei der Deutschen Forschungsgemeinschaft. (DFG Infobrief 1/2006). Retrieved October 31, 2006 from: http://www.dfg.de/dfg_im_profil/zahlen_und_fakten/statistisches_berichtswesen/ib/download/
- OECD (2002): International Mobility of the Highly Skilled (Policy Brief July 2002). Retrieved October 31, 2006 from: <http://www.oecd.org/dataoecd/9/20/1950028.pdf>.
- Rössel, Jörg & Landfester, Katarina (2004): Die Juniorprofessur und das Emmy Noether-Programm. Eine vergleichende Evaluationsstudie. Retrieved October 31, 2006 from: http://www.diejungeakademie.de/pdf/Juniorprofessur_%20und_Emmy_Noether.pdf.
- Stifterverband für die Deutsche Wissenschaft (2002): Brain Drain – Brain Gain. Eine Untersuchung über internationale Berufskarrieren. (Durchgeführt von der Gesellschaft für Empirische Studien: Beate Backhaus, Lars Ninke, Albert Over). Retrieved October 31, 2006 from: <http://www.ges-kassel.de/download/BrainDrain-BrainGain>.

Are Multi-Authorship and Visibility Related? Study of Ten Research Areas at Carlos III University of Madrid

Isabel Iribarren-Maestro, María Luisa Lascurain-Sánchez, Elías Sanz-Casado

*{iiribarr, esmlascura, elias} @bib.uc3m. @bib.uc3m.es
Carlos III University of Madrid, Department of Library Science and Documentation, Getafe 28903,
Madrid and Laboratory of Information Metrics Studies (LEMI) (Spain)*

Abstract

Opinions in the literature on the possible relationship between co-authorship and number of citations vary. This paper contributes to the debate with a further analysis of the subject, taking account of the number and quality of citations found for multi- (author, institution, country) and single-authored papers. The study is based on the scientific production of ten Carlos III University of Madrid departmental areas between 1997 and 2003 as reflected in the ISI Web of Science, and the number of times the respective papers were cited between 1997 and 2004. Univariate multifactorial analysis of variance (ANOVA) was used to verify the relationship between multi-authorship and visibility. The correlation between multi-institutional and multi-national authorship and the quartile of the citing journals was analyzed with correspondence analysis. The results show that while multi-institutional and multi-national authorship raise the number of citations, co-authorship and number of citations are unrelated. Correspondence analysis failed to show any correlation between the quartile of the citing journal and multi-institutional or multi-national authorship, but did reveal a relationship between citing journal quartile and departmental area.

Keywords

visibility; collaboration; citation analysis; Impact Factor

Introduction

A positive correlation between multi-authorship at whatever level (author, institution, country) and visibility in the scientific community has been reported in a number of the studies exploring the possibility of such a relationship. Bordons and Gómez (2000) adopted the premise that several authors and groups sharing ideas, technology and experience should generate higher quality papers than a single author working alone.

A similar stand was taken by Glänzel and Schubert, who sustained that studies involving international cooperation call for greater effort than research conducted in a single country, a fact that may imply higher quality. This, in turn, would grow the number of citations (Glänzel, 2000 and Glänzel and Schubert, 2001). Other authors have shared this opinion: Beaver (1986) reported that on average, multi-authored research is more visible than the single-authored variety in terms of number of citations, and that the former tends to be of a higher quality than the latter; Garfield, in turn, argued that hypothetically multi-authorship guarantees higher quality due to the prior peer reviewing inherent in research by two or more scholars (Garfield, 1986) and that multi-national or multi-institutional articles produce a greater impact (Garfield, 1996). In an extensive review of the literature on the relationship between multi-authorship and impact factors, Bordons and Gómez (2000) cited several articles that found a positive correlation between the two variables (Bridgstock, 1991; Bordons, García Jover and Barrigón, 1993, Katz and Hicks, 1997, Van Raan, 1997).

Not all scientists accept this notion, however. Some have associated the larger number of citations with the substantial rise in self-citations implicit in multi-authorship, along with the citations by each of the co-author's colleagues (Rousseau 1992; Leimu and Koricheva, 2005). Other authors have questioned the existence of a relationship between multi-authorship and quality, observing that there are no significant differences in the number of citations between single- and multi-authored papers (Oromaner, 1975; Avkiran, 1997). Several experts have explained the lack of any such correlation on the grounds that in some cases multi-authoring is based on a mentor-apprentice relationship rather than on peer partnering between researchers or large research groups (Bayer, 1982).

In light of the diversity of opinion on the dependence between these variables, the present paper poses yet another analysis of the association between multi-authorship and visibility. In this case, the study is based on the scientific production, listed on the ISI Web of Science, generated by ten Carlos III University of Madrid departmental areas between 1997 and 2003.

Objectives

The two chief objectives pursued in this study were:

- To analyze whether multi-authorship affects the number of citations.
- To analyze whether multi-authorship affects citation quality (defining such quality to be equivalent to citing journal impact).
-

Methodology

The pre-research for this study included a review of all the papers listed on the Web of Science and having at least one author with UC3M affiliation. The papers so identified were then grouped by department and a series of departmental areas were established to account for variations in university structure. After an analysis of the presence of each area's production in ISI databases, the ten whose activity was seen to be most optimally reflected were selected. These areas were: "Materials Science and Engineering and Chemical Engineering" - INGMAT; "Economics" - ECO; "Business Economics" - EMP; "Statistics and Econometrics" - EST; "Physics" - FIS; "Computer Science" - INF; "Electrical, Electronic and Robot Engineering" - INGEEAU; "Mechanical Engineering" - INGMEC; "Mathematics" - MAT; and "Communications Technology" - TECCOM.

The 1997-2003 timeframe was chosen because it concurs with a period when this relatively new institution was consolidating its reputation.

The indicators used to reach the objectives proposed are listed below:

- Relationship between multi-authorship (authors, institutions, countries) and the number of paper citations. Acknowledging the possibility that in addition to multi-authorship in scientific papers, area/department prestige may also have a bearing on the number of citations, the methodology proposed by Avkiran (1997) – who compared multi- and single-authored paper citations – was extended to include the variable "department".
The study was conducted using univariate multifactorial analysis of variance (ANOVA). With this technique, the variability in the results of an experiment can be broken down into independent components, which is particularly useful where more than one input (independent variables) affects the output (dependent variable) (Pérez López, 1996). In addition, it can be used to study the behaviour of the dependent variable in the various groups established by combining the values of the independent variables (Ferrán Aranaz, 1997).
In the present study ANOVA was used to explore whether the level of multi-authorship and the identity of the department producing the papers (independent variables) affect the number of paper citations (dependent variable); three ANOVA trials were run, one for each level of participation (country, institution, author).
- Relationship between multi-institutional and multi-national authorship and the quartile of the journals citing the results of UC3M research between 1997 and 2004. The journal Impact Factor in the year of publication was found for each paper, analyzing not only the journals publishing Carlos III University of Madrid papers, but the entire category in which they were classified to establish each year's quartile. For the intents and purposes of the study, journals classified in more than one category – and in different quartiles in each – were regarded to be in the highest quartile.
Correspondence Analysis (CA) was used for this stage of the study. This method deduces the relationships between different categories by defining their similarities and grouping them accordingly (Carrasco and Hernán, 1993). The CA values obtained were plotted on bubble charts where, in addition to the similarities between variables, a third measure is shown, namely the relative weight acquired by each value upon analysis.

Results and discussion

The results obtained from the analysis of bibliometric indicator data are shown below. First, the results obtained by correlating multi-authored papers and the number of citations are discussed. This is followed by a review of the visibility of such cooperative research measured in terms of the quartile of the publications where the respective papers were cited.

Multi-authorship vs No. of citations

Statistical analysis (ANOVA) was conducted to explore the relations between multi-authorship (authors, institutions, countries) and the number of citations and establish the existence or otherwise of correlations among the different variables. The results of testing the hypothesis that the variables “departmental area” and/or “number of participating countries/institutions/authors” have no effect on the number of paper citations were as follows:

Table 1. ANOVA results: citations vs multi-national authorship and departmental area

Variable	Degress of Freedom	F	Significance
Corrected Model	19	6,63	0,00
Intercept	1	172,86	0,00
DEPARTMENT	9	10,38	0,00
MULTI-NATIONAL	1	5,77	0,016
DEPARTMENT * MULTI-NATIONAL	9	0,705	0,704
Error	1456		
Total	1476		
Corrected Total	1475		

Computed using alpha = ,05
R Squared = ,080 (Adjusted R Squared = ,068)

The value obtained in the multi-national authorship trial (Table 1) revealed that both the departmental area and the number of participating countries had a bearing on the number of citations: $F_{9;1456;DPT} > F_{9;1456;0.05}$ and $F_{1;1456;MULTI-NATIONAL} > F_{1;1456;0.05}$. On the contrary, no significant interaction was detected between the two variables (departmental area and country): $F_{9;1456;DPT}$ and $MULTI-NATIONAL < F_{9;1456;0.05}$.

Table 2. ANOVA results: citations vs multi-institutional authorship and departmental area

Variable	Degress of Freedom	F	Significance
Corrected Model	19	6,46	0,00
Intercept	1	153,237	0,00
DEPARTMENT	9	7,33	0,00
MULTI-INSTITUTIONAL	1	4,185	0,041
DEPARTMENT * MULTI-INSTITUTIONAL	9	0,612	0,787
Error	1456		
Total	1476		
Corrected Total	1475		

Computed using alpha = ,05
R Squared = ,078 (Adjusted R Squared = ,066)

The ANOVA value obtained in the multi-institutional authorship trial (Table 2) indicated that both the departmental area and the number of participating countries had a bearing on the number of citations: $F_{9,1456;DPT} > F_{9,1456;0,05}$ and $F_{1,1456;MULTI-INSTITUTIONAL} > F_{1,1456;0,05}$.

As in the preceding analysis, no significant interaction was found between the two variables (area/department and institution): $F_{9,1456;DPT}$ and $MULTI-INSTITUTIONAL < F_{9,1456;0,05}$. These results concurred with data reported by Persson, Glänzel and Danell (2004), who compared two studies on citations, one conducted 20 years after the other. They found both a greater degree of multi-national authorship and more citations in the more recent of the two. Garfield (1996), along with the authors of three other studies focused on specific countries (Bordons, García Jover and Barrigón, 1993; Katz and Hicks, 1997; Van Raan, 1997), also observed that multi-national or multi-institutional articles produced a greater impact.

Table 3. ANOVA results: citations vs multi-author authorship and departmental area

Variable	Degress of Freedom	F	Significance
Corrected Model	18	6,46	0,00
Intercept	1	31,159	0,04
DEPARTMENT	9	5,422	0,00
MULTIAUTHOR	1	0,019	0,890
DEPARTMENT * MULTIAUTHOR	8	0,269	0,976
Error	1457		
Total	1476		
Corrected Total	1475		

Computed using alpha = ,05
R Squared = ,074 (Adjusted R Squared = ,062)

The assessment conducted at the author level (Table 3) revealed that the identity of the departmental area producing the paper had an effect on the number of citations: $F_{9,1456;DPT} > F_{9,1456;0,05}$, whereas co-authorship had no significant bearing on the dependent variable: $F_{1,1456;MULTIAUTHOR} < F_{1,1456;0,05}$. Finally, the interaction between the two variables (area/department and author) was not significant in this case either: $F_{9,1456;DPT}$ and $MULTIAUTHOR < F_{9,1456;0,05}$. Avkiran (1997) and Oromaner (1975) reported similar results, observing no differences in single- and co-authored paper citations, unlike Abt (1984), Lawani (1986) and Beaver (1986), who did report such a relationship. Lawani even proposed co-authorship as a measure of quality.

With the above confirmation that the number of citations was affected not only by the departmental area responsible for the papers, but by the number of countries and number of institutions involved as well, the next step was to analyze whether such multi-authorship was related to the JCR ranking (quartiles) of the journals where the papers were cited.

(Multi-national and multi-institutional) authorship vs citing journal quartile

Correspondence analysis is a multi-dimensional indicator that provides information on the relationship between two variables. Here it was used to determine the relationship between the type of multi-institutional authorship (national or international) of papers published by the various research departments and the visibility of the journals in which they were cited (citing journal quartile).

The CA in Figure 1 shows the relationships between papers, classified by whether one or several institutions or one or several countries were involved in their preparation, and the quartiles of the journals where they were cited.

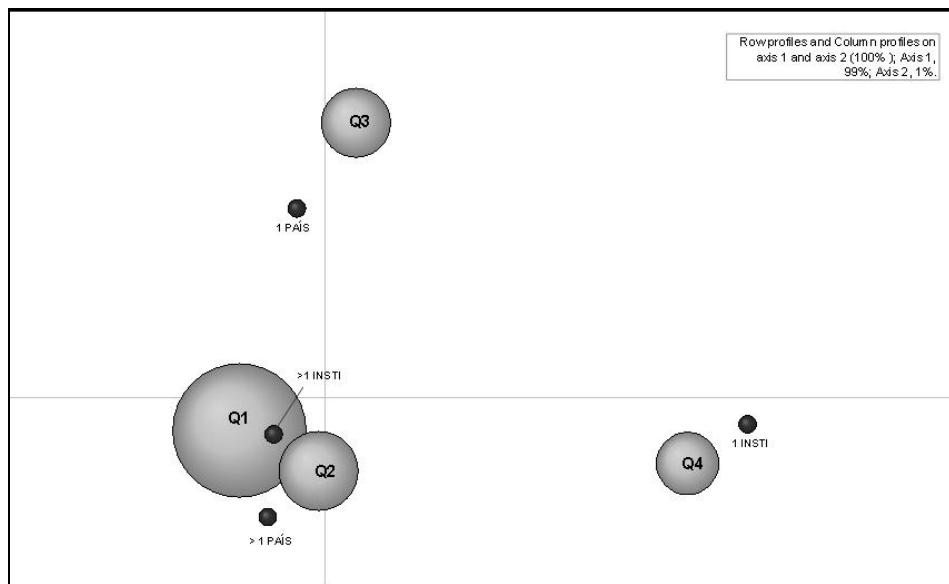


Figure 1. CA, papers by type of mulit-authorship vs citing journal quartile

According to the graph in Figure 1, multi-institutional and multi-national papers were cited primarily in first and second quartile journals, with a preponderance of citations in journals classed in the quartile with the highest visibility. In the case of multi-institutional papers from a single country, the visibility of the citing papers was smaller, as they were published in journals scattered across the first three quartiles. When no multi-institutional authorship was involved, most of the citations were in journals in the least visible (fourth quartile). These same variables were studied at the departmental area level to verify whether the above pattern was found in all the areas analyzed. The results are shown in Figure 2 for multi-institutional, and Figure 3 for multi-national authorship.

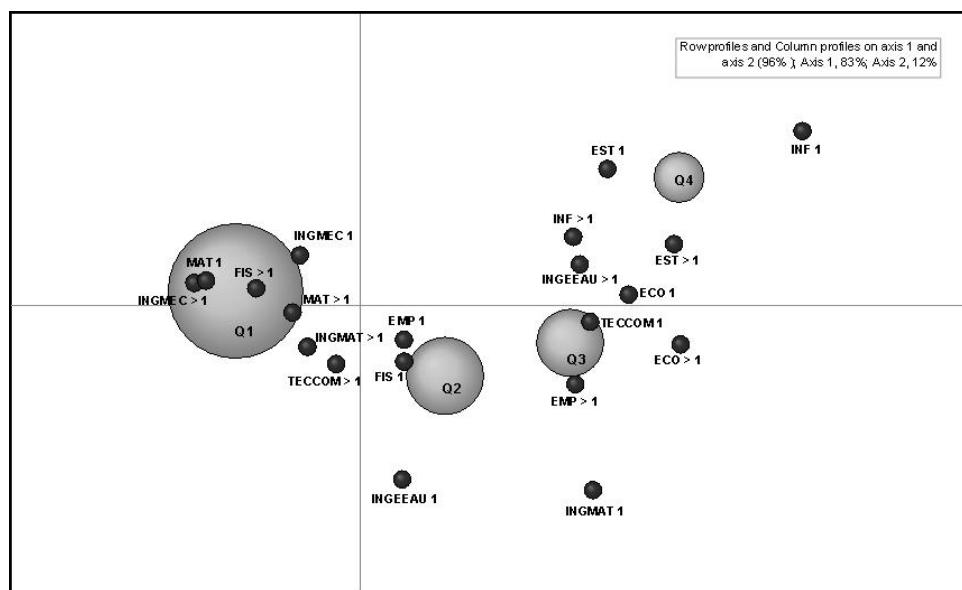


Figure 2. CA, papers involving multi- or single-institutional authorship by departmental area vs citing journal quartile

This figure shows that multi-institutional authorship is not closely related to citations in journals in the highest quartiles, for the papers found in the area around Q1 have both single- and multi-institutional authors. Moreover, all the Mathematics (MAT) and Mechanical Engineering (MEC) papers are positioned around Q1, along with all the multi-institutional papers produced by the Physics (FIS),

Materials (INGMAT) and Communications Technology (TECCOM) departmental areas. Q2 is located close to the y axis, an indication that its behaviour is similar in connection with all the groups of papers, as well as with intra-institutional Business Economics (EMP) and Physics (FIS) areas papers. Finally, quartiles 3 and 4 are surrounded by papers authored by only one as well as papers involving two or more institutions.

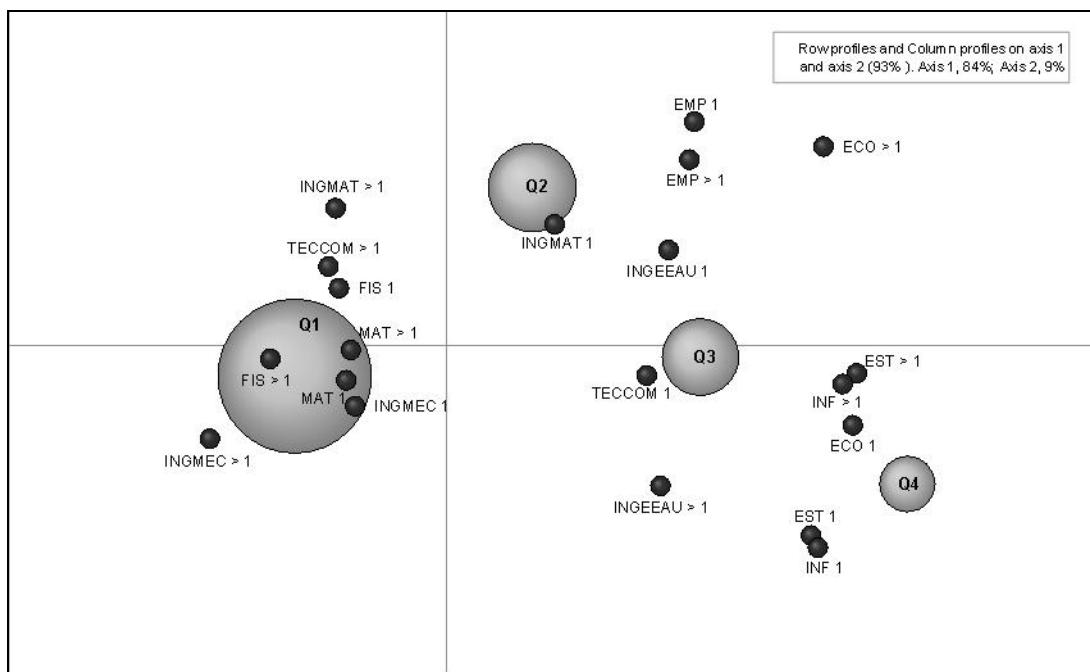


Figure 3. SCA, papers involving national or multi-national authorship by departmental area vs citing journal quartile

Figure 3 shows the relationship between papers with and without multi-national authorship and the visibility of the citing publications. The papers published by certain UC3M departments, regardless of whether they were authored multi-nationally, were cited in highly visible journals. This was the case of the Physics (FIS), Mathematics (MAT) and Mechanical Engineering (INGMEC) Departments. In other areas, such as Communications Technology (TECCOM) or Materials (INGMAT), a strong relationship was observed between multi-national authorship and visibility in first quartile journals. In others, however, closer to the social sciences (Statistics (EST), Economics (ECO), Business Economics (EMP) and Information Technology (INF), multi-national authorship was found to be unrelated to impact in terms of citations in highly visible journals.

Conclusions

Different conclusions can be drawn from the findings described above about the relationship between multi-authorship and visibility of the papers published by Carlos III University of Madrid researchers. Multifactorial ANOVA analysis confirmed that both the number of countries and the number of institutions involved in the preparation of a paper have a positive effect on the number of citations. These results concur with the findings reported by other authors (Bordons and Gómez, 2000; Glänzel, 2000; Glänzel and Schubert, 2001; Garfield, 1986; 1996), according to which multi-institutional and multi-national authorship enhances paper quality. The explanation advanced is that the science contained in these studies is the result of discussion among researchers with different ideas, knowledge and methodologies, and often even working in different disciplines.

Nonetheless, in this paper no relationship was observed between the number of authors signing a paper and the number of times it was cited; consequently, the involvement of more than one author did not entail a larger number of citations. That notwithstanding, sight should not be lost of the fact that the present analysis was focused on an academic institution, so many of the multi-authored papers were the outcome of researcher training. Hence, although the analysis addressed the multi-authorship

variable only (considering multi-institutional and multi-national papers as well), many multi-authored papers involved no actual (national or international) multi-institutional cooperation. Young researcher training is one of universities' basic responsibilities. In this regard, Carlos III University is a young institution, where many research teams are in the consolidation stage and departments are constantly growing: therefore, in some of the papers, multi-authorship is indicative more of researcher training than of cooperation among two or more experienced scientists.

The analysis of the relationship between multi-institutional and multi-national authorship on the one hand and citing journal visibility on the other showed that such multi-authored papers were cited more often in journals in the first two quartiles (Q1 and Q2) than intra-institutional or national papers. These findings lead to the general conclusion that the variables multi-institutional / multi-national authorship and visibility are positively correlated, whereby the existence of the former entails an increase in the latter. Nonetheless, when the Carlos III University departments were analyzed individually, the scientific quality of some of them was found to be independent of the existence or otherwise of multi-authorship. Indeed, most of their papers were routinely cited in journals in the upper positions of the respective Journal Citation Reports subject categories. Conversely, papers published by other departments were not cited in journals with the highest visibility even when multi-institutional or multi-national authorship was involved.

References

- Abt, H. (1984). Citations to single and multiauthored papers. *Publications of the Astronomical Society of the Pacific*, 96 (583), 746-49.
- Avkiran, N. K. (1997). Scientific collaboration in finance does not lead to better quality research. *Scientometrics*, 39 (2), 173-84.
- Bayer, A. E. (1982). A bibliometric analysis of marriage and family literature. *Journal of Marriage and the Family*, 44 (3), 527-38.
- Beaver, D. d. B. (1986). Collaboration and teamwork in Physics. *Czech. J. Phys.*, 36, 14-18.
- Bordons, M., García Jover, F. & Barrigón, S. (1993). Is collaboration improving research visibility? *Research Evaluation*, 3 (1), 19-24.
- Bordons, M. & Gómez, I. (2000). Collaboration Networks in Science. In: B. Cronin & H. B. Atkins (Eds.): *The Web of Knowledge: A festschrift in honour of Eugene Garfield*. New Jersey: ASIS, pp. 197-213.
- Bridgstock, M. (1991). The quality of single and multiple authored papers - an unsolved problem. *Scientometrics*, 21 (1), 37-48.
- Carrasco, J. L. & Hernán, M. A. (1993). *Estadística multivariante en las ciencias de la vida*. Madrid: Cibest; Ciencia 3.
- Ferrán Aranaz, M. (1997). *SPSS para Windows: programación y análisis estadístico*. Madrid: McGraw-Hill.
- Garfiel, E. (1986). Which medical journals have the greatest impact? *Annals of Internal Medicine*, 105, 313-20.
- Garfield, E. (1996). Fortnightly Review: How can impact factors be improved? *British Medical Journal*, (313), 411-13.
- Glänzel, W. (2000). Science in Scandinavia: a bibliometric approach. *Scientometrics*, 48 (2), 121-50.
- Glänzel, W. & Schubert, A. (2001). Double effort = double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, 50 (2), 199-214.
- Katz, J. S. & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. In: B. C. Peritz & L. Egghe. *Proceedings of the Sixth Conference of the International Society for Scientometrics and Informetrics*. (Jerusalem). Israel: AHVA Coop. Printing Press Ltd., 163-75.
- Lawani, S. M. (1986). Some bibliometric correlates of quality in scientific research. *Scientometrics*, 9, (1-2), 13-25.
- Leimu, R. & Koricheva, J. (2005). Does scientific collaboration increase the Impact of Ecological Articles? *BioScience*, 55 (5), 438-43.
- Oromaner, M. (1975). Collaboration and impact - career of multi-authored publications. *Social Science Information sur les Sciences Sociales*, 14 (1), 147-55.
- Persson, O., Glänzel, W. & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*, 60 (3), 421-32.
- Pérez López, C. (1996). *Econometría y análisis estadístico multivariante con Statgraphics: técnicas avanzadas*. Madrid: RA-MA.
- Rousseau, R. (1992). Why am I not cited or why are multiauthored papers more cited than others. *Journal of Documentation*, 48 (1), 79-80.
- Van Raan, A. F. J. (1997). Science as an international enterprise. *Science and Public Policy*, 24 (5), 290-300.

A Hybrid Mapping of Information Science¹

Frizo Janssens*, Wolfgang Glänzel**,** and Bart De Moor*

**Frizo.Janssens@esat.kuleuven.be, Bart.DeMoor@esat.kuleuven.be*

K.U. Leuven, Dept. of Electrical Engineering ESAT-SCD, Kasteelpark Arenberg 10, B-3001 Leuven (Belgium)

***Wolfgang.Glanzel@econ.kuleuven.be*

K.U. Leuven, Steunpunt O&O Indicatoren, Dekenstraat 2, B-3000 Leuven (Belgium)

****glaenzw@iif.hu*

Hungarian Academy of Sciences, ISPR, Nádor u. 18, H-1051 Budapest (Hungary)

Abstract

Previous studies have shown that hybrid clustering methods that incorporate textual content and bibliometric information can outperform clustering methods that use only one of these components. In this paper we apply a hybrid clustering method based on Fisher's inverse chi-square to integrate full-text with citations and to provide a mapping of the field of information science. We quantitatively and qualitatively assess the added value of such an integrated analysis and we investigate whether the clustering outcome is a better representation of the field by comparing with a text-only clustering and with another hybrid method based on linear combination of distance matrices. Our dataset consists of almost 1000 articles and notes published in the period 2002–2004 in 5 representative journals. The optimal number of clusters for the field is 5, determined by using a combination of distance-based and stability-based methods. Term networks present the cognitive structure of the field and are complemented by the most representative publications. Three large traditional sub-disciplines, particularly, information retrieval, bibliometrics/scientometrics and more social aspects, and two smaller clusters about patent analysis and webometrics, can be distinguished.

Keywords

mapping of science; hybrid clustering; text mining; fisher's inverse Chi-square method; information science.

Introduction

The long-term goal of this research is an accurate unsupervised clustering of science or technology fields, towards the detection of new emerging fields or hot topics. The idea of combining bibliometric or citation information with textual content is not new for it has already been pursued to obtain improved performance in information retrieval (e.g., Calado et al., 2003), bibliometric mapping of science (Mullins et al., 1988; Snizek et al., 1991; Braam et al., 1991; Glenisson et al., 2005; Janssens et al., 2006b), clustering (a.o., Modha & Spangler, 2000; Wang & Kitsuregawa, 2002), and classification (e.g., Joachims et al., 2001; Calado et al., 2006).

Sometimes textual information can indeed indicate similarities that are not visible to bibliometric techniques, and vice versa. As an example, we encountered two papers with bibliographic coupling similarity equal to 0, but with more than 95% textual cosine similarity. Both papers were of *Ding* and *Foo* and were published in *Journal of Information Science* (Appendix: Ding et al., 2002a; Ding, 2002b). The reason why both papers were not bibliographically coupled is that they mostly cited literature not published in periodicals or serials. However, both papers were correctly identified as being very similar by the textual cosine similarity, as they were follow-up papers, namely part I and II of “*Ontology research and development. A review of ontology generation*”. As an aside, there was

¹. This work was supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc & fellow grants. Flemish Government: Steunpunt O&O Indicatoren; FWO: PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame. Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011). EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.

actually one cited reference common to both papers, but the cited work was published more than 10 years before the papers under investigation, which is a common threshold for bibliographic coupling or co-citation analyses.

On the other hand, based on text alone, true document similarity might be obscured by differences in vocabulary use, or spurious similarities might be introduced as a result of textual pre-processing like stemming, or because of polysemous words or words with little semantic value. For instance, documents about music information retrieval might erroneously be linked to patent-related research based on common terms that are used in both contexts, such as *title*, *record*, *creative*, *business*, etc.

In an earlier study, the concept structure of (library and) information science (IS) was obtained by full-text mining of almost 1000 articles and notes published in the period 2002–2004 in 5 representative journals with strong focus on information science (Janssens et al., 2006a). The optimum solution for this text-based clustering of IS was found for six clusters. Besides two clusters in bibliometrics, one cluster was found in information retrieval, another one containing general and miscellaneous issues, and webometrics and patent studies were identified as small but emerging clusters within IS. In that study only the 'pure' text corpus was analysed, excluding any bibliographic or bibliometric components. However, the authors have assessed the performance of clustering and classification algorithms using various data integration schemes on a dataset with bioinformatics-related publications (Janssens et al., 2006b). The best outcome was obtained by hybrid methods that exploited both text and citations. An integration method based on Fisher's inverse chi-square and another one based on linear combination of distance matrices were among the best methods and significantly outperformed corresponding text-only and link-only methods, as well as a concatenation of vectors. In the present study we use these methods to map IS by using the full-text information as well as citations, and we compare the results with the text-only clustering.

Dataset

The document set used for our study consists of 914 full-text articles or notes, published between 2002 and 2004 in one of five journals. Table 1 shows the distribution of the 914 documents over the selected journals. This data set is the same as introduced by Janssens et al. (2006a), except for the exclusion of 24 publications. By matching of the articles with the Web of Science (WoS) database of *Thomson Scientific* (Philadelphia, PA, USA), we noticed that 22 were actually not listed as article or note, but that there was one letter among them, 6 were reviews, 11 editorial materials were included, as well as 4 biographical-items. Finally, two duplicate publications were detected in the original set. The exemption of 22 unique papers (or 2.3%) for this analysis, will not distort results much, particularly because the documents were removed from clusters in a reasonably stratified manner (12 from the largest Bibliometrics cluster, 5 from Information Retrieval (IR), 3 from the Social cluster, and 1 each from Webometrics and Bibliometrics2). Moreover, the optimal number of clusters for text-only clustering of the remaining 914 publications still amounts to 6. For comparing hybrid and text-based clustering, the latter one has first been redone on the smaller data set.

Table 1. The distribution of the 914 articles or notes over the 5 selected journals

Journal	Number of papers	%
<i>Information Processing & Management</i>	139	15.2
<i>Journal of the American Society for Information Science and Technology (JASIST)</i>	306	33.5
<i>Journal of Documentation</i>	82	9
<i>Journal of Information Science</i>	131	14.3
<i>Scientometrics</i>	256	28
Total	914	100

Methodology

Text representation

All textual content was indexed with the Jakarta Lucene² platform (Hatcher & Gospodnetić, 2004) and encoded in the vector space model using the TF-IDF weighting scheme (Baeza-Yates & Ribeiro-Neto, 1999). Text pre-processing comprised the following steps:

- Text extraction from full-text papers, preceded by Optical Character Recognition if necessary.
- Automatically removing acknowledgements and cited references from article content.
- Neglecting of author names, stop- and template words, URLs and e-mail addresses and stemming all remaining terms with the Porter (1980) stemmer.
- Phrase and synonym detection and reintroducing author names being part of a phrase (see Dunning, 1993; Manning & Schütze, 2000).

The dimensionality of the *term-by-document* matrix was reduced to 150 factors by latent semantic indexing (LSI) (Deerwester et al., 1990; Berry et al., 1995). Text-based similarities were calculated as the cosine of the angle between the vector representations of two papers (Salton & McGill, 1986). Figure 1 presents an overview of the textual analysis.

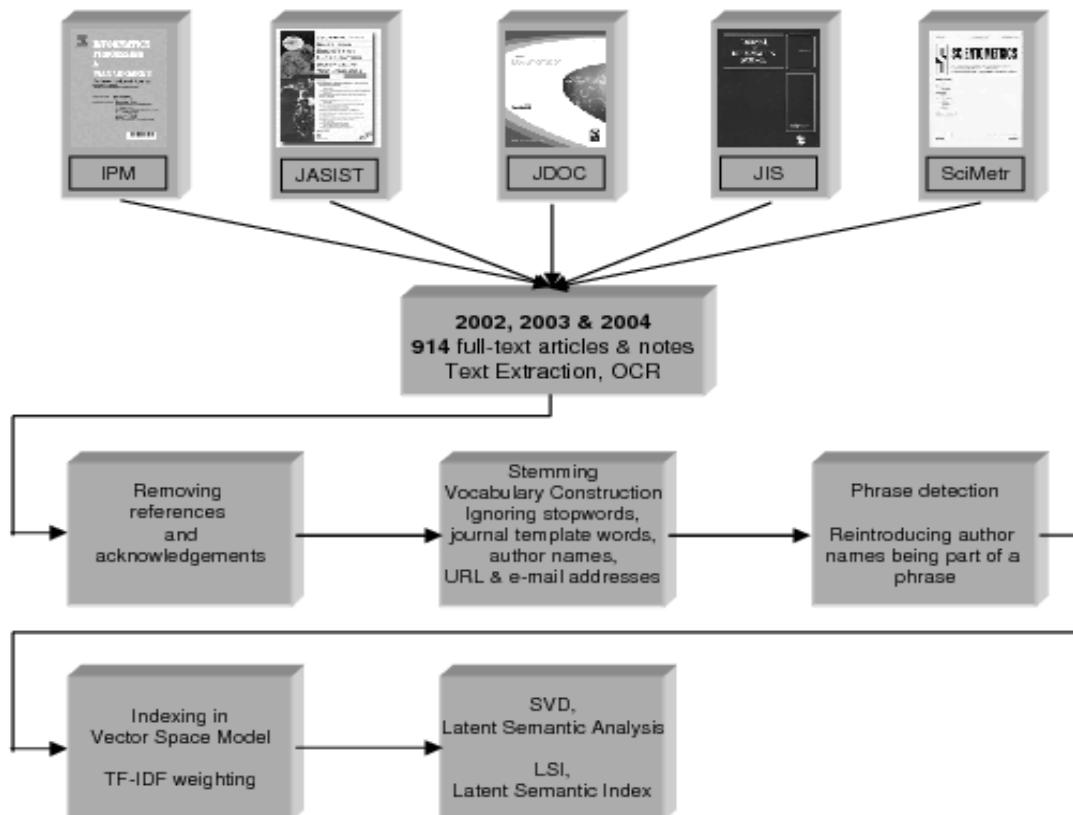


Figure 1. Overview of the textual analysis

Citation analysis

The cosine measure used to quantify the text-based similarity between any two documents can analogously be used with Boolean input vectors indicating the cited references in an article, or indicating all citing articles. If the vectors are normalised this corresponds to bibliographic coupling (Kessler, 1963) and co-citation, respectively, which are two citation-based measures of similarity. In the present study we use bibliographic coupling and combine it with the text-based similarities in order to obtain a hybrid data source that can be used by our clustering algorithm.

² <http://lucene.apache.org/>, visited in November 2006.

Clustering

To subdivide the IS papers into clusters we used the agglomerative hierarchical clustering algorithm using Ward's method (Jain & Dubes, 1988). It is a 'hard' clustering algorithm meaning that every publication is assigned to exactly 1 cluster. We determined the optimal number of clusters by observing the dendrogram, the curves with mean Silhouette coefficient for various numbers of clusters (Kaufman & Rousseeuw, 1990), and by using the stability-based method of Ben-Hur et al. (2002). In the latter method, the optimal number of clusters k is determined by inspection of a stability diagram (see Figure 4 for an example). The plot shows, for 2 up to 25 clusters, the cumulative distribution of the mutual similarities between 200 pairs of clustering solutions for random sub-samples of the data set, each comprising 777 documents (sampling ratio of 85%). For higher k the distances between the curves decrease and form a band. For the optimal number of clusters the largest k is chosen such that partitions into k clusters are still stable. This comes down to looking for a transition curve to the band of wide distributions.

The Silhouette value for a document ranges from -1 to +1 and measures how similar it is to documents in its own cluster vs. documents in other clusters (Rousseeuw, 1987). The mean Silhouette value for all documents is a measurement of the overall quality of a clustering solution.

The requisite input for many clustering algorithms includes mutual distances between all objects (documents). These distances can be based on the text, on the citations, or on a combination of both information sources. In the next subsections we describe linear combinations of distance matrices as well as Fisher's inverse chi-square method.

Weighted linear combination of distance matrices

For both data sources, i.e. the normalised *term-by-document* matrix A and the normalised *cited_references-by-document* matrix B , square distance matrices D_T and D_{BC} can be constructed as follows:

$$D_T = O_N - A^T \cdot A$$

$$D_{BC} = O_N - B^T \cdot B,$$

with N the number of documents and O_N a square matrix of dimensionality N with all ones. 'BC' refers to bibliographic coupling. These distance matrices D_T and D_{BC} can be combined into an integrated distance matrix D_i by a weighted linear combination (linco) as follows:

$$D_i = \alpha \cdot D_T + (1 - \alpha) \cdot D_{BC}$$

The resulting D_i can then be used in clustering or classification algorithms. A comparable methodology was described as the toric k-means algorithm by Modha & Spangler (2000). Although a very attractive, easy and scalable integration method, caution should be taken as a linear combination might neglect differences in distributional characteristics of various data sources (Janssens et al., 2005).

Fisher's inverse chi-square method

As a plain linear combination might not be optimal for integrating both textual and bibliographic coupling (BC) information, we developed a methodology based on Fisher's inverse chi-square method. Fisher's inverse chi-square is an omnibus statistic from statistical meta-analysis to combine p -values from multiple sources (Hedges & Olkin, 1985). In contrast to the weighted linear combination procedure, this method can handle distances stemming from different metrics with different distributional characteristics and avoids domination of any information source.

All text-based and link-based document distances in D_T and D_{BC} , as described above, are transformed to p -values with respect to the cumulative distribution function of distances for randomised data. This randomisation is a necessary condition for having valid p -values. In our setting, a p -value means the probability that the similarity between two documents could be at least as high just by chance.

For a correct application of Fisher's inverse chi-square method the input test statistics should be continuous. However, this is not the case for the sparse BC since most pairs of scientific articles do not have any reference in common. For this problem, we defined a slight, rank-preserving modification to

the original formula for BC by adding a constant 0.01 to the numerator. The new 'dense BC' between two papers x and y is then $(N_{xy} + 0.01) / \sqrt{N_x \cdot N_y}$, with N_x and N_y the number of references in paper x and paper y , respectively, and N_{xy} the number of references in common. The advantage of this formula is that it leads to a much larger set of possible values. In practice, however, it is still a finite set of discrete values, so we also superimposed Gaussian noise (standard deviation of 0.0025). The random noise will not deteriorate results since the error to be expected from for instance missing references in the WoS database is much higher.

If the p -values for the textual data (p_1) and for link data (p_2) are calculated, an integrated statistic p_i can be computed as $p_i = -2 \cdot \log(p_1^\lambda \cdot p_2^{(1-\lambda)})$. If the null hypothesis is true (i.e., in the case of randomised data), the distribution of $(p_1^\lambda \cdot p_2^{(1-\lambda)})$ is uniform and the integrated statistic has a chi-square distribution with 4 degrees of freedom (Hedges & Olkin, 1985). The complement of the integrated p -value, $(1 - p_i)$, is the new integrated document similarity measure that can be used in clustering or classification algorithms.

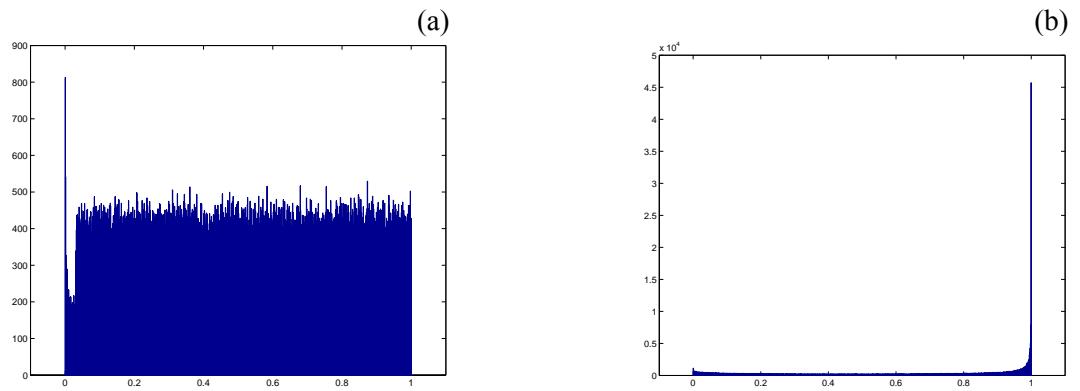


Figure 2. Histogram of p -values corresponding to (a) bibliographic coupling and (b) textual pairwise document distances for the real data w.r.t. randomised data.

In Figure, the distribution of p -values for the 'real data' is not uniform because otherwise there would be no structure in the data as the distribution of distances would be the same as for randomised data. The peaks at 0 and 1 indicate that for the 'real data', with respect to the random data, more document pairs have a very small or a very large distance.

The weight λ can be used to tune the relative importance or 'quality' of both information sources; however, choosing a good value for λ is not straightforward. We propose to define λ by choosing a value x for the *smallest but still significant* BC link (e.g., $x = 0.03$) and a value y for the smallest text-based similarity that is also still significant (e.g., $y = 0.1$). x and y can be based on visual inspection of the histogram of similarities, in combination with some experience. Next, convert the distances $(1-x)$ and $(1-y)$ to p -values p_x and p_y , respectively, and choose λ such that both *weakest still significant links* have the same contribution in p_i , by asking that $p_x^\lambda = p_y^{1-\lambda}$. This way we compensate for the fact that significant similarities are not as numerous in both datasets.

Fisher's inverse chi-square method can also be applied if SVD is used as a pre-processing step for either the textual data (LSI), either for the citation-based component, or for both. The random document vectors should then first be projected in the same space of reduced dimensionality, before calculating the distribution of document similarities. After application of SVD, intuitively defining the *smallest but still significant* distance by an expert becomes more difficult. However, a parameter sweep can still be performed and the difficulty of defining λ is compensated by the augmented performance after applying SVD, especially on the textual data.

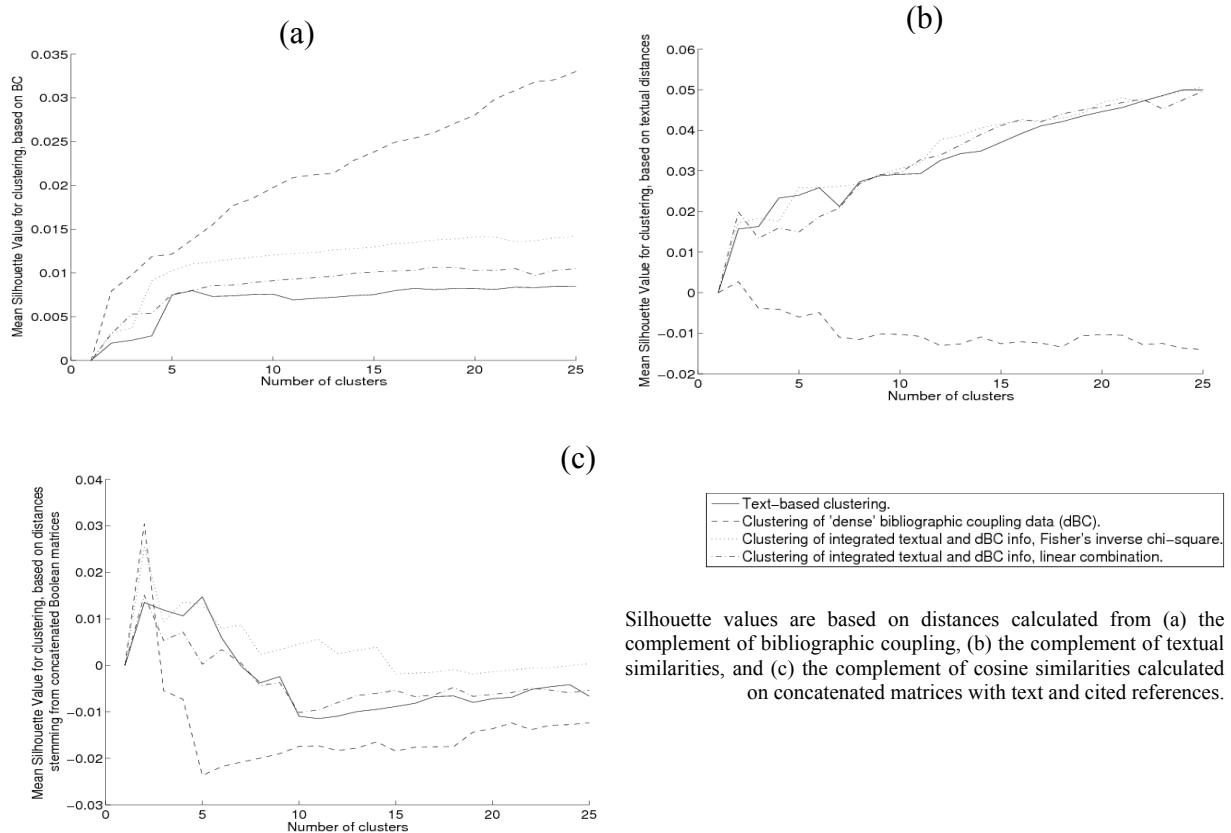
Term networks

For visualisation of clustering results we determined for each cluster the best words or phrases according to mean TF-IDF weights (see Figure 5 for an example). Each cluster has its own 'central node', represented by a diamond, which also indicates the number of members. Each central node points to the best 20 keywords for the cluster. When a keyword is among the best 20 for more than one cluster, it is only repeated once but connected to all corresponding cluster nodes. The grey level and thickness of an arc reflect the importance of a word for a cluster. Two terms are connected if both occur next to each other in one or more papers of the same cluster (only considering important words); the more co-occurrences, the closer the terms. Pajek was used for visualisation (Batagelj & Mrvar, 2002).

Results

Optimal number of clusters

As Silhouette values are intrinsically based on distances (Rousseeuw, 1987), depending on the chosen source of distances different Silhouettes can be calculated. In each case (a), (b) and (c) from Figure 3, we used the complement of cosine similarity as distance measure, but each time with a different input matrix. In (a), the *cited_references-by-document* matrix was used, whereas the *term-by-document* matrix was the input for (b). Finally, for (c), integrated distances were calculated from both matrices concatenated.



Silhouette values are based on distances calculated from (a) the complement of bibliographic coupling, (b) the complement of textual similarities, and (c) the complement of cosine similarities calculated on concatenated matrices with text and cited references.

Figure 3. Silhouette curves with mean Silhouette coefficient for clustering solutions of 2 up to 25 clusters for text-only clustering, link-only clustering, integrated clustering with Fisher's inverse chi-square method, and integrated clustering by linear combination of document similarities.

Table 2. Optimal number of clusters for Fisher's inverse chi-square method as perceived by the stability-based method (Figure 4) and by different mean Silhouette curves in Figure 3 using link-based (a), text-based (b) and integrated distances (c).

Evaluation method	Number of clusters
Mean Silhouette value based on BC (Figure 3(a))	≥ 4
Mean text-based Silhouette value (Figure 3(b))	≥ 5
Mean 'hybrid' Silhouette value (Figure 3(c))	4 or 5
Stability diagram (Figure 4)	3, 4 or 5

In the experiments of Figure 3, the integration weight was set to 0.5 for both linco and Fisher's inverse chi-square method for simplicity of comparing, but conclusions with regard to number of clusters remain the same (see also Table 2). For the optimal number of clusters for hybrid clustering by Fisher's inverse chi-square method, the curve with citation-based Silhouettes (Figure 3(a), curve for 'Fisher's inverse chi-square') hints towards 4,5,6 or more clusters, whereas the text-based Silhouettes show a local maximum for 5 clusters (b). Figure 3(c) suggests 4, or maybe 5 clusters, but not more. By observing the stability diagram in Figure 4, a solution with 5 clusters seems clearly more stable than 6 clusters, while not differing that much in stability from 3 or 4 clusters. Based on these findings, we chose 5 as the optimal number of clusters for the inverse chi-square integrated clustering. On the dendrogram (not shown), five clusters could also be considered as a nice cut-off point.

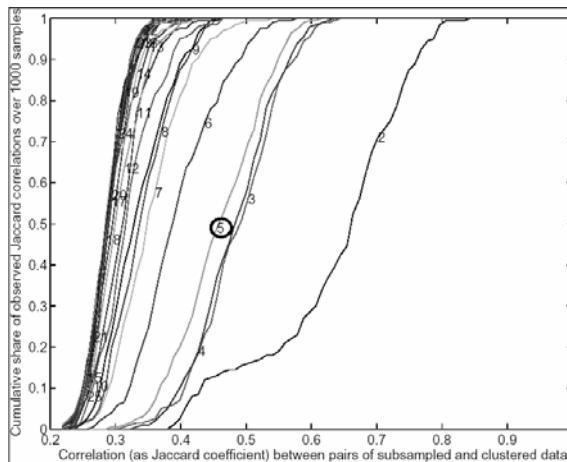


Figure 4. Stability diagram for determining the number of clusters for hybrid clustering using Fisher's inverse chi-square method (Ben-Hur et al., 2002).

Comparing Fisher's inverse chi-square method with linco, text-only & link-only clustering

When using the same composite procedure for determining the optimal number of clusters in the case of the linco method, two observations could be made. First, when clustering the linearly combined distance matrices, the optimal number of clusters was 8, compared to 5 for Fisher's inverse chi-square method. Secondly, linco came up with very large, noisy clusters for any solution with less than 8 clusters. E.g., when asking for 5 clusters, the largest cluster even contained 722 out of 914 documents. Figure 3 also presents a more detailed comparison of the performances of linco, Fisher's inverse chi-square method, text-only and link-only clusterings. Main observations are summarised in Table 3. In (a), not surprisingly, the link-based clustering of dBC values performs best. However, when validating with textual or integrated distances (b&c), this link-only clustering performs very poorly. Integration of text and cited references leads to better Silhouettes than pure text-based methods in (a). From the same figure it is also clear that Fisher's inverse chi-square method does a better job than linco, perhaps an illustration of textual information dominating citations in case of plain linear combinations. Furthermore, linco provided somewhat less stable clusterings than Fisher's inverse chi-square method.

Table 3. General appreciation of clustering different data types by observing Silhouette curves in Figure 3. A lower value indicates a better appreciation, 1 is best and 4 is worst. Different values are possible for different ranges of cluster numbers, indicated between brackets.

	Text-based clustering	Clustering of dBC	Fisher's inverse chi-square	Linear combination
Mean Silhouette value based on BC (see Figure 3(a))	4	1	2	3
Mean text-based Silhouette value (see Figure 3(b))	3 ($c=2, c>10$) 1 or 2 otherwise	4	1 or 2	3 ($c=3..6$) 1 or 2 otherwise
Mean 'hybrid' Silhouette value (see Figure 3(c))	1 ($c=3$ or 5) 4 ($c=2$) 2 or 3 otherwise	1 ($c=2$) 4 otherwise	2 ($c=2, 3$ or 5) 1 otherwise	2 or 3

A little counterintuitive but very beneficial is that, when the validation relies on pure text-based Silhouettes (b), Fisher's inverse chi-square does at least equally well as the pure text-based clustering (which actually *plays a home game* here), except for a four clusters solution. The linco method is the best one on a very coarse level of aggregation with only two clusters, but then goes down. From 7 clusters onwards linco again does as good as or even better than text-based clustering and for more than 10 clusters it competes with the Fisher curve. Thus, based on evaluation with textual Silhouettes, Fisher's inverse chi-square method in general again outperforms linear combination.

In (c), which represents the most natural way of evaluating integrated clusterings, namely by basing the Silhouettes on integrated data, Fisher's inverse chi-square method is again the method of choice. Surprisingly, the clustering of linearly combined data is not better here than the text-only clustering, maybe another illustration of textual data dominating citations. Interestingly, the local maximum of the text-based solution at 6 six clusters in figure (b), as also described by Janssens et al. (2006a), also decreases to 5 clusters in (c), being evaluated with integrated Silhouette values.

Hybrid mapping of the field by using Fisher's inverse chi-square method

Figure 5 presents the cognitive structure of IS as a term network consisting of, for each of 5 clusters, the best 20 stemmed terms or phrases from titles or abstracts according to mean TF-IDF scores. We have labelled the clusters based on their most significant terms and most 'representative' publications (see Table 4). In order to determine these 'medoid' papers closest to the corresponding centroids, we looked at the largest similarities to the mean cluster profile (centroid). Three large and two smaller clusters can be distinguished. Publications in the three larger classes are concerned with rather traditional sub-disciplines of the IS field, particularly, with Information Retrieval (IR), Bibliometrics/scientometrics and with what we called "Social" aspects. The latter term is probably not the best solution but it clearly refers to the fact that many of the papers in this cluster deal with user and community relevant questions, their composition or special demands, etc. The two smaller classes represent relatively new and emerging topics in IS, namely, Patent analysis and Webometrics. Hence, the hybrid clustering result contains the same topics as found by the text-based clustering (Janssens et al., 2006a), except for the merger of the two Bibliometrics clusters. Figure 5 also visualises the interconnections between clusters. Clusters 3 and 5 are connected through the science-technology interface as represented among others by national science and technology indicators, patent citations, industry research, patenting universities and inventor-author coactivity. Interdisciplinary research in the intersection of science and technology – here represented by the stem *nanotechnolog* – is also one of the bridges between the two paper sets. Citations and their equivalents on the Web (in-links/out-links) form the important connection between the Bibliometrics cluster and the Webometrics cluster, which, in turn, is strongly liked to the general/Social cluster through the Web use. The stem *queri* finally connects Webometrics with IR. Here, *search engin*, *web crawler* and *algorithm* form a strong interface.

Table 4. For each of 5 clusters the two medoid papers, the publications with largest cosine similarity to the mean cluster profile.

Cluster 1. Information Retrieval (312 documents)
Schlieder, T. & Meuss, H. (2002). Querying and ranking XML documents. <i>JASIST</i> , 53, 489-503.
Huang, C. K., Chien, L. F., & Oyang, Y. J. (2003). Relevant term suggestion in interactive Web search based on contextual information in query session logs. <i>JASIST</i> , 54, 638-649.
Cluster 2. Webometrics (63 documents)
Thelwall, M. & Harries, G. (2004). Do the Web sites of higher rated scholars have significantly more online impact? <i>JASIST</i> , 55, 149-159.
Thelwall, M. & Harries, G. (2003). The connection between the research of a university and counts of links to its web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. <i>JASIST</i> , 54, 594-602.
Cluster 3. Patent (31 documents)
Bhattacharya, S. (2004). Mapping inventive activity and technological change through patent analysis: A case study of India and China. <i>Scientometrics</i> , 61, 361-381.
Meyer, M., Sinilainen, T., & Utecht, J. T. (2003). Towards hybrid Triple Helix indicators: A study of university-related patents and a survey of academic inventors. <i>Scientometrics</i> , 58, 321-350.
Cluster 4. Social (272 documents)
Hargittai, E. (2002). Beyond logs and surveys: In-depth measures of people's Web use skills. <i>JASIST</i> , 53, 1239-1244.
Marchionini, G. (2002). Co-evolution of user and organizational interfaces: A longitudinal case study of WWW dissemination of national statistics. <i>JASIST</i> , 53, 1192-1209.
Cluster 5. Bibliometrics (236 documents)
Al Qallaf, C. L. (2003). Citation patterns in the Kuwaiti journal Medical Principles and Practice: The first 12 years, 1989-2000. <i>Scientometrics</i> , 56, 369-382.
Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. <i>Scientometrics</i> , 60, 421-432.

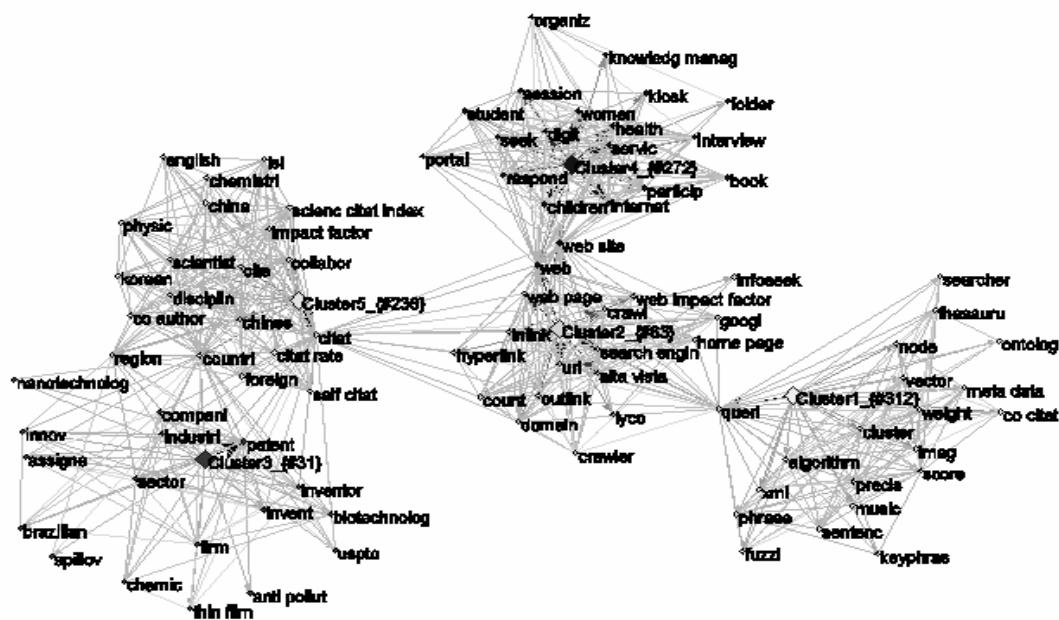


Figure 5. Term networks with for each of five clusters the best 20 stemmed terms or phrases from titles or abstracts according to mean TF-IDF scores.

The question arises of what the added value is of the combination of the two methods, the text-based and the bibliometrics aided approach. From the technical viewpoint, the appropriate choice of weight, automatically determined to be 0.43 by the method described above, results in a somewhat better

evaluation of clustering. If we compare the text-only approach and the hybrid solution, we clearly see a measurable improvement by the combination.

In Figure 6(a), the centroids of the six clusters of the text-only approach are compared with those of the five clusters of the hybrid method. Certain shifts around the merged Bibliometrics cluster can be observed. This change, however, also concerns other clusters. The centroid of the new merged Bibliometrics cluster is located nicely between the former two centroids. The Patent cluster is still the most distant one and has grown from 19 to 31 papers. Thus, some patent-related publications had been put in one of the bibliometrics clusters by the text-based algorithm, whereas the incorporation of citations has led to a more clear demarcation between patent and bibliometric studies. In the text-only setting, the Patent cluster was closer to Bibliometrics1 than to Bibliometrics2 and Bibliometrics1 was even combined with Patent before being combined with Bibliometrics2 (Janssens et al., 2006a). The present hybrid results correspond more to our intuition: there is only one Bibliometrics cluster and the Patent cluster is only merged with bibliometrics in a later stage. This leads immediately to the question of 'migrated' papers. More than a quarter of the papers was assigned to a different cluster according to the hybrid scheme.

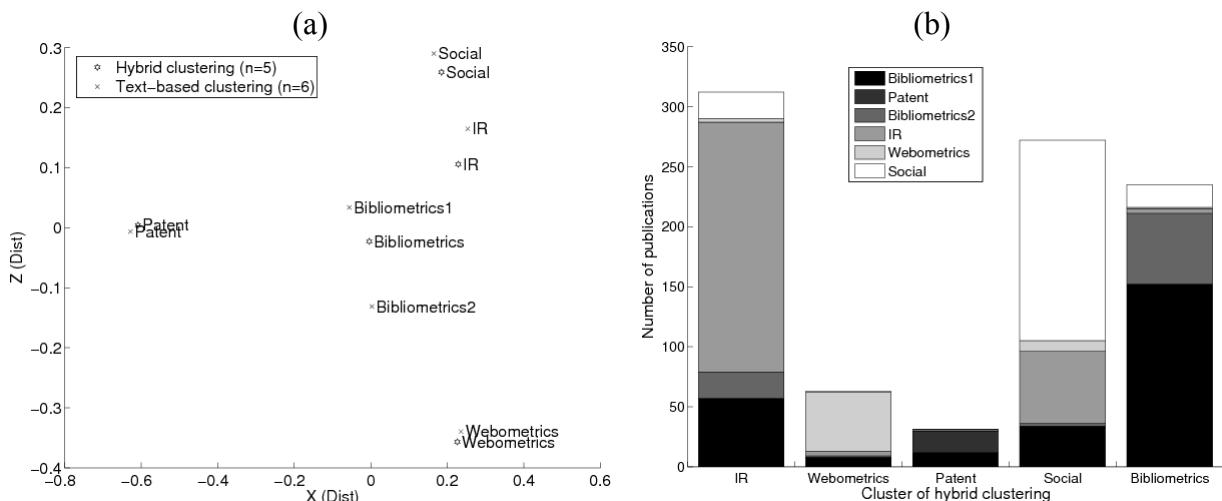


Figure 6. (a). Multidimensional scaling (MDS) plot comparing the cluster centers (centroids) of the six clusters found by the text-based clustering, with the five cluster centers of the hybrid clustering.

(b). The overlap of each of 5 clusters determined by hybrid clustering with Fisher's inverse chi-square method, with the text-based clusters.

Figure 6(b) visualises the overlap between hybrid and text-based clusters. By checking paper assignment to clusters according to the two methods manually, we found that many of these 'migrated' papers were originally misplaced in the text-based approach, like the 'new' patent papers discussed above. Nonetheless, incorrectly assigned papers still occur in the combined classification, too, but this is probably unavoidable as the adopted hard agglomerative hierarchical clustering algorithm has intrinsic weaknesses. One of the disadvantages is that wrong choices (merges) that are made by the algorithm in an early stage can never be repaired (Kaufman & Rousseeuw, 1990). To distinguish the good from the bad migrations, we sorted all migrated documents according to descending difference in text-based silhouette values for hybrid minus text-based clustering. The prevalence of positive values indicated that there are more correct than spurious migrations. A few of many examples of good migrations are the following. A paper of *Nie* about "Query expansion and query translation as logical inference" migrated from the text-based Bibliometrics1 cluster to the hybrid IR cluster (Appendix: Nie, 2003). Next, the paper of *Faba-Perez* et al. about "'Sitation' distributions and Bradford's law in a closed Web space" was put in the Webometrics cluster instead of Bibliometrics1 (Appendix: Faba-Perez et al., 2003). Finally, "Knowledge integration in virtual teams: The potential role of KMS" by *Alavi & Tiwana* changed from Bibliometrics1 to the more Social cluster (Appendix: Alavi et al., 2002).

On the other hand, less evident migrations could also be observed. For example, "Empirical evidence of self-organization?" (Appendix: van den Besselaar, 2003) moved from Bibliometrics1 to IR, and the same goes for a paper by Leydesdorff, "Indicators of structural change in the dynamics of science: Entropy statistics of the SCI Journal Citation Reports" (Appendix: Leydesdorff, 2002).

Figure 7 visualises the effect of migration after merging the two bibliometrics clusters through incorporating text and citations. 24 new papers appear in the very centre of the new Bibliometrics cluster as consequence of migration. Other documents of the former Bibliometrics1 and Bibliometrics2 clusters are not included in the new one, among which the patent related publications. There is still another reason for the 'success' of the hybrid or bibliometrics aided classification beyond any technical considerations. Any lexical (text-based) approach is usually based on rather rich vocabularies and peculiarities of natural language. The result is, according to our observations, a rather 'smooth' or gradual transition between what is related and what is not. The relationship between documents is, therefore, somewhat fuzzy and not always reliable. On the other hand, if strict citation-based criteria are applied, that is, if non-periodical references and occasional coupling links are removed, the resulting *citations-by-document* matrix becomes extremely sparse. In this case, rejection of relationship tends to be unreliable. The above-mentioned modification to the original formula for bibliographic coupling by adding a constant 0.01 to the numerator helps smoothing the 'singularity', but is not able to overcome it. This might explain the low efficiency of coupling- (and co-citation)-based clustering techniques. The combination of the two techniques helps to improve the reliability of relationship and therefore of the clustering algorithm as well.

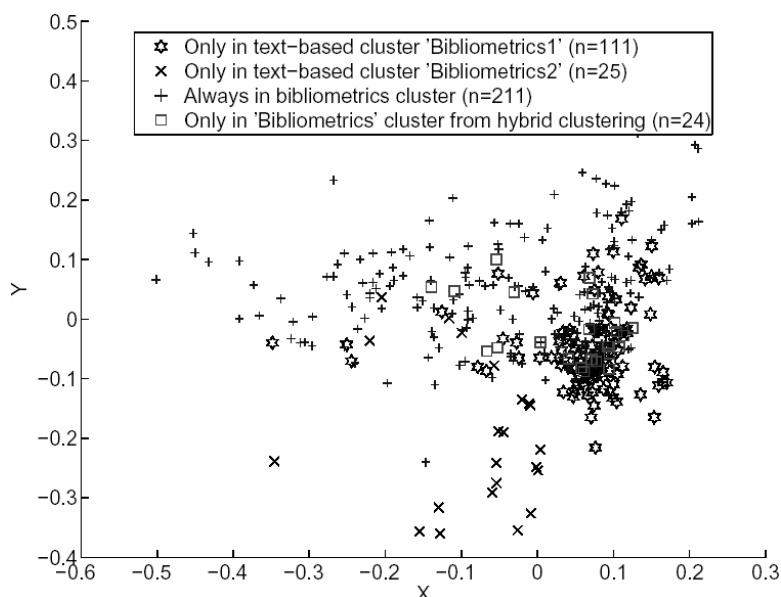


Figure 7. Multidimensional scaling (MDS) plot only considering documents in the two bibliometrics clusters of the text-based solution and documents in the bibliometrics cluster of the hybrid clustering. A distinction is made between documents that were only once assigned to a bibliometrics related cluster, and documents that were consistently assigned to bibliometrics.

Conclusion

The field of information science was subdivided into 5 classes by using a hybrid clustering method based on Fisher's inverse chi-square, incorporating the full text of scientific publications and their cited references. Three large clusters could be distinguished: one containing research in information retrieval, another one about bibliometrics/scientometrics, and, finally, a collection of more 'socially' directed topics. The two smaller classes, patent analysis and webometrics, represented relatively new and emerging topics in IS.

In order for the inverse chi-square method to be applicable, we used a slightly modified formula for bibliographic coupling and also superimposed a noise component. A method was proposed to determine the integration weight λ for tuning the relative importance of two information sources.

The number of clusters was semi-automatically determined by applying a combination of distance-based and stability-based methods. However, the optimal number is still a difficult issue and depends on the adopted validation and on the chosen similarity measures, as well as on the input data, be it mere text, just citations or a combination. For a text-only clustering, 6 clusters seemed to be optimal, whereas for an integrated clustering by linear combination even 8 could be perceived as the best choice.

This latter algorithm, although being a very attractive, easy and scalable integration method that had previously not been shown to be inferior to Fisher's inverse chi-square method, was in the present setting outperformed with regard to the Silhouette coefficient and stability.

By integrating text and citations, Fisher's inverse chi-square method also did quantitatively and qualitatively better than the pure text-based method. We compared the six clusters of the text-only approach with the five clusters of the hybrid method. Quite some papers had 'migrated' to another cluster. Many of these were originally misplaced in the text-based approach, so we clearly observed an improvement by the combination. On the other hand, incorrectly assigned papers still occurred in the combined classification as well. However, this is unavoidable as the adopted hard agglomerative hierarchical clustering algorithm has intrinsic weaknesses. We think that, in order to gain even better performance, a transition should be made towards fuzzy clustering algorithms.

References

- Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Cambridge: Addison-Wesley.
- Braam, R. R., Moed, H. F. & van Raan, A. F. J. (1991). Mapping of Science by Combined Cocitation and Word Analysis. 2. Dynamic Aspects. *JASIS*, 42, 252-266.
- Batagelj, V. & Mrvar, A. (2002). Pajek - Analysis and visualization of large networks. *Graph Drawing*, 2265, 477-478.
- Ben-Hur, A., Elisseeff, A. & Guyon, I. (2002). A stability based method for discovering structure in clustered data. In *Pacific symposium on biocomputing* (vol. 7, pp. 6-17). Retrieved November 15, 2006 from: <http://helix-web.stanford.edu/psb02/benHur.pdf>.
- Berry, M., Dumais, S. T. & O'Brien, G. W. (1995). *Using linear algebra for intelligent information retrieval*. SIAM Review, 37(4), 573-595.
- Calado, P., Ribeiro-Neto, B., Ziviani, N., Moura, E. & Silva, I. (2003). Local versus global link information in the Web. *ACM Transactions on Information Systems*, 21, 42-63.
- Calado, P., Cristo, M., Goncalves, M. A., de Moura, E. S., Ribeiro-Neto, B. & Ziviani, N. (2006). Link-based similarity measures for the classification of Web documents. *JASIST*, 57, 208-221.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6), 391-407.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61-74.
- Glenisson, P., Glänsel, W., Janssens, F. & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, 41, 1548-1572.
- Hatcher, E. & Gospodnetić, O. (2004). *Lucene in Action*. New York: Manning Publications Co.
- Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego: Academic Press.
- Jain, A. & Dubes, R. (1988). *Algorithms for clustering data*. New Jersey: Prentice Hall.
- Janssens, F., Leta, J., Glänsel, W. & De Moor, B. (2006a). Towards mapping library and information science. *Information Processing & Management*, 42(6), 1614-1642.
- Janssens, F., Tran Quoc, V., Glänsel, W. & De Moor, B. (2006b). Integration of textual content and link information for accurate clustering of science fields. In V. P. Guerrero-Bote (Ed.), *Proc. of the I Intl. Conf. on Multidisciplinary Information Sciences and Technologies* (InSciT2006) (pp. 615-619), Mérida, Spain.
- Joachims, T., Cristianini, N. & Shawe-Taylor, J. (2001). Composite kernels for hypertext categorisation. In *Proceedings of the 18th International Conference on Machine Learning* (ICML) (pp. 250-257).
- Kaufman, L. & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley and Sons Inc.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14, 10-25.
- Manning, C. D. & Schütze, H. (2000). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

- Modha, D. S. & Spangler, W. S. (2000). Clustering hypertext with applications to web searching. *ACM Conference on Hypertext* (pp. 143-152).
- Mullins, N., Snizek, W. & Oehler, K. (1988). *The structural analysis of a scientific paper*. In A. F. J. vanRaan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 81–105). New York: Elsevier Science.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- Salton, G. & McGill, M. J. (1986). *Introduction to modern information retrieval*. New York: McGraw-Hill, Inc.
- Snizek, W., Oehler, K. & Mullins, N. (1991). Textual and nontextual characteristics of scientific papers: Neglected science indicators. *Scientometrics*, 20(1), 25–35.
- Wang, Y. & Kitsuregawa, M. (2002). Evaluating contents-link coupled web page clustering for web search results. In *Proc. of the 11th intl. conf. on Information and knowledge management* (CIKM) (pp. 499-506).

Appendix: Bibliographic sources of papers referred to in the text as subject of analysis:

- Alavi et al. (2002). *JASIST*, 53(12):1029-1037.
- Ding et al. (2002a). *Journal of Information Science*, 28 (2):123-136.
- Ding (2002b). *Journal of Information Science*, 28(5):375-388.
- Faba-Perez et al. (2003). *Journal of Documentation*, 59(5):558-580.
- Leydesdorff (2002). *Scientometrics*, 53(1):131-159.
- Nie (2003). *JASIST*, 54(4):335-346.
- van den Besselaar (2003). *JASIST*, 54(1):87-90.

Response Surface Methodology and its Application in Evaluating Scientific Activity¹

Evaristo Jiménez-Contreras*, Rafael Bailón Moreno*, Daniel Torres-Salinas**, Rosario Ruiz Baños*, Rafael Ruiz-Pérez*, Mercedes Moneda Corrochano* and Emilio Delgado López-Cózar*

*evaristo@ugr.es

Departamento de Biblioteconomía y Documentación, Universidad de Granada, Granada-18071 (Spain)

**torressalinas@gmail.com

Centro de Investigación Médica Aplicada (CIMA), Universidad de Navarra, Pamplona-31008 (Spain)

Abstract

The possibilities of the Response Surface Methodology (RSM) has been explored within the ambit of Scientific Activity Analysis. The case of the system “Departments of the Area of Health Sciences of the University of Navarre (Spain)” has been studied in relation to the system “Scientific Community in the Health Sciences”, from the perspective of input/output models (factors/response). It is concluded that the RSM reveals the causal relationships between factors and responses through the construction of polynomial mathematical models. Similarly, quasi-experimental designs are proposed, these permitting scientific activity to be analysed with minimum effort and cost and high accuracy.

Keywords

evaluation of scientific activity; response surface methodology; experimental design

Introduction

The analysis of scientific activity focusing on the economic input/output model—especially when dealing with institutions—is classical and almost the foundation of scientific evaluation (Martin & Irvine, 1983). This model implies that the system under study has easily defined borders affected by a set of factors or variables called *inputs* and which represent the resources of the system (funding, researchers, equipment, etc.). This system in turn generates or responds to products resulting from their scientific activity, called *outputs*, such as publications or patents.

The relationships which link inputs with outputs are complex and difficult to describe with elemental mathematical models. Therefore, the need arises for tools that are capable of more complex modelling and that achieve maximum refinement of the role of each variable in the system as well as the of synergistic and/or antagonistic interrelationships between the same variables.

The Response Surface Methodology (RSM) emerged in the 1950s (Box & Wilson 1951; Box & Hunter 1951) within the context of Chemical Engineering in an attempt to construct empirical models able to find useful statistical relationships between all the variables making up an industrial system. This methodology is based on experimental design with the final goal of evaluating optimal functioning of industrial facilities, using minimum experimental effort. Here, the inputs are called factors or variables and the outputs represent the response that generates the system under the causal action of the factors or variables. In recent years it is being applied successfully in other scientific fields such as biology, medicine, and economy. Myers et al. (2004) has exhaustively reviewed the literature in the sense, describing the developments and applications of this methodology. Very recently, RSM has been used even to validate new experimental methods (Jurado et al. 2003).

Objectives

In this presentation, we seek to explore the possibilities of the Response Surface Methodology within the scope of the analysis of scientific activity.

¹ This study is part of a project funded by the initiative of the *Centro de Investigación Médica Aplicada de la Universidad de Navarra*. Color Version of all figures are available from the authors.



Figure 2. System Departments of Health Sciences of the University of Navarre

For this, we shall consider the case of the system “Departments of Health Sciences of the University of Navarre”. The University of Navarre will be represented by a system in which the factors (inputs) are the human resources as well as the economic resources while the response (outputs) are the scientific production (*Figure 2*)

Materials and methods

Description of the method of the response surfaces

The designs of the response surface methodology (RSM) are those in which problems are modelled and analysed; in these problems the response of interest is influenced by different variables. The RSM is widely used as an optimisation, development, and improvement technique for processes based on the use of factorial designs—that is, those in which the response variable is measured for all the possible combinations of the levels chosen of the factors or variables. The main effect of a factor is defined as the variation in response caused by a change in the level of the factor considered, when the other ones are kept constant. There is an interaction (dependence) between the variables when the effect of one factor depends on the behaviour of another. The application of the RSM becomes indispensable when, after the significant factors affecting the response have been identified, it is considered necessary to explore the relationship between the factor and dependent variable within the experimental region and not only at the borders. Response surfaces are recommended for these types of factorial designs for their effectiveness and quick execution. This consists of correlating the k variables put into action through a second-degree polynomial expression of the following form:

$$y_{obs} = b_0 + \sum_{i=1}^k b_i x_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^k b_{i,j} x_i x_j + \sum_{i=1}^k b_{i,i} x_i^2 + e$$

where y_{obs} is the dependent variable, and x_i the factors or variables with which we wish to correlate it. The expression contains a first-degree term that represents a linear relationship considered as the principal, another term in which the variables cross each other to represent the influence of some over others, and finally a second-degree term that refines the previous one and gives maximums and minimums—i.e. optimal values of the dependent variable. The symbols b_0 , b_i , $b_{i,j}$ are constants and e a term of error or residual between the observed and calculated value. The experimental values are adjusted to the above equation by a polynomial regression and the usual statistics can be used to determine the goodness of the fit.

The SRM implies, apart from the use of a second-degree polynomial model, a very reduced experimental design called Central Composite Design (CCD). The CCD is formulated on the basis of the factorial designs adding the star points and the central point, and three types of different structures can be used (*Figure 3*).

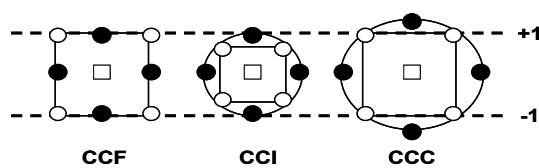


Figure 3. Different structures of CCD: central side (CCF), inscribed (CCI) and circumscribed(CCC)

Regardless of the structure of the composite central design that is used, for each factor or variable, experiments will be performed for 5 different values or levels: -2 , -1 , 0 , $+1$ y $+2$. Therefore, not all the combinations possible will be made, but rather only those that fulfil a geometric CCD design, i.e. only the points indicated.

In certain applications, the variables cannot take any combination of values, due to certain restrictions. Figure 4 is an example of an experimental window where only in the shaded area, limited by restricting lines, is the design feasible. To facilitate the setting up and fit of the model, a new group of components are defined, these being called pseudocomponents.

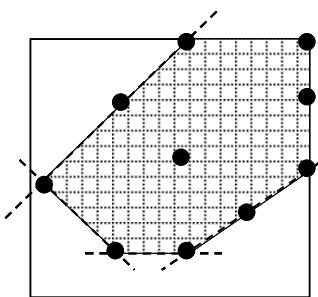


Figure 4. Example of design with restrictions

From the resulting values, for each of the variables, the coefficients of the polynomial equation are determined (b_0 , b_i , $b_{i,j}$) and the equation can be simplified according to the influence of the factors in the final response. The resulting equation is used as a model of a given system to determine the response of y as a function of the different values of x_1 and x_2 within the defined area in the CCD.

Material: Area of Health Sciences of the University of Navarra

An evaluation was made of the international scientific production of 50 departments of the University of Navarra (UNAV) related to Biomedicine and Health Sciences in the period 1999-2005. The production data and citations were taken from the Web of Science and those of impact from the Journal Citation Reports corresponding to this period. The information on economic and human resources was provided by the this university. Table 1 presents the variables that have been analysed in this work.

Table 6. Variables analysed in the evaluation of the UNAV

Indicators
<i>Nº of researchers</i>
<i>Funding through research projects</i>
<i>Nº of works in the databases of Web of Science</i>

Overall, the UNAV produced a total of 2,229 works that have received a total of 19,716 citations. Some 41% of their works have been published in journals in the first quartile. The economic resources identified come from the funding of 534 research projects classified into 5 typologies: Europeans (4%); International (1%); Internal (17%), and Regional (40%). It is assumed that the human resources for the period analysed had an annual mean of 764 full-time researchers, of which 485 were doctors and the rest pre-doctoral students.

Software

For the calculation of the response surfaces, a specific program was used, Modde v. 4, of the company Umetrics of Sweden (www.umetrics.com)

Results and interpretation of the results

For the calculation of the response surface of the system “University of Navarre” (UNAV), we used a CCF design with restriction, as shown in Figure 5. The cloud of points represents the group of

departments in the area of Health Sciences of UNAV. The points highlighted are those departments that have the characteristics closest to the CCF type of design with restrictions.

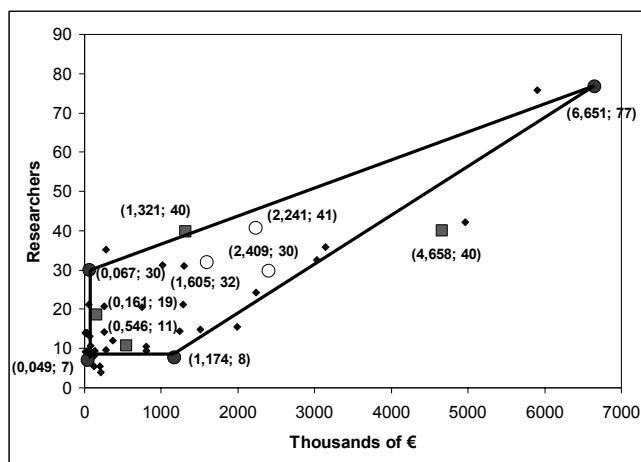


Figure 5. CCF design with restrictions for the UNAV system

The factors used are the number of researchers, S, and the funding, in the form of decimal logarithm, $\log F$ (F is expressed in thousands of €). The response is evaluated as production, P, of scientific articles listed in the Web of Science. The best fit corresponds to a linear response with respect to the number of researchers, while with respect to the logarithm of the funding the response is simultaneously linear and quadratic. There is also a response with respect to the interaction researchers-funding, which signifies that there is a synergetic effect between the two factors.

$$P = 233 - 2.6S - 191\log F + 44(\log F)^2 + 1.25S\log F \quad R^2 = 0.865 \quad Q^2 = 0.722$$

The goodness of the response surface represented by Eq. 1 is acceptable (R^2 and Q^2 are well) On the other hand, the F test of Snedecor confirms also that the fit is satisfactory at the significance level of 5%.

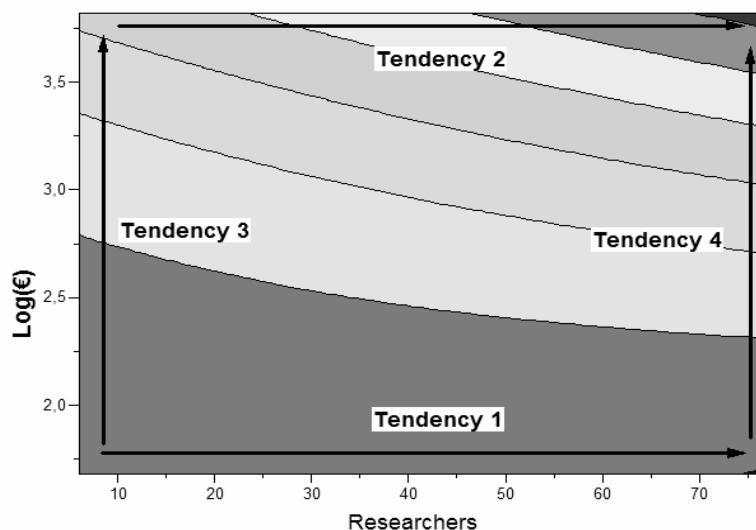


Figure 6. Flat or contour representation for the system UNAV

However, perhaps the most interesting aspect, from our viewpoint, is the generation of a graphic model that synthesises the weight of the variables chosen and their influence on the results as these variables are changed. There are two basic representations of the model: flat and contour (Figure 6), and three-dimensional or superficial (Figure 7).

From the flat and the three-dimensional representation, it is now very easy to explain the behaviour (response) of the scientific system of the UNAV according to whether the factors affecting the production of articles of the departments are affected or not.

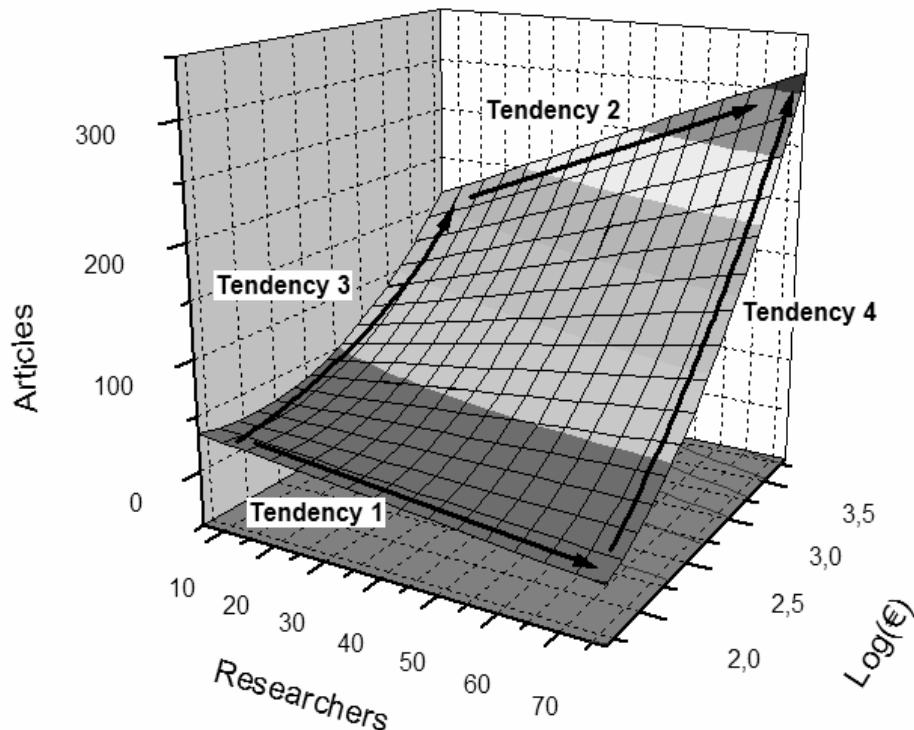


Figure 7. Three-dimensional representation of the surface of the system UNAV

Movement one (Tendency 1) shows what happens in the system when, under low funding, the number of researchers increases. Although it may prove unexpected, the result predicts a fall in scientific production. The explanation, however, proves attractive, when the resources are scarce, the increase in staff would prove counterproductive inasmuch as, with decreasing research funds as a consequence of the increase in researchers to attend, the capacity of producing new works tends to diminish (dark-blue fringe), as insinuated in the lower-right corner. Nevertheless, the capacity of the model in this sphere should not be exaggerated due to the scarcity of the data at this level, to their variability

Movement three (Tendency 3) shows the evolution in the situation of increasing funding with comparatively lower increases in staff. The possible situations covered by the blue and green segments show a progression in the results that even triple those obtained with low funding. Finally, great increases of investment are accompanied by a greater exponentially greater response, especially in the final part of the graph.

Finally, *movements two and four* (Tendency 2 and 4), which begin with few researchers having abundant financing and many researchers with little funding, the two groups converging in the form of many researchers with much funding. This inevitably marks a similar trajectory that culminates at the maximum limit of the results found in the case of the UNAV. However, the trajectories are not identical; in the first case the path is longer, given that it begins from a more deficient situation. In this sense, the general topography of the sample surface shows that it is far more effective to have fewer

human resources with better funding, than the contrary case. In other words, the economic variable is determinant in the human.

Conclusion

Although we know that the impact factor (IF) of a journal does not predict the IF of an author or a particular work, what seems evident is that the prestige itself of the journal attracts citations in that we group only a certain number of works. The authors that publish in high-impact journals, which have more capacity to select from among the many works sent to them, are more visible to the scientific community, this constitutes the other determining element and closes the *virtuous* circuit of research with impact.

References

- Martin, B.R. & Irvine, J. (1983): Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, 12, 61-92.
- Box, G.B.P. & Wilson, K.B. (1951). On experimental attainment of optimum conditions. *Journal of the Royal Statistical Society*, 13, 1-45.
- Box, G.E. P & Hunter, J.S. (1951): Multifactor experimental designs for exploring response surfaces. *Journal of the Royal Statistical Society*, 13(1), 195-240.
- Myers, R.H., Montgomery, D.C., Vining, G.G., Kowalski, S.M., and Borror, C.M. (2004), Response surface methodology: A retrospective and current literature review. *Journal of Quality Technology*, 36, 53-77.
- Jurado-Alameda, E., Bravo-Rodríguez, V., Bailón-Moreno, R., Nuñez-Olea, J. & Altmajer Vaz, D. (2003). Bath-Substrate-Flow Method for Evaluating the Detergent and Dispersant Performance of Hard-Surface Detergents. *Industrial & Engineering Chemistry Research*, 42, 4303-4310.

The Role of Ethnic Ties in International Collaboration: The Overseas Chinese Phenomenon¹

Bihui Jin*, Ronald Rousseau**, Richard P. Suttmeier*** and Cong Cao***

*jinbh@mail.las.ac.cn

National Science Library of the Chinese Academy of Sciences 33 Beisihuan Xilu, Zhongguancun, Beijing, 100080 (P.R. China)

**ronald.rousseau@khbo.be

KHBO, Industrial Sciences & Technology, Zeedijk 101, B-8400 Oostende (Belgium)
& University of Antwerp, IBW, Universiteitsplein 1, B-2610 Wilrijk (Belgium)

***petesutt@uoregon.edu, ccao@darkwing.uoregon.edu

University of Oregon, Eugene, OR 97403 (USA)

Abstract

The term ‘Overseas Chinese Phenomenon’ is used here to refer to the fact that scientists of Chinese descent play an important role in international collaboration between mainland China and the rest of the world. In this paper, we review international collaboration between ethnic Chinese scientists in eight countries – USA, England, Germany, France, Japan, Canada, Australia, and South Korea – and colleagues in China itself. Our analysis shows that while ethnic ties play an important role as a bridge between China and the country of residence, policies of the Chinese government with respect to international collaboration and overseas Chinese reinforce the growth of ethnically based co-authorship.

Keywords

international collaboration; overseas chinese scholars; ethnic ties; government policy; ethnic collaboration index (ECI); knowledge transfer

Introduction

Since opening its doors to the outside world in 1978, China has experienced a series of organizational reforms and policy adjustments. Coincidentally, a large number of Chinese students and scholars have gone abroad, studying in Western technologically developed countries, and engaging in academic exchange and scientific collaboration with international colleagues. These activities also have had a profound influence on the development of China’s science and technology, including international scientific collaboration.

From the beginning of the 1990s, the number of science and technology articles published by Chinese in international periodicals has shown an exponential growth (Jin & Rousseau, 2004, 2005).² This increase brought China into a quantitative expansion phase (Jin & Rousseau, 2005). During the ten year period from 1996 to 2005, the doubling time of all Chinese SCI papers was 3.97 years, indicating that on average every four years the number of published articles has doubled. Among these, articles written by Chinese scientists through international collaboration show a doubling time of 3.81 years, an increase rate slightly higher than that for all Chinese SCI papers.

In 1996, Chinese international collaboration led to the publication of 3,017 papers, reaching 15,069 in 2005, a five-fold increase over ten years. The number of Chinese papers with international collaboration in 2005 alone was almost equal to the sum of all such articles published during the 1994

¹The authors thank Zhang Wang, Zhou Qiuju, Yang Liying, Yang Liangbin and Wang Dan for collecting and counting data. Research for this article has been funded by a Major State Basic Research Special Program China under grant (No 2004CCC00400) and by the US National Science Foundation (OISE-0440422).

² All data used in this article originate from the Web of Science.

– 1997 period. This rapid increase in the number of international collaboration can also be observed from the fact that today, of all Chinese SCI papers, about one in four is internationally co-authored.

Many factors, both internal and external, have caused the developments mentioned above. Besides political, social and geographical reasons (Wagner et al., 2001; Wagner & Leydesdorff, 2005; Katz, 1994; Beaver, 2001; Zitt et al., 2000), we would like to add another one, namely ethnic ties. Our recent survey of Chinese-American collaborative papers in about one hundred periodicals indicates that among 3,603 such papers, 72.3% have at least one author working in the US who is either a Chinese scientist or a scientist of Chinese descent (Jin et al., 2007). This high percentage indicates that overseas Chinese living, studying or working in the United States are playing a vital role in Chinese-American scientific collaboration.

This article expands the scope of the previous investigation to eight countries – USA, Japan, Germany, England, Australia, Canada, France and South Korea – in an attempt to prove that ethnic ties play an essential role in the collaboration pattern of mainland China with other countries. As far as we know, few such studies have been performed before although we have the contribution of B.M. Webster (2004) who, in studying the impact of ethnic minority researchers on the scientific output of the UK, found that ethnic participation in British science varies across different ethnic groups, with the Chinese best represented in relation to their share in the total population. Basu and Lewison (2006) wrote an article, somewhat similar in spirit as ours, studying the output of the scientific community of Indian origin in the USA. Finally, Bassecoulard et al. (2003) studied the scientific production of Madagascar and compared this with that of Malagasy scientists in the diaspora.

Data collection and methods

Target countries

SCI data indicate that for the period from 2001 to 2005, USA, Japan, Germany, England, Australia, Canada, France and South Korea were the top eight countries in terms of the number of collaborative papers with mainland China. Table 1 lists the total number of China's collaborative papers with authors of these eight countries (collaborative articles including authors from China and another country will be referred to as "co-papers"). In 2005, the number of Chinese co-papers totaled 15,069, of which 12,101, or 80.3%, were with these eight countries. The reader may notice that the sum (14,314 for the year 2005) of Table 1 is significantly higher because an article written in collaboration between China and two (or more) of these eight countries is counted twice (or more), namely once for each collaborating country.

Table 1. China's co-papers with its eight most important partner countries (2001-2005)

Partner country	Number of co-papers		Partner country	Number of co-papers	
	2005	2001-2005		2005	2001-2005
USA	5,722	20,815	AUSTRALIA	976	3,837
JAPAN	2,303	9,342	CANADA	1,100	3,831
GERMANY	1,377	5,622	FRANCE	832	2,935
ENGLAND	1,327	4,806	SOUTH KOREA	677	2,377

Ethnic Chinese

The term 'ethnic Chinese' in this article refers to people of Chinese descent living outside mainland China (but not in Taiwan or Singapore). They are recognized by the fact that they have a Chinese family name. They may or may not use an English first name. Ethnic Chinese will also be referred to as 'overseas Chinese'.

Data sources

All data are taken from the Web of Science, the literature type is 'article', and the time span is 2001-2005. Not all records are used however. Details are provided in the following section, where we describe how groups are constructed.

Collaboration types and groupings

The role played by overseas Chinese played in co-papers can be analyzed from several points of view. Therefore, we distinguish the following five types of co-papers.

Type A are those articles where at least one (ethnic) Chinese author has an address outside China. Some co-authors may be 'foreigners'.

Type B articles are those articles written with international collaboration but in which all Chinese authors have only addresses in mainland China. So, these articles do not involve overseas Chinese co-authors as partners, only 'foreigners'.

Type Aa are those articles where at least one ethnic Chinese author has two addresses: one in China and one in another country.

Type C are those articles where all authors are mainland and ethnic Chinese and at least one has an address in another country.

Type D are those where the first author, or the corresponding author is an overseas Chinese.

Figure 1 illustrates the relation between these five types of co-papers.

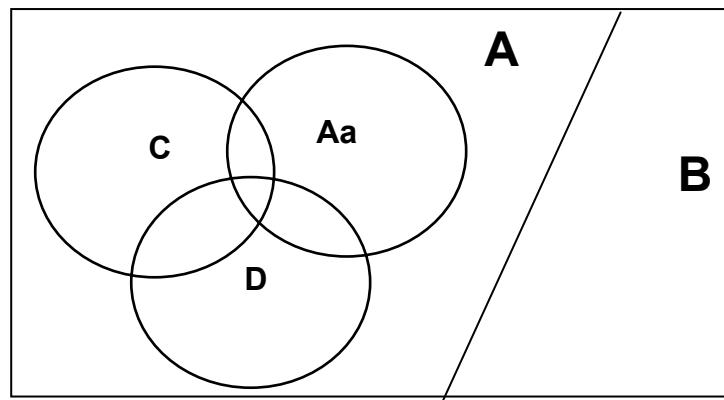


Figure 1. The five types of co-papers

From the different characteristics of the papers surveyed, we considered the following three data sets for analysis.

Set I contains a sample from all co-papers. Two factors are taken into account for determining this sample: the ability to obtain the original articles and the workload since all articles in this sample required visual inspection of the original publication. Articles included in this sample therefore had to satisfy the following three requirements: a) being included in the SCI, b) being available in the National Science Library of the Chinese Academy of Sciences, and c) leading to an adequate sample size. These restrictions are such that we were able to include all articles satisfying our requirements for all the countries other than the USA. For the USA we had to use a partial sample. This data set contains 18,879 papers. If an article involves n collaborating countries then it counts n times.

Set II consists of all type C co-papers (i.e., all authors are ethnic Chinese) involving the eight selected countries and published between 2001 and 2005. This gives a total of 11,782 items in Set II. If an article involves n collaborating countries then it counts n times.

Set III consists of all type D co-papers involving the eight selected countries and published between 2001 and 2005. Articles with more than ten authors (usually big international projects) are omitted from this set. This leads to a total of 9,836 items in which the first or corresponding author is an overseas Chinese. This number again includes double (or more) counting: if an article involves n collaborating countries then it counts n times.

Table 2. The number of items in different data set (2001-2005)

Data sets	Items		Note
Set I	18,879		Sample data
<i>Partner countries</i>	Total co-papers (a)	Sample data (b)	Percentage included: b/a
<i>China with USA</i>	20,242	5,248	25.9%
<i>China with England</i>	7,962	2,053	25.8%
<i>China with Japan</i>	9,362	3,658	39.1%
<i>China with Germany</i>	4,997	2,773	55.5%
<i>China with France</i>	2,880	1,318	45.8%
<i>China with South Korea</i>	2,351	1,007	42.8%
<i>China with Canada</i>	3,749	1,508	40.2%
<i>China with Australia</i>	3,729	1,314	35.2%
Set II	11,782		100 %
Set III	9,836		100 %

Observations

Type A and Type B

The occurrence of overseas Chinese among co-authors determines the distinction between type A and type B. This distinction was made by visual inspection of the published articles. In this way, through many hours of work, we were able to overcome restrictions imposed by the structure of the Web of Science.

From this data, it is possible to introduce an “ethnic collaboration index” (ECI) as the ratio of type A articles over all co-papers, an index which, of course, could in principle be calculated for other ethnic groups as well. Table 3 gives the ECI per country over the period 2001-2005.

Table 3 ECI for ethnic Chinese per country (period 2001-2005)

Country	ECI	Country	ECI	Country	ECI	Country	ECI
USA	0.721	Canada	0.551	Japan	0.475	France	0.303
Australia	0.561	England	0.481	Germany	0.403	South Korea	0.283

As further depicted in Figure 2, China’s collaboration with the USA is more likely to yield Type A articles, while the majority of co-papers with Australia and Canada also are of the A type. Additional data (not included in this article), show that the overall ECI has been increasing over time, especially for the USA, England and Canada. Only Australia and Germany show an opposite trend: there the growth rate of type B articles is larger than that of type A articles, leading to a decrease in ECI.

The relative number of overseas Chinese scholars in different nations and disciplines in each partner country directly affects the ECI. Here, statistics per domain were collected for the fields of physics, chemistry, biology and engineering. Figure 3 illustrates our findings. A value larger than one means that in this domain there are more co-papers with overseas Chinese than for the average of all fields. The opposite is true for a value smaller than one. We observe that engineering always (in all eight countries) has been the field in which more collaboration has occurred between Chinese scientists and their overseas Chinese colleagues than in other fields. Physics is generally weaker (except for Japan), while chemistry is generally stronger. The strength of the overseas Chinese in America and Japan is quite even over the four domains.

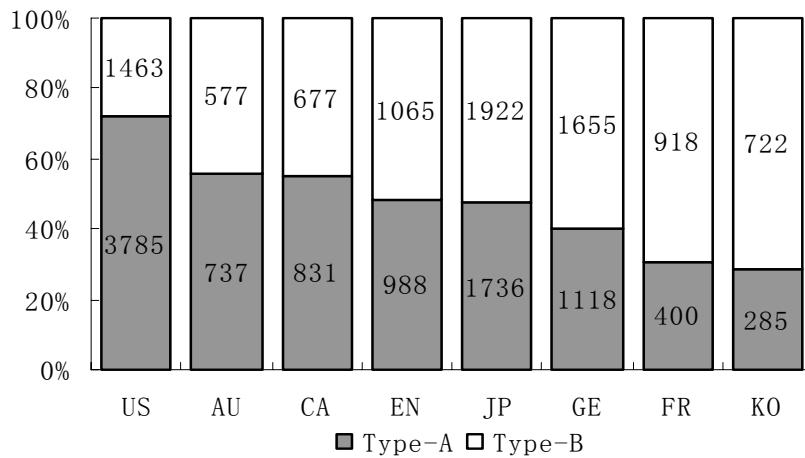


Figure 2. Illustration of the Ethnic Collaboration Index (ECI)

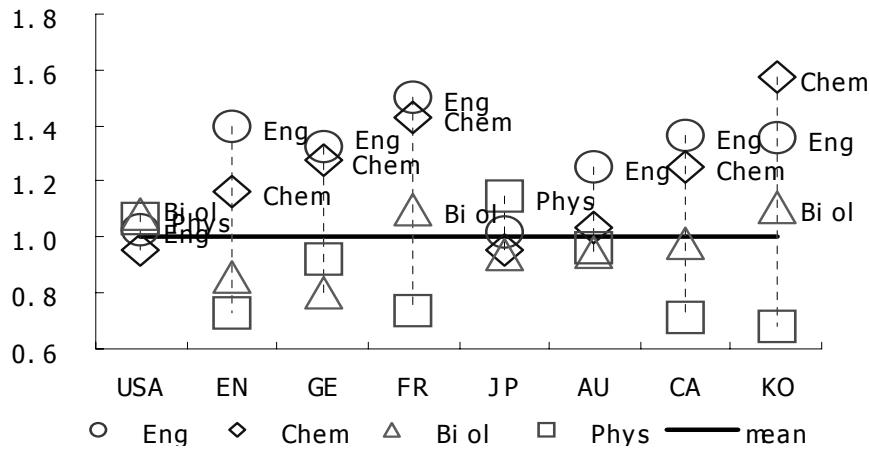


Figure 3. Relative strengths of Overseas Chinese ties in four fields

Type- Aa: papers by amphibious authors

The defining characteristic of Type Aa papers is that they include an overseas Chinese author who has two addresses, one is a Chinese organization address and the other is an organization address in a partner country. We will refer to such an author as an “amphibious author”. In a sense an amphibious author embodies Chinese ethnic collaboration in one person.

Amphibious authors are usually experienced and accomplished scientists. They play an extremely significant role in Chinese international collaboration and in reducing scientific gaps between China and advanced countries. According to our survey (see Figure 4), this type of paper shows a considerable increase in all countries (except for Germany). The largest growth has occurred with Canada (203%), but the USA shows a high growth rate for this type of paper (186 %) as well. In 2005, Type Aa articles yielded more than 50% of all Type A articles in all countries, except England. These data prove that amphibious authors occupy a key position in the international collaboration network of China.

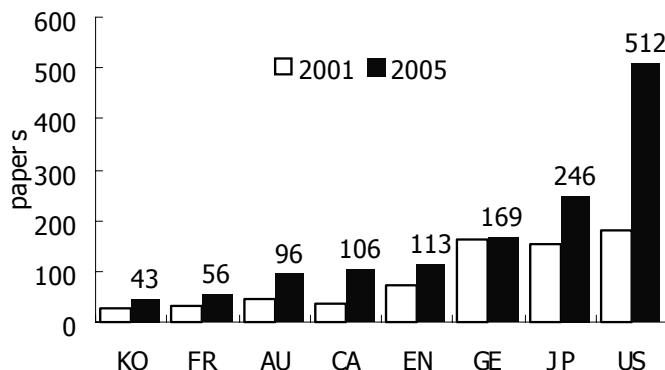


Figure 4. Type Aa articles: a comparison between the year 2000 and the year 2005

Type C

Type C articles are characterized by the fact that all authors are Chinese: some living in mainland China, some living abroad. As there are no non-Chinese authors involved, these articles suggest an independence of ethnic Chinese from their hosts. It is not clear whether this is positive or not.

Table 4. Yearly distribution and percentages of articles of Type C

Country	2001	2002	2003	2004	2005	Type C / All co-papers	
						2001	2005
US	844	945	1243	1464	1942	29.8%	34.7%
CA	75	98	138	139	132	20.8%	32.8%
AU	45	51	58	81	114	22.0%	29.2%
EN	144	215	222	301	373	12.1%	17.9%
FR	114	141	162	201	212	12.1%	13.9%
GE	103	153	215	289	356	14.6%	9.7%
JP	105	163	211	257	280	8.1%	9.1%
KO	21	23	40	57	55	7.3%	8.1%

Citation analysis of such articles (to be performed in a future article) will provide an answer. Table 4 shows that the number of type C articles has increased in absolute terms and relative to the total number of articles with at least one overseas Chinese). As seen in Table 4, this type of article is most prevalent with the United States, closely followed by Canada, and lowest in Japan and Korea.

Figure 5 offers another look at the total set of co-papers. All these articles involve mainland Chinese authors, as well as authors with an address in another country. Articles in the B group do not have ethnic Chinese co-authors, while authors in the C group do not have non-Chinese co-authors. Yet, type C articles form only a minority in the total number of co-papers. Most articles involve ‘foreign’ co-authors as local partners, i.e. scientists who are not Chinese or of Chinese descent. If we denote them as type F, we see that they are always included in type B publications and in all type A publications which are not type C. These foreign co-authors are the mainstream partners for international collaborations with mainland China. Depending on the country between 65 and 92% (see Table 4) of all co-papers involve these local partners.

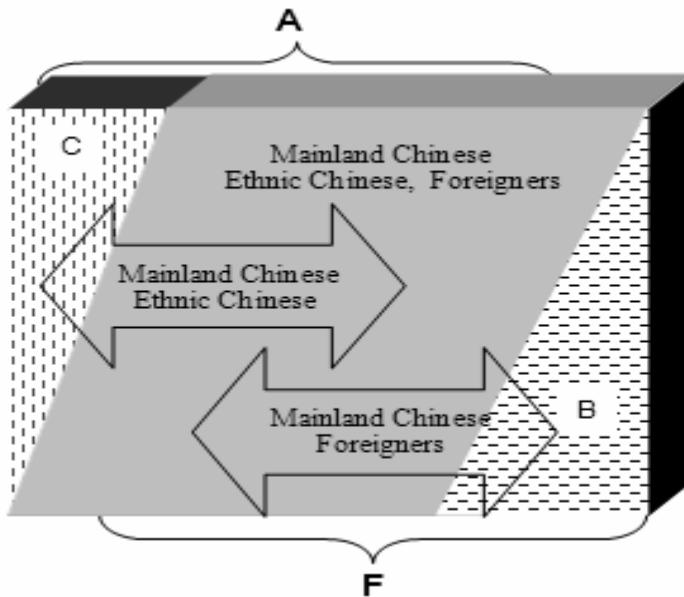


Figure 5 Another look at the relation between different types of articles

Type D articles

Table 5. Distribution of type D articles *

Country	2001	2002	2003	2004	2005	Growth between 2001 and 2005	
						Type D	all co-papers
US	707 (27.2%)	765	994	1195	1286 (26.1%)	0.82	0.90
JP	210 (16.1%)	254	262	309	333 (15.9%)	0.59	0.61
EN	118 (10.6%)	195	165	205	208 (11.6%)	0.76	0.61
CA	96 (21.1%)	126	164	197	207 (22.2%)	1.16	1.05
AU	105 (23.7%)	114	166	191	171 (20.0%)	0.63	0.93
GE	78 (17.0%)	114	155	114	132 (12.1%)	0.69	1.38
FR	32 (10.8%)	33	52	45	61 (10.4%)	0.91	0.98
KO	41 (18.6)	64	55	63	54 (10.1%)	0.32	1.43

* The figure in parentheses is the percentage of Type D articles in the total number of co-papers in that particular country.

Usually the corresponding author is the most important one. Except for certain fields (e.g., mathematics) this is often also the first author. In this section we pay special attention to the share of ethnic Chinese in this ‘backbone’ role. We first note that the absolute number of type D articles increased between 2001 and 2005. Yet, over this five year period, the proportion of type D articles among all co-papers shows a slight decline. When all co-papers are considered, our data show that Type D papers constitute only 19.2% of the total.

Data on Type C articles show that, among the co-papers of which all authors are Chinese (living in mainland China or in another country) a mainland Chinese collaborator is first author in about 74.3% of the cases in 2005. Only in a minority of cases (about 25.7%) is the overseas Chinese the first author. This shows that in this type of collaboration the overseas Chinese does not play the leading role, but merely acts as a bridge between China and his/her country of residence. We may conclude from this observation that ethnic Chinese is one factor of facilitating international contacts between countries.

Explanations of the observed results

Ethnic ties: A power in collaboration

On December 10, 1976, at Stockholm's concert hall, when receiving the Nobel Prize in physics, Samuel C.C Ting (Ding Zhaozhong), a scholar of Chinese descent, delivered his Nobel banquet acceptance speech in Chinese – the first time Chinese was used for such a ceremony. In this way, he hoped that the whole world would hear the voice of a “Chinese nation of science”. Since China’s reform and opening-up, Samuel Ting has been a regular visitor to China, leading to enhanced opportunities in scientific collaboration between China and the United States, which illustrates especially well the importance of ethnic ties.

But, as students of transnational relations are discovering, the concept of ethnic tie is not without considerable ambiguity in today’s globalized world. On one hand, it points to shared identities based on language and culture. But the transnationalism which characterizes most overseas Chinese scientists is not so simple: we see the rise of bi- (or multi-) lingualism, and bi-(or multi-) culturalism further compounded by new forms of professional attachments and changes in immigration status and/or citizenship. In short, transnationalism tends to produce multiple identities.

Ethnic ties can thus be seen as a way of managing these multiple identities. As used here, we view ethnic ties as a form of “homophily,” a principle which states that contact between similar people occurs at a higher rate than among the dissimilar people (McPherson et al., 2001). Thus, on one hand, we see overseas Chinese scientists drawn to established affective relations binding families, relatives, teachers, former collaborators and students together, affective orientations which can and do spill over to a broader sense of service to China’s national development. On the other hand, ethnic ties are also used to advance career objectives, reputation, and prestige. In this study, we are not able to sort out the complex ways in which this happens, but it is clear from the data that ethnic ties – the *Overseas Chinese Phenomenon* – do play a very important role in China’s international scientific collaborations. In this, we believe that China is not unique and look forward to further work on how, and if, ethnic ties play a role in international collaboration for other countries, such as India and South Korea.

Social basis of the Overseas Chinese Phenomenon

The ‘Overseas Chinese phenomenon’ in China’s international collaboration in science and technology has developed from a deep social basis over many years. Chinese scientists leaving China, permanently or temporarily, are certainly influenced by Chinese traditional culture, write the same language, and have close social relations with China. These factors have constructed a social basis binding overseas Chinese to a China with which they share many characteristics. This social basis is also where the opportunity lays for overseas Chinese to carry on international collaboration with China. From the foreign country’s perspective, the presence of a significant community of ethnic Chinese makes it easier to collaborate with scientists in mainland China. Without this community China’s contribution to international collaboration would not be as high as it is today.

At present, there is a large number of overseas Chinese scientists located in many countries. When these scientists retain ties with their country of origin, they are able to contribute to not only the scientific welfare of their adapted country but also that of their country of origin.

Visible hands: policy effect

The appearance of the ‘overseas Chinese phenomenon’ in Chinese international collaboration can be attributed to another important reason, namely the vital role played by the Chinese government (Xiang, 2005) Since the end of 1990’s, the Chinese government has gradually issued a series of policy measures to attract overseas Chinese scholars. Over the years these policies are similar to a visible hand, guiding overseas Chinese scholars to join in Chinese international collaboration. The ‘Overseas Chinese Phenomenon’ observed in this paper can at least partially be explained by the effect of this policy.

Over the period from 1978 until the end of 2005, the total number of Chinese citizens studying abroad reached 933,400, of whom 232,900 students already returned. At present, 512,800 Chinese citizens are abroad studying, conducting collaboration research or on an academic visit (China Education and Research Network, 2006). The Chinese government has adopted many kinds of policies and measures for engaging this group of citizens to participate in scientific research with Chinese colleagues. For example, the *Fund for Distinguished Young Scholars* established by the *National Natural Science Foundation of China*, has funded 431 research projects over the period 2000-2005. Since 1997, the Chinese Academy of Sciences has incorporated the *Outstanding Talents Program* as a part of its *Hundred Talents Program*. By 2004, altogether 850 ethnic Chinese returned to mainland China through this program (Bureau of Comprehensive Planning of Chinese Academy of Sciences, 2004). In 1998 the Ministry of Education started implementing the 'Yangtze River Scholars Program', with the objective of attracting back to China a large number of outstanding young and middle-aged academic professionals, and giving them the opportunity to help China uplift the educational level of all types of institutes for higher education. In 2005, 88 scientists were appointed as short-term professors (for short courses). They came from many countries and regions, but only three of them were of non-Chinese decent. That year the United States alone provided 61 'Yangtze River Scholars'.

Conclusion and discussions

In this article we studied some aspects of China's international scientific collaboration. We believe that the observed data can be explained by two main factors: ethnic ties and the influence of government policies. We do not intend to make an analysis of government policies now, but focus on the issue of ethnic ties instead. The 'Overseas Chinese Phenomenon' refers to the fact that overseas Chinese have become a vital factor in helping Chinese scientists to establish international collaboration channels, and in finding international collaboration partners. The 'Overseas Chinese Phenomenon' in international collaboration seemingly is serving as a mechanism of knowledge transfer in the developmental process of Chinese science, and is likely to remain important even when China is fully integrated into the world of international science.

Our analysis shows that ethnic overseas Chinese play an active role in promoting international collaboration and act as bridges between China and their country of residence (permanent or temporary). Their role helps China to reduce the gaps that still exist between the country and leading developed countries. There are, however, two sides on this coin. As most co-papers involve non-ethnic Chinese from the partner country, China benefits, but developed countries benefit as well as Webster (2004) has argued. The United States of America, in particular, has been the beneficiary of the presence of 62,500 PhDs of Chinese descent working within its borders as of 2003 (US NSF, 2006). As Wagner and Juma (2005) have argued, international scientific cooperation is indeed the key to success in a globalized, networked world, and ethnic ties can be a surprisingly important mechanism in promoting that cooperation.

Is the role of ethnic ties a universal phenomenon existing in international collaboration of all or most developing countries? We hypothesize that it is, and hope that our colleagues looking at scientific development in other countries investigate this further. We even suggest that it is perhaps stronger for developing countries, but that it exists between all ethnic groups, living in a developing country or not. In this article we have shown that ethnic ties are certainly a powerful factor in contemporary Chinese international collaboration. As such it is an example of the homophily principle which states that contact between similar people occurs at a higher rate than among dissimilar people (McPherson et al., 2001).

In this article we have studied quantitative aspects of the Overseas Chinese Phenomenon. As we know, China is in a quantitative expansion phase (Jin & Rousseau, 2005) We have not addressed qualitative issues here, but it would certainly be interesting to investigate whether these collaborative articles are of high quality, if they attract more citations than the average Chinese article, which of the types discerned here attracts the most citations, and from whom. This is left for future research.

References

- Bassecoulard, E, Ramanana-rahary, s. & Zitt, M. (2003). The ultra-periphery of science: three contrasting views of the Malagasy contribution in terms of domestic research, the diaspora and special topics. In G. Jiang, R. Rousseau & Y. Wu (Eds.) *Proceedings of the 9th International Conference on Scientometrics and Informetrics* (pp. 10-21). Dalian: Dalian University of Technology Press.
- Basu, A., & Lewison G. (2006). Visualization of a Scientific Community of Indian origin in the US: A case study of Bioinformatics and Genomics. Paper available at E-LIS, ID code 6268.
- Beaver, d. DB(2001). Reflections on scientific collaboration (and its study): past, present, and future. *Scientometrics*, 52, 365-377.
- Bureau of Comprehensive Planning of Chinese Academy of Sciences (ed.) (2004). *Statistical yearbook of Chinese Academy of Sciences, 2004*. Beijing: Science Press, p.196.
- China Education and Research Network. Retrieved November 15, 2006, from:
<http://www.edu.cn/20060605/3193432.shtml>
- Jin, BH & Rousseau, R. (2004). Evaluation of research performance and scientometric indicators in China. In: *Handbook of Quantitative Science and Technology* (Moed, Glänzel & Schmoch, Eds.). Kluwer: Dordrecht; p. 497-514.
- Jin, BH & Rousseau, R. (2005). China's quantitative expansion phase: exponential growth but low impact. In P. Ingwersen & B. Larsen (Eds.) *Proceedings of ISSI 2005* (pp. 362-370). Stockholm: Karolinska University Press.
- Jin, BH, Suttmeier, R.P., Zhang, W., Cao, C. , Wang, D. & Zhou, QJ. (2007). Sino-US collaboration in science & technology: A bibliometric analysis. *Science Focus* (to appear).
- Katz, J.S. (1994), Geographical proximity and scientific collaboration. *Scientometrics*, 31, 31-43.
- Mcpherson, M., Smith-lovin, L. & Cook, J.M. (2001). Birds of a feather: homophily in social networks. *Annual Review of Sociology*, 27, 415-444.
- US NSF, Division of Science Resources Statistics, Scientists and Engineers Statistical Data System. Retrieved November 15, 2006 from: <http://www.nsf.gov/statistics/>
- Wagner, C.S., Brahmakulam, I., Jackson, B., Wong, A. & Yoda, T. (2001). *Science and Technology Collaboration: Building Capacity in Developing Countries?* MR-1357.0-WB. Santa Monica (CA):RAND, p. XIV.
- Wagner, C. S. & Juma, C. (2005). The Case against Scientific Protectionism. *Scidev.Net Science and Development Network* (31 October 2005).
- Wagner, C.S. & Leydesdorff, L. (2005), Network structure, self-organization and the growth of international collaboration in science. *Research Policy*. 34, 1608-1618.
- Webster, B.M. (2004). Bibliometric analysis of presence and impact of ethnic minority researchers on science in the UK. *Research Evaluation*, 13, 69-76.
- Xiang, B. (2005).Promoting Knowledge Exchange Through Diaspora Networks (The Case of the People's Republic of China). Report to the Asian Development Bank. ESRC Centre on Migration, Policy and Society (COMPAS), University of Oxford. Available at: <http://www.adb.org/Documents/Reports/GCF/reta6117-prc.pdf>
- Zitt, M., Bassecoulard, E. & Okubo, Y. (2000). Shadows of the past in international collaboration: collaboration profiles of the top five producers of science. *Scientometrics*, 47, 627-657.

Is there a Convergent Structure of Science? A Comparison of Maps using the ISI and Scopus Databases[†]

Richard Klavans * and Kevin W. Boyack **

* *rklavans@mapofscience.com*
SciTech Strategies, Inc., Berwyn, PA 19312 (USA)

** *kboyack@sandia.gov*
Sandia National Laboratories, P.O. Box 5800, MS-1316, Albuquerque, NM 87185 (USA)

Abstract

This article compares two maps of science that are built from different, but highly representative sets of the world-wide scientific literature. The analysis in this article extends existing work in this area in three major ways. First, we provide quantitative comparisons of the ISI and Scopus databases for 15 areas of science. Second, we illustrate how these differences have an impact on the resultant map of science. Third, we argue that these differences do not affect the fundamental shape and structure of science; the differences create local differentiation and improve our understanding of local relationships. We conclude with a discussion about the value of generating a convergent map of science.

Keywords

map of science; journal coverage; ISI-Scopus comparison; co-citation analysis

Introduction

Maps of science are visual representations of the relationships between different areas of science. These maps allow us to better understand the relationships between mathematics, physics, chemistry, biochemistry, biology, earth sciences, medical sciences, social sciences, computer sciences and engineering. Accurate maps of science can significantly contribute to our understanding of how science is structured and how it evolves. Maps of science, if they accurately reflect the underlying structure of scientific behavior, can play a central role in education and the communication of scientific issues to the general public.

The intent of this paper is to explore the possibility that there is a convergent structure to science. More specifically, will different databases or methodologies generate pictures representing the structure of science that are structurally equivalent? Convergence, if it exists, can have a significant effect on education. Maps of science can have the same role in education as maps of the world. For example, can you imagine learning about world history without a map of the world? Yet this is what we do today in the sciences; we teach about each area of science as if it exists as an isolated country.

We will start the exploration of convergence by comparing maps based on two databases. These databases have slightly different coverage. The ISI database, which has been the standard in this field for the past 30 years, covers the scientific literature, the social sciences and the humanities. The Scopus database, which is significantly larger in size and scope, covers more of the international literature, more of the engineering literature and excludes the humanities.

The paper is organized into four sections. In the first section, we present our methodology for generating a map or model of science from a database of scientific papers. Using the ISI and Scopus databases we generate two separate maps of science, and then compare them in a qualitative way. We then compare the two maps more quantitatively by analyzing journal and paper coverage for fifteen areas of science. The final section discusses the value of a convergent map of science and the need for additional research on this topic.

[†] Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000. Color versions of all figures are available from the authors.

Mapping Methodology

Maps of science have been published for several decades. Early maps (Griffith, Small, Stonehill, & Dey, 1974) were necessarily small, used severe thresholds, and only represented a small fraction of the available papers. As time has passed, the resources for generating maps have increased such that maps of millions of papers are now possible. For this study, we followed the same basic procedures that we have previously used to generate large maps of science. Although the procedure has been previously described in the literature (Klavans & Boyack, 2006b), we give a brief discussion of each step here.

- 1) Given that our goal is a convergent map of “all of science”, the first step is to identify databases of scientific articles that represent activity in a broad set of scientific disciplines and include extensive citation data. Historically, the ISI databases provided by Thomson Scientific had been the only databases that met this initial criterion. ISI databases cover a broad set of scientific disciplines and have been the standard in the field for conducting international comparisons of scientific publication. Elsevier, the largest publisher of scientific journals, introduced a competitive database in 2005. Their Scopus database covers many of the major scientific journals that ISI covers. In addition, Scopus appears to have greater coverage of selected scientific areas (computer science, engineering, clinical medicine and biochemistry). It is also claimed to have greater coverage of the international literature, especially from Asia and the Far East. We have thus generated individual maps from both the ISI and Scopus citation databases.
- 2) Selecting an appropriate time slice of data is a necessary step in any mapping exercise. There are two main approaches: a narrow time slice, and a broad time slice. A broad time slice assumes that the structure of science is extremely stable over time. Evolution of science is then shown on this structural framework (Chen, 2006). In this study, we use a relatively narrow time slice, specifically because of the belief that the structure of science may not be stable over time. We limit each of our maps to a single year of data, the 2004 indexing year, for both the ISI and Scopus maps. One year is sufficiently long to damp out the effects of single issues of particular journals and different publication rates, but short enough to create a representative map, or snapshot, of science.
- 3) References, or cited papers, were used as the basic unit of analysis. There are three general approaches that are commonly used to generate the structural elements in a map of science. The first is to use the journal as the unit of analysis, and corresponding maps represent the disciplinary structure of science. A second is to use current papers as the unit of analysis; corresponding maps represent themes, or topics of research, for that year of data. A third approach is to use the reference papers as the unit of analysis. In this case the corresponding maps represent paradigms that researchers build upon.

We did not choose to generate a disciplinary map of science in this study. The methodology for generating disciplinary maps has required that each journal occupy one, and only one position (Boyack, Klavans, & Börner, 2005). We believe that this can violate the very nature of the phenomena. Many journals cover multiple disciplines (especially journals where the strategy is to report on developments in all of science). We do not believe that a disciplinary map, based on the restriction of single positions for every journal, will provide the most accurate representation of the structure of science.

In this study, we choose to generate a paradigm map (clustering the references) instead of a thematic map (clustering the current articles). Either would serve the purposes of this study. However, given our assumption that current themes or topics change more rapidly than the underlying paradigms that people use, we expected that thematic maps might be less convergent over time. We plan to explore these issues in future studies.

4) Thresholds are commonly used so that only the most important references are included in a map of science. We use an extremely low threshold² so that all disciplines are well represented. For a discussion of the effect of thresholds on disciplinary bias, see (Klavans & Boyack, 2006b). The threshold resulted in 1,895,118 unique references from the ISI database (out of a possible 12,509,925). The threshold resulted in 2,100,129 unique references from the Scopus database (out of a possible 13,273,040).

5) There are many alternative measures of paper-paper relatedness that have been proposed in the literature (cf. Jones & Furnas, 1987). We use a distance measure that was recently shown to be the most accurate measure available (Klavans & Boyack, 2006a).

6) The references are clustered using the distance measure and an average link clustering algorithm (Klavans & Boyack, 2006b). Each cluster represents a research community – a group of researchers using a specific approach to a problem. These clusters are sometimes referred to as specialities by other researchers.

Previous attempts to cluster extremely large sets of scientific references have used single link clustering because of computational efficiency. Average link clustering is preferable, but requires approximately n^2 calculations. We were able to improve the computation efficiency of an average link clustering algorithm to $n \log n$ time through the use of the distance measure, which is calculated using an interim dimensional reduction step (Klavans & Boyack, 2006a, 2006b).

7) We follow a hierarchical clustering procedure, first suggested by Small (Small, Sweeney, & Greenlee, 1985), to continue to cluster the clusters using the same measure of relatedness and clustering algorithm. In essence, this requires a repeating of steps 5 and 6 above. No information is thrown away at each subsequent level of clustering – the original co-citation counts are aggregated to the appropriate clusters and levels. We stopped the hierarchical clustering when there were less than 1000 nodes (this represented four levels of clustering). This higher level of aggregation represents paradigms.

8) Current papers (those indexed in 2004) are assigned to the paradigms, or clusters of references, using the references in the current papers. 864,961 current papers from 8,408 journals in the ISI database were assigned to the 1,895,118 clustered references. There were 1,081,216 current papers from 11,877 journal or conference titles assigned to the 2,100,129 clustered references in the Scopus database.

9) A visualization algorithm was used to generate a layout of paradigms (Davidson, Wylie, & Boyack, 2001; Klavans & Boyack, 2006b), thus creating visual maps for the ISI and Scopus models. We selected an edge cutting setting that generated similar pictures, in terms of white space and node spacing, from both databases. Pajek (Batagelj & Mrvar, 1998) was used to generate the final pictures of each map.

Maps of Science

Figure 1 is a comparison of the maps of science from the Scopus and ISI databases for 2004. The nodes represent paradigms, or clusters of references. The size of the node corresponds to the number of current papers that were assigned to each of the paradigms. The lines between nodes represent strong relationships between clusters of scientific references. Only the primary relationships, selected by the visualization software, are shown in these graphs.

² For references published in the year prior to the indexing year (in this case, 2003), any reference with 3 or more co-citations with any other reference is included. For all older references, a threshold of 4 co-citations is used.

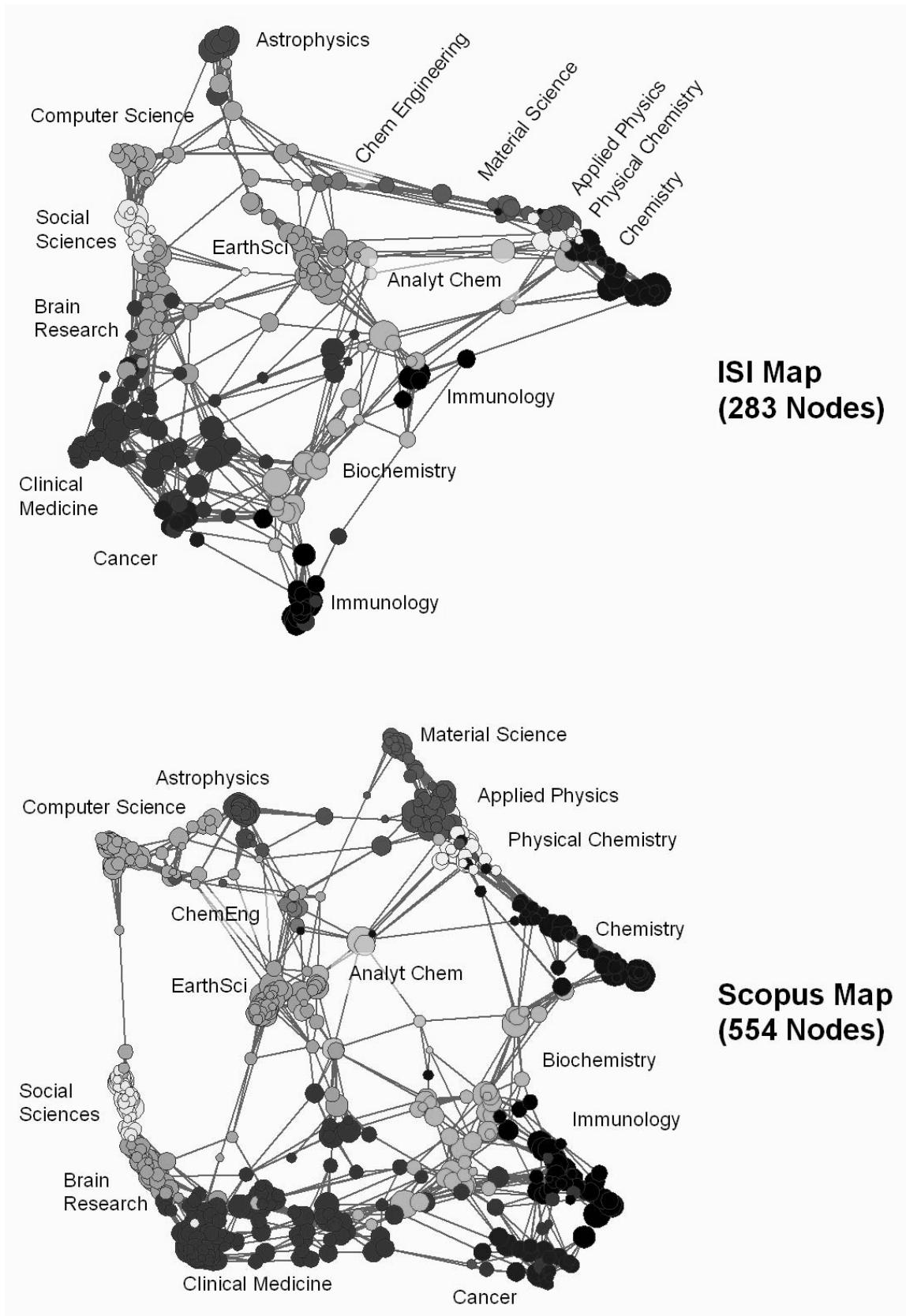


Figure 1. A comparison of the ISI and Scopus maps of science.

The ISI map has 283 nodes, while the Scopus map has significantly more (554 nodes). It is likely that several factors contributed to the Scopus map having twice as many nodes as the ISI map. First, there were 11% more reference papers in the Scopus database that met the co-citation threshold. In the absence of any other major differences, we would expect the Scopus map to have ~11% more nodes than the ISI map. Yet, there are other differences. Primary among these is the distribution of scientific vs. technical journals in the two databases. We expect that reference papers from technical journals tend to form smaller clusters than those from scientific journals. This may be so for several reasons – the technical literature 1) has fewer cites per paper, on average, than the scientific literature, 2) is more specialized and thus cites more of the periphery and less of the core scientific base, and 3) may cite more work from smaller journals than does a paper that is more scientific in nature. These factors combine to generate a reference map (the Scopus map) with significantly increased differentiation of the scientific literature.

In order to compare the maps more easily, we decided to split them up into multiple categories, so that the category sizes, shapes, and connections could be visually compared. Each map was divided into 15 different areas using a manual process of examining the paradigms, their dominant journal constituents, and the distribution of journals from current papers assigned to the paradigms. Each paradigm was manually assigned to one of the 15 categories.

The relative locations of disciplines that appear in the upper part of the map (*Computer Science*, *Astrophysics*, *Material Science*, *Applied Physics*, *Physical Chemistry*, *Chemical Engineering*, *Chemistry* and *Analytical Chemistry*) are shown in Figure 2. The first pair of maps in Figure 2 illustrates how the databases generate slightly different shapes and connections for *Computer Science*. The ISI database suggests that *Computer Science* is more connected and closer to the shape right below it (*Social Science*). The Scopus database suggests the opposite – *Computer Science* is more distant and less connected to *Social Science*, with a branch that is tightly linked to *Social Science* and a separate area that is located between *Clinical Medicine* and *Earth Science*. This difference may be due to the inclusion of proceedings. The ISI database does not include proceedings. The remaining journals in *Computer Science* have a strong relationship with *Social Science* (as shown on the left). The proceedings literature in *Computer Science*, which is significantly larger than the journal literature in *Computer Science* (Boyack, 2007; Gläzel, Schlemmer, Schubert, & Thijs, 2006), has a very weak and distant relationship with the *Social Sciences*. The Scopus Map illustrates the effect of adding these two literatures together. First, the proceedings literature links tightly with journal literature (as expected). But once combined, there is only a small section of *Computer Science* that is close to social science. The overall effect is to make these areas more distant from one another.

The shape and relative location of *Applied Physics* also differs significantly in the two maps. In Figure 2, the ISI map suggests that *Applied Physics* is a cluster of nodes that are separated from *Computer Science* and *Astrophysics* by *Chemical Engineering* and *Material Science*. The Scopus map suggests that *Applied Physics* plays a larger and more central role. There is a branch of *Applied Physics* that connects more directly to *Astrophysics*. *Material Science* appears on one side of *Applied Physics* (directly above), while *Chemical Engineering* appears off to one side. The differences in these two shapes seem to be a result of the inclusion of more journals and proceedings in *Applied Physics*, especially from Asia and the Far East. Greater coverage in *Applied Physics* results in morphological changes in the shapes shown in Figure 2.

Material Science, *Applied Physics* and *Physical Chemistry* also have different locations in the two maps. These disciplines are located along a line in upper right of the ISI map. The Scopus map, however, pulls out *Material Science* as a more distinct group, has a much larger domain for *Applied Physics*, and makes a wider separation between *Applied Physics* and *Physical Chemistry*. The remaining disciplines in the upper right part of these two maps – *Chemistry*, *Chemical Engineering* and *Analytical Chemistry* – are very similar in location and shape.

Figure 3 further illustrates how the maps are similar in the relative location of disciplines. All of the seven disciplines listed in this figure, which shows the lower portion of both maps, have the same

relative placement. The only significant difference in location is *Cancer*. ISI places *Cancer* close to *Clinical Medicine*. Scopus suggests that *Cancer* is more differentiated and linked more tightly to *Immunology*. There is also a significant difference in the shapes for *Immunology*. ISI generates two smaller shapes associated with *Immunology*, both of which are branched off of *Biochemistry*. The Scopus database shows a much larger and interconnected shape for *Immunology*.

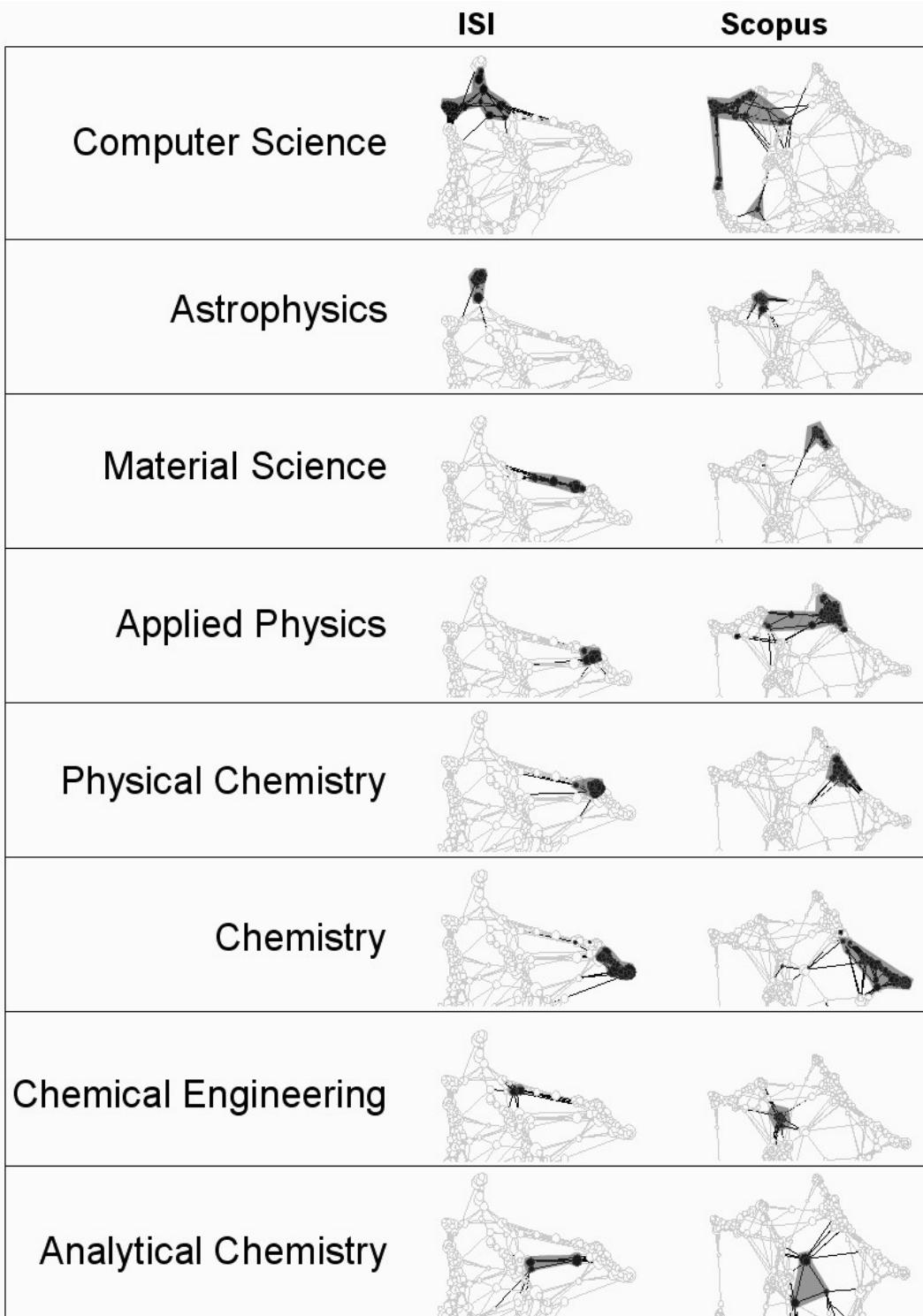


Figure 2. Locations and sizes of selected disciplines from the upper sections of both maps.

Additional research is needed to determine the reasons for the differences noted here. At this stage of the analysis, we suspect that the differences are due to aggregation and coverage. The more aggregated categories in the ISI map can easily hide the detail that would show relationships that appear in the Scopus Map. Disaggregating the larger nodes may help to reveal these relationships. We also suspect that lower coverage of an area of science will tend to result in a more distorted map; a greater coverage should reveal a more accurate picture of the shape and structure of science. The following section explores these issues more quantitatively.

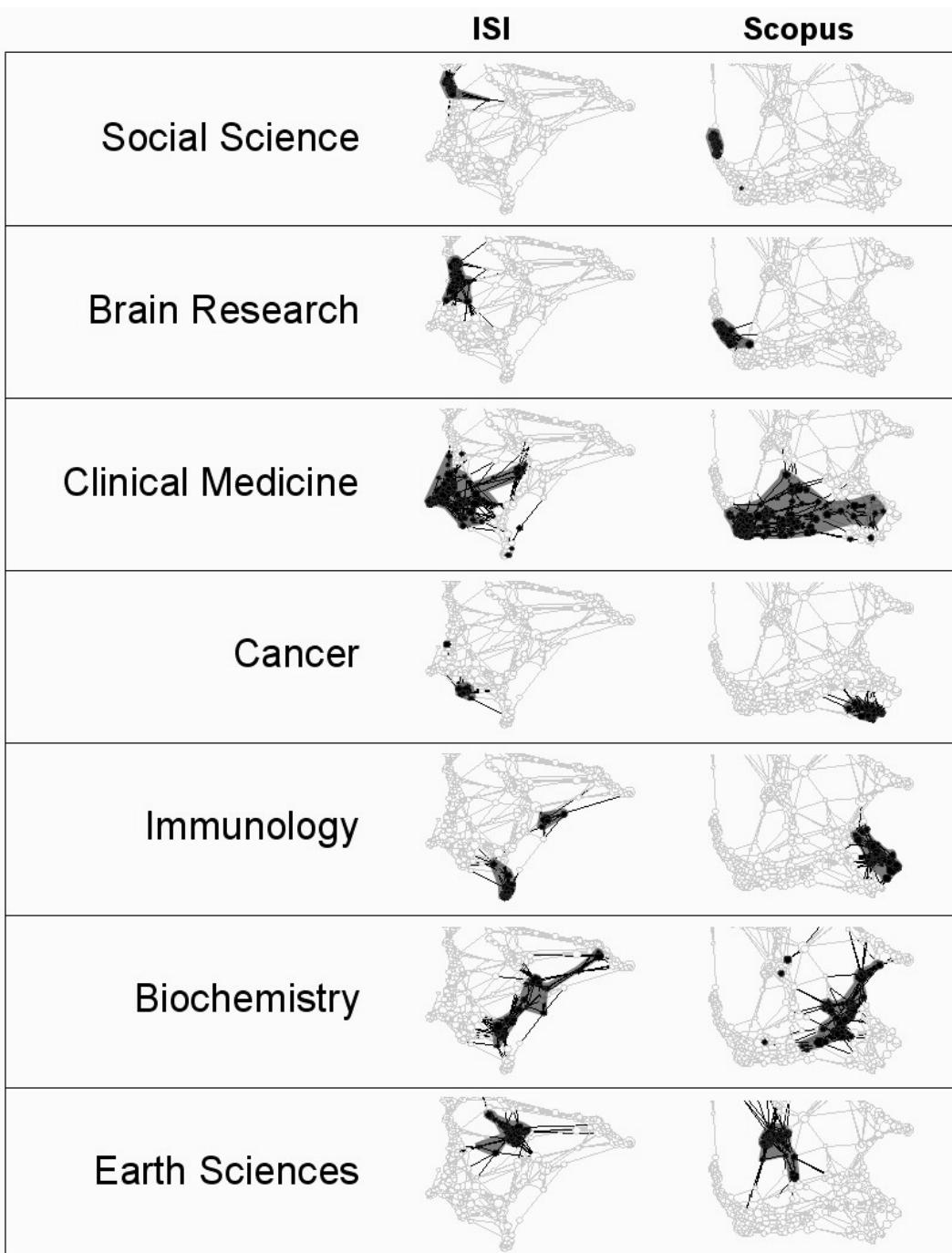


Figure 3. Locations and sizes of selected disciplines from the mid and lower sections of both maps.

Quantitative Analysis

As noted above, we based our study on data from two sources: ISI and Scopus. For the ISI map, we used the combined 2004 citation indexes (Science, Social Science, and Arts & Humanities) from Thomson Scientific. These databases cover approximately 9,000 journals, of which 8,408 were

represented in the current paper assignments to our ISI paradigm map (see step 8 in the methodology section). However, these databases have only limited coverage of conference proceedings (especially proceedings in computer science). ISI is very restrictive in which journals they include in this database. Journals without sufficient evidence of scientific merit (such as the lack of a peer review procedure) are not included. While there may be controversies about the inclusion or exclusion of individual journals, there has been general consensus that the ISI database has a highly representative set of world-wide scientific literature.

It is important to note that ISI does have a separate Proceedings database that was not included in this study. There were two reasons for this exclusion. First, although the coverage of this database has been studied (Glänzel et al., 2006), it is not yet a standard procedure to include it in science maps. We are aware of only one instance in which the ISI Proceedings database has been included in a map of all of science (Boyack, 2007). The second reason is cost. The databases from ISI are costly, and there was not sufficient budget to include the Proceedings database in this study.

The 2004 Scopus database is larger. The number of titles (journals, proceedings, trade publications or book series) is far greater, over 14,000, of which 11,877 were represented in the current paper assignments to our Scopus paradigm map. The Scopus database was recently introduced into the marketplace and has not been subjected to the same critical analysis as the ISI database. We know relatively little about what it includes or excludes, except for claims and counter-claims in the press and a few preliminary comparisons (Jacso, 2005 and references cited therein).

We compared the ISI and Scopus journal coverage by matching journal titles between our ISI and Scopus science maps. The results shown in Figure 4 are preliminary, as we expect to find more matches over time, but are representative of the overlapping and unique coverage of the two data sources. Of the journals that are included in the current paper assignments in our models, there are 6,887 in common between the two sources. Thus, 82% of the ISI titles are covered by Scopus. Of the remaining 18%, nearly 10% are from the Arts & Humanities index, leaving only 694 science and social science journals that are not found in the Scopus model. Of the titles that are unique to Scopus, 3,910 are journal titles (based on information found at the Scopus website), and 1,080 are conference proceedings, trade publications, book series, etc. Two additional notes are in order. First, the overlap calculations are based on only those journals for which the databases have cited references. Scopus indexes approximately 1,400 Medline titles for which it does not have the cited references. Thus, those journals are not included in the Scopus map and overlap calculations. Second, the ISI Proceedings database typically indexes 1,200+ conference titles each year. We presume that there would be substantial overlap between the ISI and Scopus conference titles, at least among the larger conferences. If the ISI Proceedings database were added to our ISI map, it would likely balance out the conference coverage between the two models.

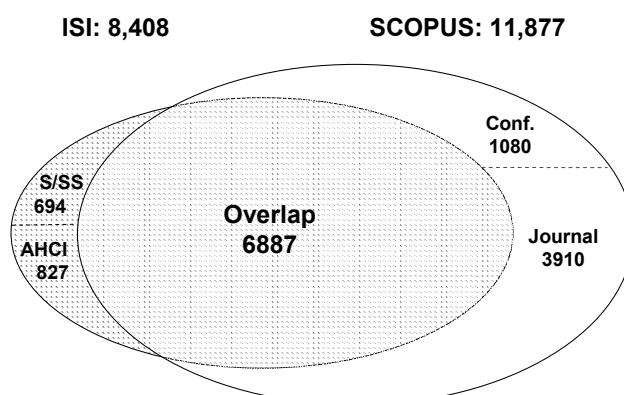


Figure 4. Overlapping and unique coverage of the ISI and Scopus databases for 2004. Only those journals included in the current paper assignments of our models are included in these numbers.

In addition to overall numbers of journals, Table 1 compares the numbers of journals and articles for the fifteen broad scientific areas shown in Figure 1. The areas where the ISI database has greater coverage are at the top, while the areas where the Scopus database is stronger are at the bottom. ISI has more journals and papers in two of fifteen areas of science: *Social Sciences* and *Astrophysics*. There are three areas in which the coverage is roughly comparable, more so in numbers of papers than journals: *Immunology*, *Chemistry*, and *Brain Research*. In the other ten areas, the Scopus database has substantially greater coverage. At the extreme is *Computer Science*, where the Scopus map has 816 more journals and 63,808 additional articles published in 2004 than the ISI map.

Table 1. Journal and current paper coverage for the Scopus and ISI models by scientific area.

Area	Journals Scopus	Journals ISI	Diff (SC-ISI)	Current Scopus	Current ISI	Diff (SC-ISI)
<i>Social Sciences</i>	2,338	2,350	-12	76,231	79,260	-3,029
<i>Astrophysics</i>	102	122	-20	29,777	30,102	-325
<i>Immunology</i>	569	434	135	72,760	71,516	1,244
<i>Chemistry</i>	427	360	67	69,250	67,135	2,115
<i>Brain Research</i>	747	655	92	66,140	62,829	3,311
<i>Materials Science</i>	465	261	204	32,121	27,360	4,761
<i>Analytical Chemistry</i>	88	70	18	15,959	11,185	4,774
<i>Chemical Engineering</i>	221	99	122	15,901	6,985	8,916
<i>Earth Sciences</i>	1,540	1,105	435	107,377	96,326	11,051
<i>Applied Physics</i>	371	266	105	83,680	69,403	14,277
<i>Cancer</i>	385	175	210	39,440	24,093	15,347
<i>Biochemistry</i>	556	357	199	77,639	55,151	22,488
<i>Physical Chemistry</i>	234	108	126	52,384	26,570	25,814
<i>Clinical Medicine</i>	2,181	1,209	972	206,818	165,115	41,703
<i>Computer Sciences</i>	1,653	837	816	135,739	71,931	63,808
All Areas	11,877	8,408	3,469	1,081,216	864,961	216,256

We have labeled the first area in Table 1 as *Social Sciences*. However, this means two different things in our two maps. In the ISI map, the majority of the Arts & Humanities journals and papers are located in the *Social Sciences* area. However, the Scopus map does not include any Arts & Humanities information. Since the two maps have nearly identical numbers of journals and papers in their *Social Sciences* areas, this suggests that the Scopus map has, in fact, greater coverage of the social sciences that is roughly equal in size to the Arts & Humanities portion of the ISI map. That the Arts & Humanities would cluster with the social sciences is not surprising given the category map of Moya-Anegón et al. (2004) showing arts, history, and philosophy as an appendage attached to the social sciences.

Table 2 compares the number of current papers and reference papers for the two maps shown in Figure 1. Here we introduce the concept of reference intensity, which is the number of reference papers per current paper in a particular area of the map. Reference intensity can be calculated at multiple levels. For instance, it can be calculated for each of the paradigms (283 in the case of the ISI map), or it can be calculated for the 15 areas of science identified in our maps.

Overall, the map based on the Scopus database has more current papers and references than the ISI database. The overall reference intensity (for the entire map) of the ISI data is higher. We suspect that reference intensity has an important impact on the aggregation of papers into paradigms. Higher reference intensity seems to result in greater aggregation (more links result in the algorithms deciding that there are fewer clusters). The difference in reference intensity may, in part, explain some of the structural differences noted in the previous section.

Table 2. Reference intensity for the Scopus and ISI models by scientific area.

Area	Refs Scopus	Refs ISI	Current Scopus	Current ISI	R/C Scopus	R/C ISI	Diff
<i>Brain Research</i>	179,162	56,590	66,140	62,829	2.71	0.90	-1.81
<i>Astrophysics</i>	73,731	41,746	29,777	30,102	2.48	1.39	-1.09
<i>Chemistry</i>	141,384	113,873	69,250	67,135	2.04	1.70	-0.35
<i>Earth Sciences</i>	237,687	190,873	107,377	96,326	2.21	1.98	-0.23
<i>Clinical Medicine</i>	485,165	384,014	206,818	165,115	2.35	2.33	-0.02
<i>Immunology</i>	183,074	181,258	72,760	71,516	2.52	2.53	0.02
<i>Social Sciences</i>	119,615	128,082	76,231	79,260	1.57	1.62	0.05
<i>Analytical Chemistry</i>	29,679	24,791	15,959	11,185	1.86	2.22	0.36
<i>Physical Chemistry</i>	90,937	64,236	52,384	26,570	1.74	2.42	0.68
<i>Applied Physics</i>	103,855	149,125	83,680	69,403	1.24	2.15	0.91
<i>Computer Sciences</i>	119,901	135,971	135,739	71,931	0.88	1.89	1.01
<i>Biochemistry</i>	200,249	207,510	77,639	55,151	2.58	3.76	1.18
<i>Materials Science</i>	33,173	62,810	32,121	27,360	1.03	2.30	1.26
<i>Chemical Engineering</i>	18,483	21,594	15,901	6,985	1.16	3.09	1.93
<i>Cancer</i>	83,336	131,077	39,440	24,093	2.11	5.44	3.33
All Areas	2,099,431	1,893,550	1,081,216	864,961	1.94	2.19	0.25

There is a significant variance in reference intensity for different areas of science, and the two databases do not agree on which areas have higher and lower reference intensities. (The correlation between the reference intensity for the two databases is not significant.) There does seem to be a tendency for a drop in reference intensity when additional current papers are covered. This suggests that the policy of focusing on a smaller set of highly linked documents will result in a smaller set of highly linked clusters (consistent with the ISI map). The policy to include more of the less-linked documents will result in an addition of smaller, less linked clusters (consistent with the Scopus map). Reference intensity may help to explain the structural differences in the two maps noted previously.

Table 3 shows another asymmetric pattern in journal coverage. We grouped journals according to the country associated with the publisher (using data from the Scopus web site). We then focused on two groups of nations: the major English-speaking nations (U.S., UK, and Australia) and Asia/Far East nations (16 nations – the largest being China, Japan, Russia, India, Singapore, Korea, Taiwan and Hong Kong). The overall pattern suggests that ISI focuses on journals that are published in the major English-speaking nations. Scopus has much better coverage of the journals where the publisher is located in Asia and the Far East. This pattern of geographic emphasis, however, is very sensitive to the area of science. Scopus' increased coverage of Asia/Far East journals is particularly apparent in *Chemical Engineering*, *Physical Chemistry*, *Materials Science* and *Cancer*. There are negligible differences in the percentages of *Applied Physics*, *Chemistry* and *Astrophysics* journals that are published in Asia/Far East. The actual number of journals published in Asia/Far East in these four areas, is roughly 30% greater due to the greater overall coverage of the Scopus database.

Summary

The maps presented in this paper are convergent in that there are no fundamental differences in the underlying structure of science. The relationships between basic areas of science remain quite similar. Any differences appear to be the result of coverage, reference intensity and aggregation. It appears that increased coverage (especially of proceedings and publications from smaller nations) lowers reference

intensity, resulting in a more disaggregated map that more accurately describes how world-wide science is structured.

Table 3. Journal coverage by area and publisher location for the Scopus and ISI models.

Area	%English Scopus	%English ISI	Diff (SC-ISI)	%FarEast Scopus	%FarEast ISI	Diff (SC-ISI)
<i>Social Sciences</i>	82.16	86.86	-4.70	1.75	0.73	1.02
<i>Brain Research</i>	71.55	79.64	-8.09	3.21	0.91	2.30
<i>Immunology</i>	62.97	75.06	-12.09	5.64	2.64	3.00
<i>Chemical Engineering</i>	61.74	74.36	-12.61	17.45	2.56	14.89
<i>Analytical Chemistry</i>	65.38	71.19	-5.80	5.13	6.78	-1.65
<i>Applied Physics</i>	67.84	70.21	-2.37	16.47	15.74	0.73
<i>Clinical Medicine</i>	58.82	70.03	-11.21	8.68	4.03	4.65
<i>Materials Science</i>	61.76	69.96	-8.19	17.35	10.76	6.59
<i>Cancer</i>	56.79	69.82	-13.03	9.51	2.96	6.55
<i>Biochemistry</i>	62.35	67.67	-5.32	9.61	8.46	1.15
<i>Computer Sciences</i>	58.73	65.91	-7.18	12.21	6.98	5.23
<i>Physical Chemistry</i>	62.64	65.22	-2.57	20.11	13.04	7.07
<i>Earth Sciences</i>	54.86	64.19	-9.33	9.35	6.09	3.26
<i>Chemistry</i>	61.63	62.50	-0.87	14.83	15.71	-0.88
<i>Astrophysics</i>	54.55	55.34	-0.79	20.45	19.42	1.04

These findings, however, are limited. We do not know if maps of different time periods are sufficiently convergent that we can differentiate superficial changes (those based on the differences in coverage) from more fundamental changes (those reflecting underlying changes in the structure of science). Nor do we know if thematic maps (those built from clustering the current literature) are sufficiently convergent with paradigm maps (those built from clustering the reference literature) in a way that would enable differentiation of superficial relationships from more fundamental relationships. The existence of a convergent map that can be a shared cognitive framework for understanding the structure of science is still uncertain.

We argue, however, that this direction of research is extremely valuable, especially if there is continued evidence of convergence. Convergent maps can become a shared cognitive framework that, once learned, can provide the context for better understanding of divergent maps (e.g. maps of the same phenomena but with different layouts). An example may help to illustrate the importance of a convergent map. A map of the world, showing the location of continents and oceans, is an example of a convergent map. It is a shared cognitive framework that is quite adequate for most applications (world history, contemporary policy issues, etc.). While this is a shared framework that closely corresponds to one aspect of our world, one could argue that the map of the world is not adequate if one is interested in understanding continental drift. In this case, divergent maps (showing the initial land mass of Pangea and then the breakup of the continents might be more useful. Or, one could argue that the map is inaccurate if we were interested in population sizes or wealth. The world map has been redrawn to illustrate this phenomena. Each of these divergent maps, however, assume a core map that we are all familiar with and which, once learned and remembered, we can use as a framework for understanding the divergences.

Convergent maps of science can become a critical teaching aid. They can help us understand the intellectual neighbourhood in which we exist. They can help show the directions that a set of researchers are pursuing and point out fruitful directions for future research. They can provide fundamental insights into patterns of influence and expansion. And they can contribute to our basic understanding of how the shape and structure of science has changed over time.

References

- Batagelj, V., & Mrvar, A. (1998). Pajek - A program for large network analysis. *Connections*, 21(2), 47-57.
- Boyack, K. W. (2007). *Using detailed maps of science to identify potential collaborations*. Paper presented at the 11th International Conference of the International Society for Scientometrics and Informetrics.
- Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3), 351-374.
- Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377.
- Davidson, G. S., Wylie, B. N., & Boyack, K. W. (2001). Cluster stability and the use of noise in interpretation of clustering. *Proceedings IEEE Information Visualization 2001*, 23-30.
- Glänzel, W., Schlemmer, B., Schubert, A., & Thijs, B. (2006). Proceedings literature as additional data source for bibliometric analysis. *Scientometrics*, 68(3), 457-473.
- Griffith, B. C., Small, H. G., Stonehill, J. A., & Dey, S. (1974). Structure of scientific literatures. 2. Toward a macrostructure and microstructure for science. *Science Studies*, 4(4), 339-365.
- Jacso, P. (2005). As we may search – Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89(9), 1537-1547.
- Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38(6), 420-442.
- Klavans, R., & Boyack, K. W. (2006a). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2), 251-263.
- Klavans, R., & Boyack, K. W. (2006b). Quantitative evaluation of large maps of science. *Scientometrics*, 68(3), 475-499.
- Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E., & Muñoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61(1), 129-145.
- Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the Science Citation Index using co-citations. II. Mapping science. *Scientometrics*, 8(5-6), 321-340.

Long-Term Patterns in the Aging of the Scientific Literature, 1900–2004¹

Vincent Larivière*, Éric Archambault** and Yves Gingras***

**lariviere.vincent@uqam.ca*

Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, CP 8888, Succursale Centre-ville, Montréal, Québec, H3C 3P8 (Canada) and

Graduate School of Library and Information Studies, McGill University, Montréal, Québec (Canada)

***eric.archambault@science-metrix.com*

Science-Metrix, 4572 Avenue de Lorimier, Montréal, Québec H2H 2B5 (Canada) and

Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, Montréal, Québec (Canada)

****gingras.yves@uqam.ca*

Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, CP 8888, Succursale Centre-ville, Montréal, Québec, H3C 3P8 (Canada)

Abstract

Despite a very large number of studies on the aging and obsolescence of scientific literature, no study has yet measured, over a very long time period, the changes in the rates at which scientific literature becomes obsolete. This paper aims at studying the evolution of the aging phenomenon and, in particular, how citation half-lives are changing over more than 100 years of scientific activity. It shows that the average and median age of cited literature has undergone several changes over the period. Specifically, the two World Wars had the effect of raising the average and median age of the cited literature significantly. Moreover, and contrary to a widely-held belief, the age of cited material has risen continuously since the mid-60s. Among the possible explanations for this counter-intuitive phenomenon, the most probable is the levelling off of the growth of scientific literature related to the steady-state dynamics of modern science that follows its exponential growth.

Keywords

aging; half-life; obsolescence; Price index; diasynchronous analysis; scientific publications.

Introduction

The typical citation *life-cycle* of scientific papers starts with a fast increase during their initial years on the scientific scene, followed by a peak, and then a slow but steady fall into oblivion or are incorporated in the canon of normal science. This short and intense life and its subsequent aging process have always fascinated information scientists and bibliometricians, going as far as the seminal paper by Gross and Gross (1927). Since then, there has been a very large number of studies on aging and obsolescence (see the extensive reviews by Line and Sandison, 1974 and Line, 1993), most of them made using library loans and citation indexes and, more recently, with document download data (Nicholas et al., 2006). Despite the important body of literature on the topic, no study has yet measured on a global scale how the aging process of scientific literature has changed with time.

It has been suggested that, given the accelerated pace of scientific development, the scientific literature becomes more rapidly obsolete (Line 1970, 1993; Price, 1963, 1965). Knowledge is more rapidly disseminated with electronic means and, thus, one might expect that the useful life of scientific literature gets shorter. On the other hand, others (e.g., Odlyzko, 2002) have suggested that these electronic means and online bibliographic databases would have exactly the opposite effect—that is, authors would increasingly refer to older material. This paper examines these diverging hypotheses in order to determine whether the scientific literature is becoming more rapidly obsolete or if, on the

¹ The authors wish to thank François Vallières and Gilles Renaud for the construction of the bibliometric databases and Jean-Pierre Robitaille, Professor Jamshid Beheshti and the two anonymous reviewers for their valuable comments and suggestions.

contrary, it is increasingly being referred to for longer periods of time. In order to measure this phenomenon, data on the average and median of cited literature are compiled over the 1900 to 2004 period. Overall, this paper aims at providing a better understanding of the aging process of scientific literature and of the changes it has undergone over the last 100 years.

Methods

In their 1974 review paper, Line and Sandison have characterized three types of obsolescence studies: diachronous, synchronous, and diasynchronous. While diachronous studies follow the citation of specific documents through time, synchronous studies analyse the age distribution of cited documents at a given time. Finally, diasynchronous studies compare the age distribution of cited documents at different time periods, thus allowing for the measurement of changes in the aging process of literature (Line and Sandison, 1974). This paper belongs to the diasynchronous type of study, since it aims at analysing the evolution of yearly synchronous scores computed over the 1900–2004 period.

This paper uses data from Thomson Scientific, which is the only organization having indexed citations from scientific sources over more than 100 years. For each document indexed in Thomson's databases (source items), a list of references is included. Data between 1900 and 1944 are drawn from the Century of Science, which indexes 266 distinct journal titles covering most natural sciences and medical fields. From 1945 to 1979, data are from the Web of Science, Thomson Scientific's online bibliographic database. Finally, from 1980 to 2004, data are from the CD-ROM version of the Science Citation Index, Social Sciences Citation Index, and Arts and Humanities Citation Index.

In order to mitigate the effect of errors in the data, a 100-year and a 20-year citation window were used. The 100-year citation window proved to be the best equilibrium between minimizing errors in cited document years and maximizing the number of references. However, taking into account the potential effects that a 100-year citation window might have on the computation of some indicators (for instance, the average life is highly influenced by citations to older documents), we also used a 20-year citation window, which is the citation window used by Thomson Scientific in compiling their half-life measures. Finally, all statistics in this paper are based on the three standard types of documents: Articles (including notes), Reviews, and Meeting Abstracts.

Results

Figure 1 presents the evolution over a century of the number of scientific papers included in Thomson Scientific's databases together with the number of references made in these source papers. One can immediately see two salient features of this dataset: the publication of scientific research slowed considerably during each of the two world wars. The other important feature, although less salient, is the progressive slowing down in the growth of scientific production in this dataset after 1980. However, for this same dataset, the number of references has not been levelling off, although there is a slight decrease in the growth rate of the number of references starting around 1985. As predicted by Price (1963)—and as common sense would expect—the exponential growth could not last forever, and this Figure shows that the growth has indeed started to level off around 1980.

Figure 2 presents the evolution for the 1900–2004 period of the average age and median age of cited literature—the latter being equivalent to the citing half-life—for citation windows of 20 and 100 years. One readily notices that there is an important difference between the average age and the median age—especially for the 100-year citation window. All curves show that the two world wars had a significant effect on the age of cited literature, increasing both the average and the median age of about 1.5 to 2 years. The cause of this effect is quite obvious: as the number of papers published decreased, for obvious reasons, during the two wars (Figure 1), researchers relied more on papers published before the wars, which, in turn, increased the age of cited literature.

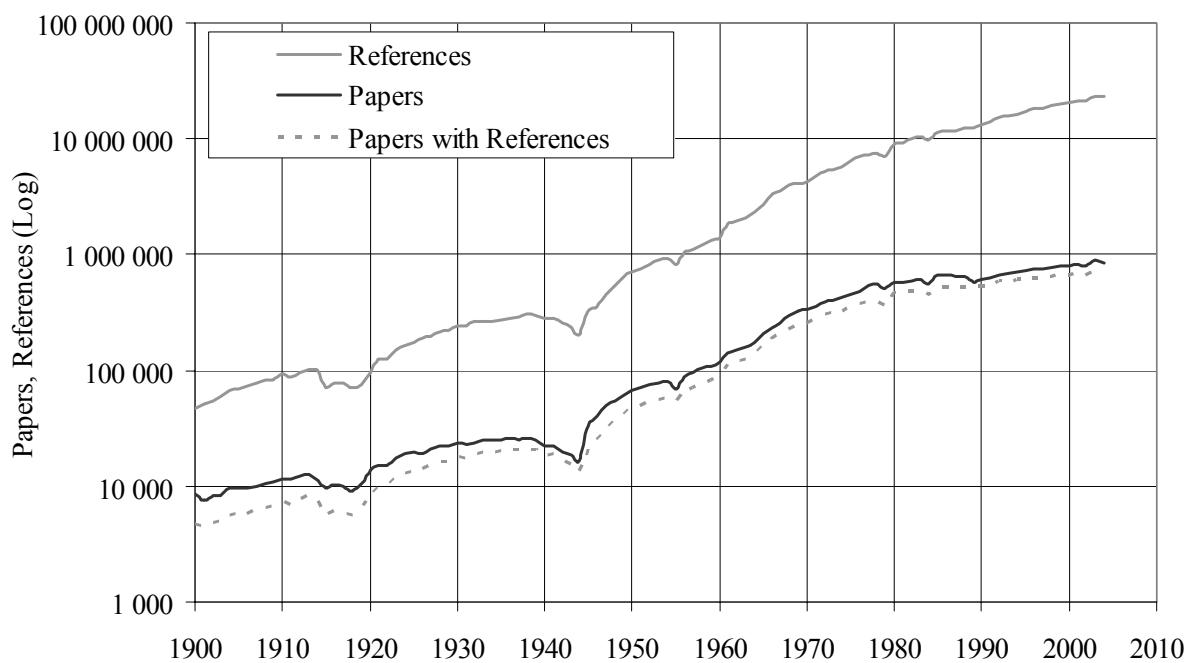


Figure 1. Number of papers, number of papers with at least one reference, and number of references, 1900–2004

The data contained in Figure 2 also show that, between 1945 and 1975, the average age of cited literature (100-year window) has been steadily decreasing, from an average of almost 12 years to 9 years. With the 20-year citation window, this decline in the age is also observed, albeit for a shorter period (1950–1960). This decrease is not observed, however, in the median age curves, which increased by about 2 years since the early fifties. All in all, the four curves show that the age of literature (average and median) has been increasing steadily since the mid-seventies².

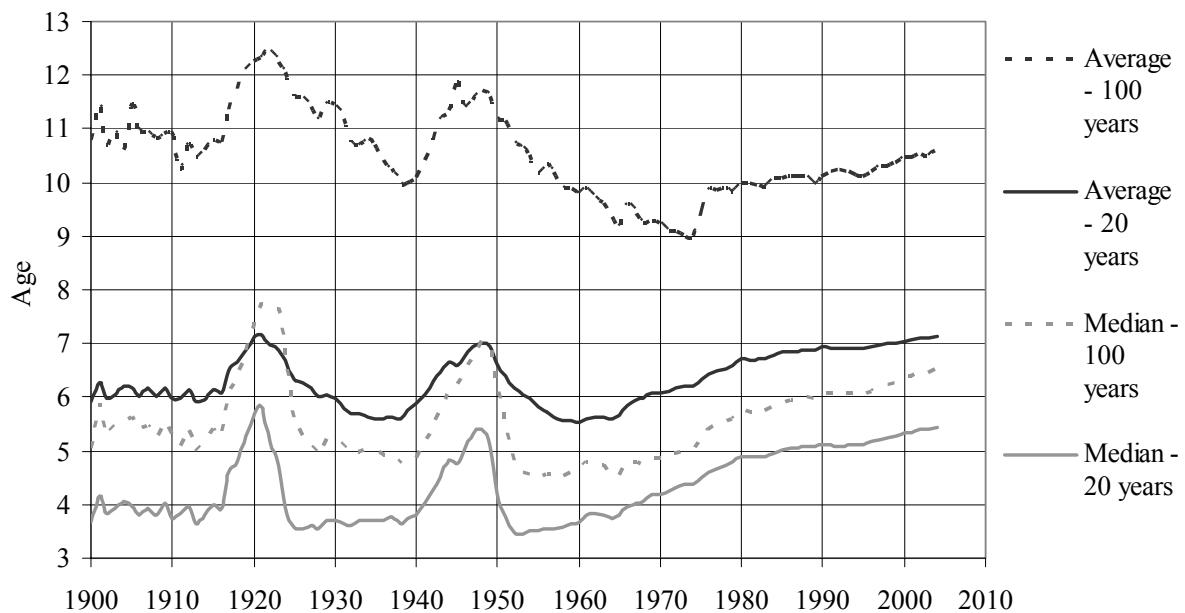


Figure 2. Average and median age of cited literature, 1900–2004

² The huge upward jump observed in the mid-seventies for the average age calculated with a 100-year citation window is caused by the creation of the AHCI by Thomson Scientific.

In order to see whether the phenomena observed is also valid for a fixed journal set, we compared the trend for two journals over the 1946–2005 period. As shown in Figure 3, the tendency towards a greater age of cited literature is also observed in *Science* and *Nature*³. For the sake of clarity, data for the 1900–1945 were removed because the small numbers of articles generated very large fluctuations. While the evolution of both curves is not identical to that of all ISI indexed journals presented in Figure 2, they are quite similar. Though the increase in the median age is less pronounced, one can observe that it has increased by about one year since the beginning of the fifties. The curves on the average age are also consistent with those presented in Figure 2.

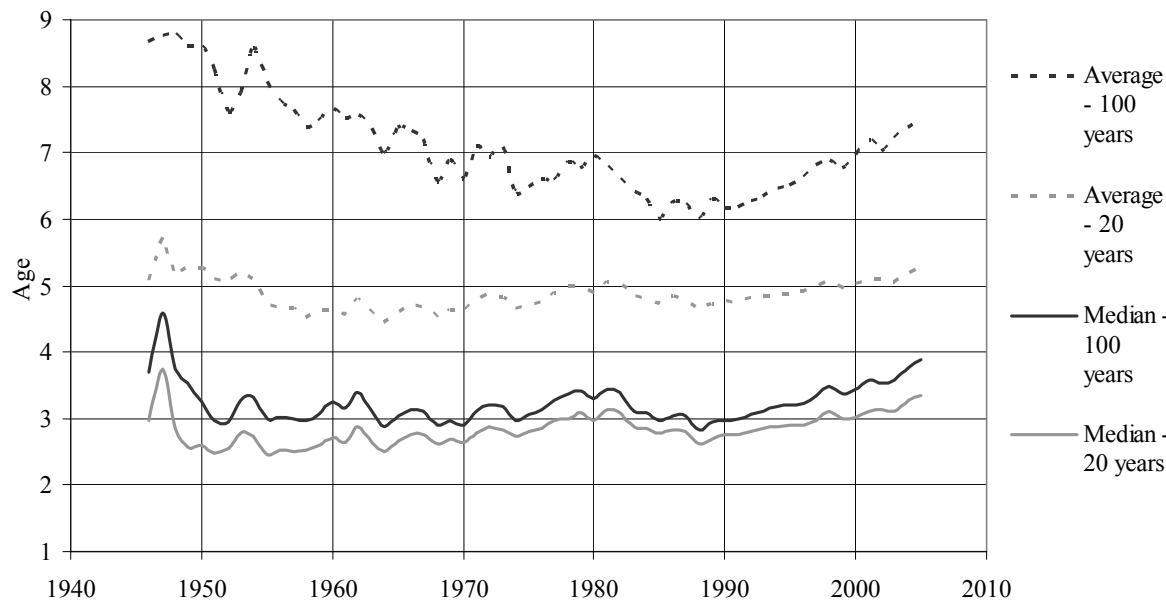


Figure 3. Average and median age of cited literature for *Science* and *Nature*, 1946–2005

Another measure of the aging of literature is the percentage of references, for a given year, to material that is five years old or younger. This measure—the Price Index—was developed by Price (1986) to distinguish fields having fast growth and an intense research front from less research-intense fields. Given the observed rise in the median age, we should expect to see a decline of the Price Index over time. Figure 4 shows that, since the mid-fifties, the share of references made to very recent literature has declined. Indeed, in addition to the huge effect of the two world wars, both the 20-year and the 100-year citation window curves show that the relative importance of recent literature among all cited material has been steadily decreasing, from 63% to 47% with a 20-year citation window, and from 53% to 41% with a 100-year citation window.

Given the fact that the number of references per paper has increased tremendously since the mid-sixties, another way of measuring changes in the aging pattern of cited literature is to analyse its evolution by *citing unit* instead of amongst all cited material. Figure 5 presents the average number of references per paper, broken down for nine discrete age classes. One can see that the increase in the number of cited documents per paper is, by and large, caused by the absolute increase of *mid-aged* or *mature* literature—that is, papers having 3 to 5, 6 to 10, and 11 to 20 years of age. On the other hand, papers aged 1 year have seen their absolute importance decrease, while these aged 2 years stagnate after the end of the Second World War, and it only started increasing (at a much smaller rate than older papers) at the beginning of the eighties. Another point worth mentioning is the fact that documents aged 0—cited documents for which the publication and the cited year are the same—did not increase at all, and even decreased over the period studied. Taken together, these findings are contrary to what is to be expected, if one accepts the widespread idea that growing competition creates a push

³ Note that the upward jump is absent from Figure 3, confirming the role of the AHCI in accounting for it in Figure 2.

towards the use of younger literature and that the Web facilitates access to preprints and newly published research.

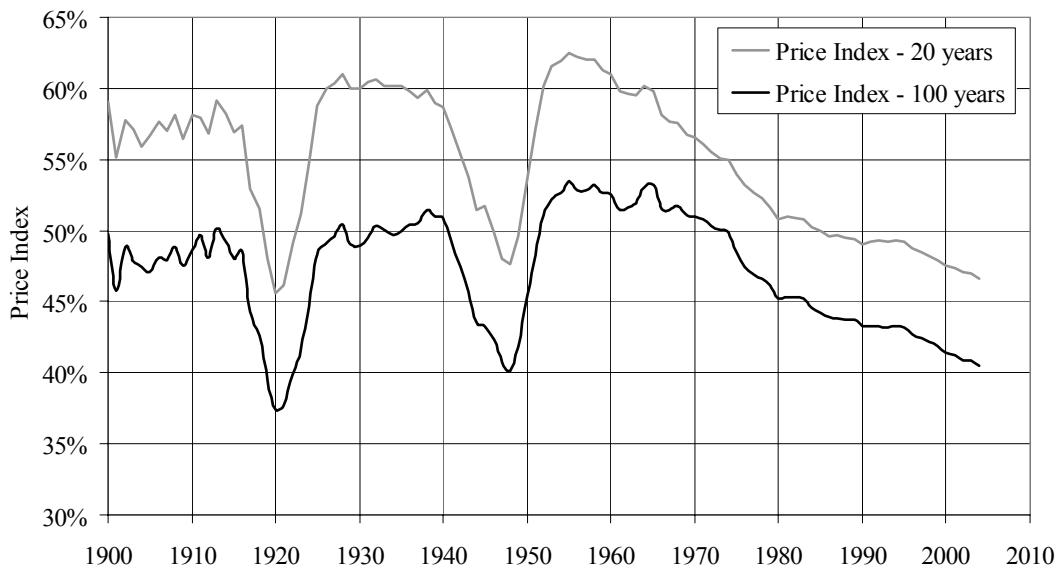


Figure 4. Price index, 1900-2004

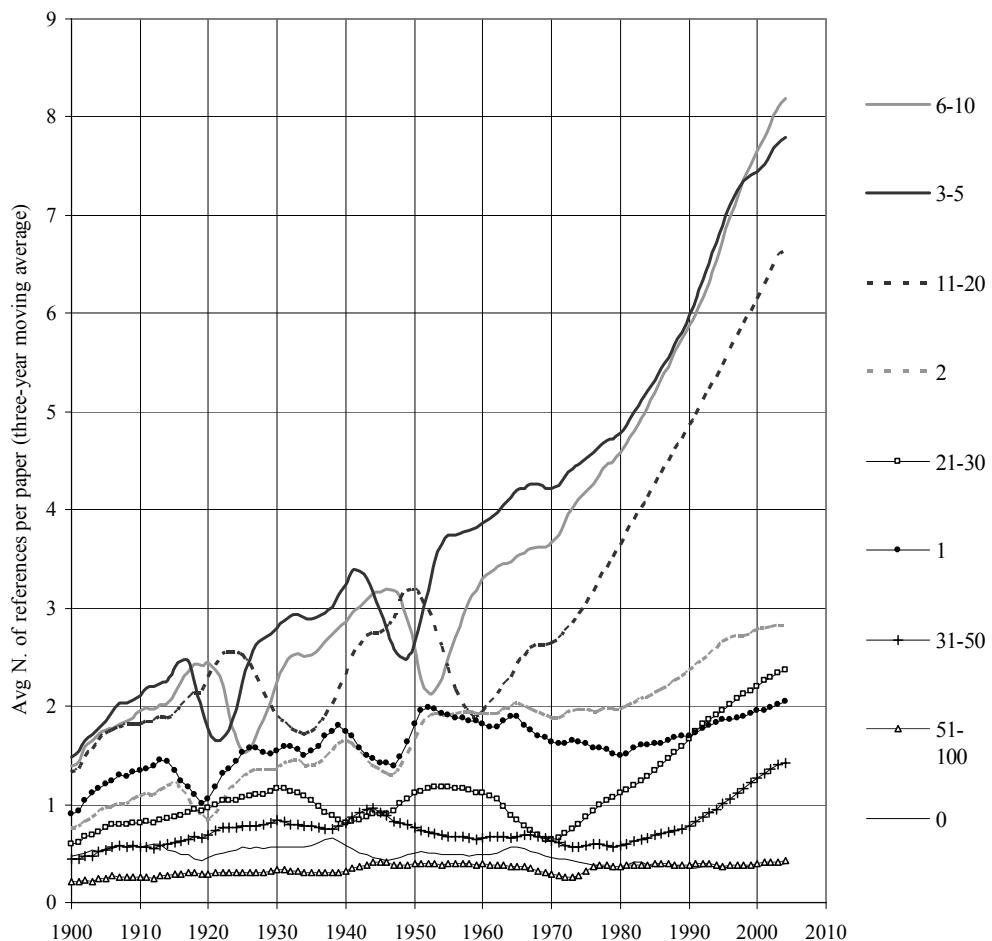


Figure 5. Average number of references per paper, by age of cited document, 1900-2004

Discussion and conclusion

The Figures presented thus far reveal a striking feature of the cited literature. Indeed, in contrast to a widely-held belief, scientific literature does not become obsolete faster nowadays and, actually, quite the opposite is observed. The useful life of scientific publications has been increasing steadily since the mid-seventies.

A first explanation for this rise in the age of cited literature can be inferred from the effects the two world wars had on the age of cited documents. As shown on Figures 1 and 2, because of the lower number of papers published during the wars (Figure 1), a significant increase in the average and median age of the documents cited by these papers can be observed (Figure 2). Given that a small decrease in the number of papers published has had a significant effect in the age of what is cited, a stabilization of the number of published papers will have a similar, albeit less pronounced, effect and increase the average and median age of cited documents.

These findings are consistent with Egghe's (1993: 199) models suggesting that, in synchronous analyses, "the higher the growth rate of the literature is, the faster it becomes obsolete." Conversely, if the rate of growth of the literature slows down, one could therefore expect a lengthening of the life of documents. Even though some countries—such as China—do currently have an exponential growth rate, in most countries, especially in North America and Europe—who account for most papers in the database—the growth has either turned to a fairly low exponential growth rate or even into linear mode, thus suggesting that these systems are in a steady state, if not slowing down (Ziman, 1994). All of this evidence points to the suggestion that the phenomena we are observing are internal characteristics of the scientific system. However, that is not to say that there are no factors at play that would be reflections of the changing behaviour of scientists. After all, there might be, underneath this trend, multiple sociological factors shaping the evolution of such a complex system.

In order to better understand the surprising growth of median age over the last quarter of the 20th century, we have distinguished, using the method developed by Larivière *et al* (2006), citations made to serials from those made to non-serials. As Figure 6 shows, the median age of cited non-serials in the natural sciences and engineering (NSE) and in medical fields (MF) is increasing more rapidly than for serials. However, for the social sciences and humanities (Figure 7), the median age of serials and non-serials is quite similar. This suggest that—at least for the NSE and the MF—part of the rise in the median age observed in Figure 2 is in fact due to the use of older non-serials. We cannot at this point suggest why non-serials show this intriguing pattern, and only further research could provide an explanation.

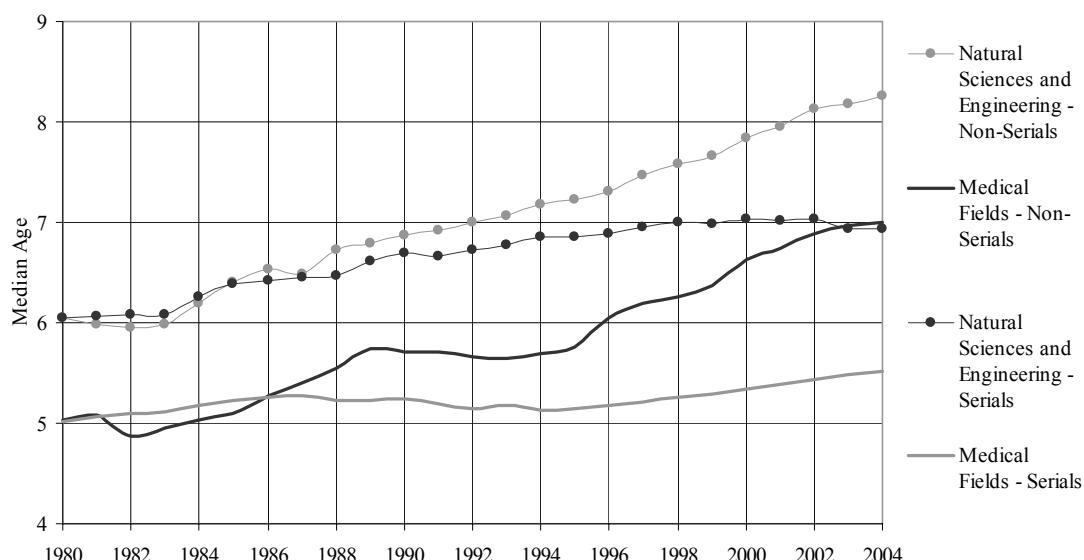


Figure 6. Median age of cited literature, by cited document, natural sciences and engineering and medical fields, 1980-2004 (100 years citation window)

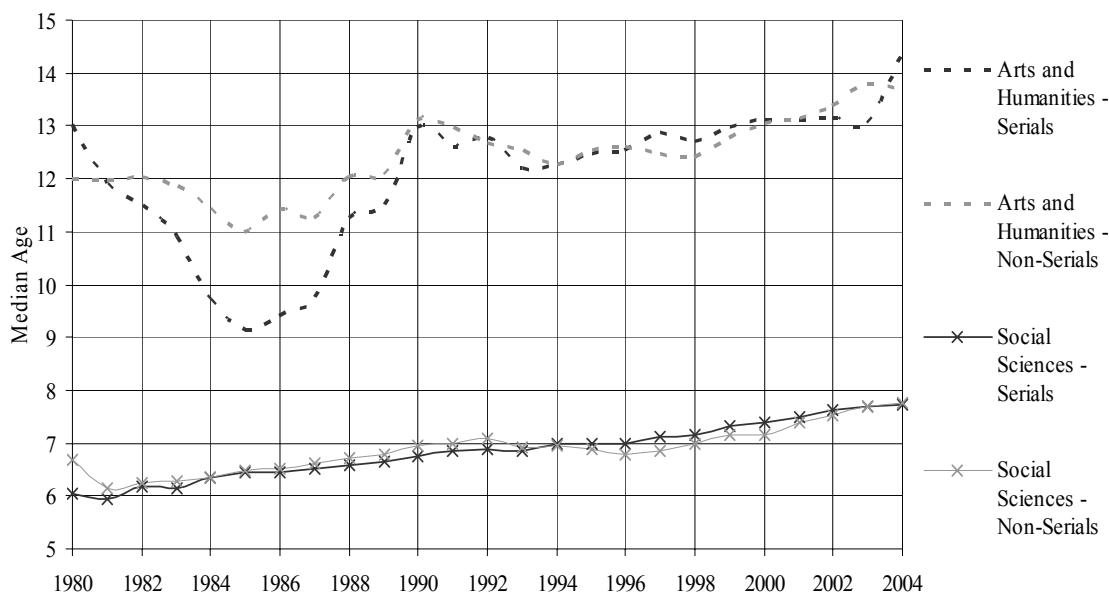


Figure 7. Median age of cited literature, by cited document, social sciences and humanities, 1980-2004 (100 years citation window)

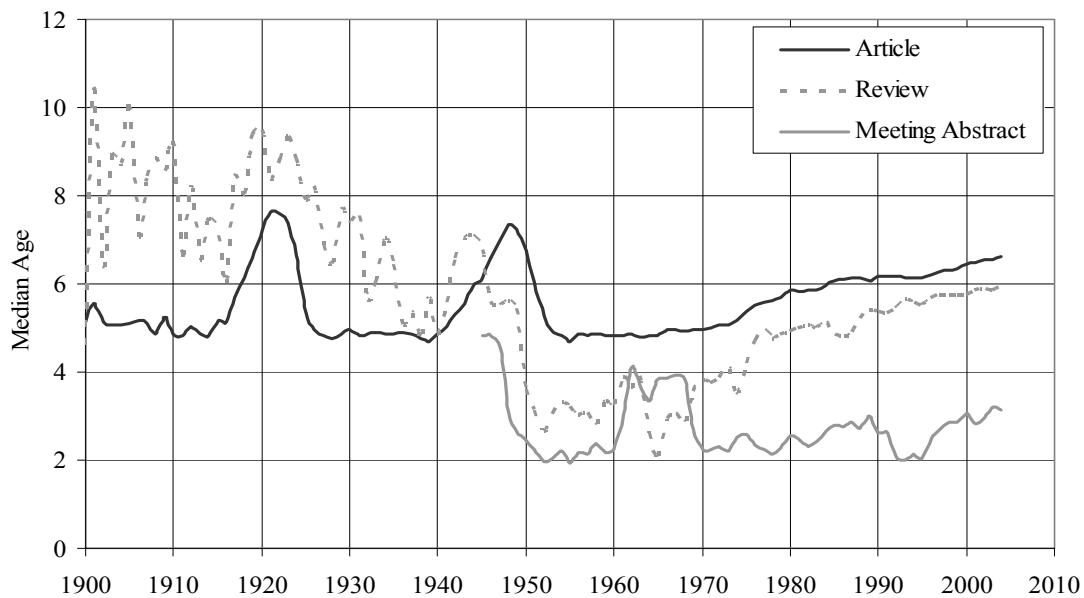


Figure 8. Median age of cited literature, by type of citing document, 1980-2004 (100 years citation window)

Another explanation suggested by a reviewer is the effect of the increasing importance of review articles in Thomson's databases. Indeed, from the sixties to this day, the share of review articles increased from about 1% to 4%. Taking into account that review articles have a higher number of references and may cite older items, their growing importance might affect the median age of cited literature presented in Figure 2. Surprisingly, Figure 8 shows that, since the mid-forties, the median age of cited literature is higher for articles than for review articles or meeting abstracts. Consequently, the presence of review articles does not explain the trend observed, since the effect of an increase of

reviews article would be to decrease the global median age of cited literature. This Figure also shows that, as one could expect, meeting abstract cite literature that is, on average, significantly younger than for articles or reviews.

Other phenomena which could contribute to the longer life of scientific literature are the explosion of online bibliographical tools containing retrospective collections of serials. Such online tools certainly help researchers access increasingly old material, which they then could cite more frequently (see Boyce et al., 2004). Although this might have contributed to the variation in the aging process in the most recent years, Figure 5 clearly shows that citing older material more frequently started as early as 1960 (e.g., citing material that is 11 to 20 years old) and grew steadily afterwards. One cannot deny that the growing availability of computerized search tools since the mid-sixties (Neufeld and Cornog, 1986) and their subsequent wider availability contributed to this change. This presents indirect evidence that online databases of historical archives—such as JSTOR—should be highly encouraged and supported. Moreover, another hypothesis that can be explored, following Luwel and Moed (1997), is the impact that increased publication delay might have on the median and average age of cited literature.

As shown by the Price Index, science as a whole has been less and less intense at the research front. This is another argument pointing at a slower *renewal* of classics and of research fields. After the golden age of science (1946–1975), scientists had solved many of the important bottlenecks they faced, and no major “scientific revolutions” have appeared since. We would, thus, now be in a period of steady-state science (Ziman, 1994). Also, it may be that major contributions were more frequent before 1975 and that scientists increasingly cite material from that last innovative period. Although these explanations are worth further research, we suggest that the principal cause of the increased age of the cited scientific information is a fairly mechanistic response to the phenomenal growth in the quantity of published material after the war and to the current slowing in the growth of science.

References

- Boyce, P., King, D.W., Montgomery, C. & Tenopir, C. (2004). How electronic journals are changing patterns of use, *The Serials Librarian*, 46, 121-141.
- Eggle, L. (1993). On the influence of growth on obsolescence. *Scientometrics*, 27, 195-214.
- Gross, P.L.K. & Gross, E.M. (1927). College libraries and chemical education. *Science*, 66, 385-389.
- Larivière, V., Archambault, É., Gingras, Y. & Vignola-Gagné, É. (2006) The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities, *Journal of the American Society for Information Science and Technology*, 57, 997-1004.
- Line, M.B. (1970) The half-life of periodical literature: apparent and real obsolescence. *Journal of documentation*, 26, 46-54.
- Line, M.B. (1993) Changes in the use of literature with time: obsolescence revisited. *Library Trends*, 41, 665-683.
- Line, M.B., Sandison, A. (1974). "Obsolescence" and changes in the use of literature with time, *Journal of Documentation*, 30, 283-350.
- Luwel, M. & Moed, H.F. (1998) Publication delays in the science field and their relationship with the ageing of scientific literature. *Scientometrics*, 41, 29-40
- Neufeld, M.L. & M. Cornog (1986). Database History: From Dinosaurs to Compact Discs. *Journal of the American society for information science*, 37, 183-190.
- Nicholas, D., Huntington, P., Dobrowolski, T., Rowlands, I., Jamali M., H.R. & Polydoratou, P. (2005). Revisiting obsolescence and journal article decay. *Information Processing & Management*, 41, 1441-1461.
- Odlyzko, A. (2002). The rapid evolution of scholarly communication. *Learned Publishing*, 15, 7-19.
- Price D.J.D. (1963). *Little science, big science*. New York: Columbia University Press.
- Price D.J.D. (1965). Networks of scientific papers. *Science*, 149, 510-515.
- Price D.J.D. (1986). Citation measures of hard science, soft science, technology, and nonscience. In Nelson, C.E. & Pollack, D.K., (Eds.), *Communication among scientists and engineers* (pp. 155-179). New York: Columbia University Press.
- Ziman, John M. (1994). *Prometeus bound: science in a dynamic steady state*. Cambridge: Cambridge University Press.

The Scholarly Database and Its Utility for Scientometrics Research¹

Gavin LaRowe, Sumeet Ambre, John Burgoon, Weimao Ke and Katy Börner

glarowe@indiana.edu, sambre@indiana.edu, jburgoon@indiana.edu, wke@indiana.edu, katy@indiana.edu
Indiana University, School of Library and Information Science, 10th Street & Jordan Avenue, Bloomington, IN
47405 (USA)

Abstract

The Scholarly Database (SDB) at Indiana University aims to serve researchers and practitioners interested in the analysis, modeling, and visualization of large-scale scholarly datasets. This database focuses on supporting large studies of changes in science over time and communicating findings via knowledge-domain visualizations. The database currently provides access to around 18 million publications, patents, and grants, ten percent of which contain full-text abstracts. Except for some datasets with restricted access conditions, the data can be retrieved in raw or pre-processed format using either a web-based or relational database client. This paper motivates the need for the database from bibliometric and scientometric perspectives (Cronin & Atkins, 2000; White & McCain, 1989). It explains the database design, setup, and interfaces as well as the temporal, geographical, and topic coverage of datasets currently served. Planned work and the potential for this database to become a global test bed for information science research are discussed.

Keywords

research database; geospatial coverage; data integration; mapping science

Introduction

Digitized scholarly datasets and sufficient computing power to integrate, analyze, and model these datasets make it possible to study the structure and evolution of science on a global scale (Börner, Chen, & Boyack, 2003; Boyack, Klavans, & Börner, 2005; Shiffrin & Börner, 2004). Results can be communicated via tables, graphs, geographic and topic maps. Frontiers emerging across different sciences can be discovered and tracked. Different funding models can be simulated and compared. School children can start to understand the symbiotic relationships among different areas of science.

The study of science on a global scale requires access to high quality, high coverage data and major cyberinfrastructure (Atkins et al., 2003) to process such data. Many studies require the retrieval and integration of data from different sources with differing data types. For example, input-output studies require input data, e.g., funding amounts and number of new graduates, and output data, e.g., the number of publications, received citations, awards, and policy changes. Unfortunately, the identification and inter-linkage of unique authors, investigators and inventors is non-trivial.

Contrary to other scientific disciplines where data is freely and widely shared, there are very few bibliometric or scientometric test datasets available. This makes it very time consuming (e.g., data download, cleaning and inter-linkage) or impossible (if datasets require access permissions) to replicate studies or reproduce results. Fortunately, some services and institutions, such as PubMed, CiteSeer, arXiv, and the United States Patent Office, provide free data dumps of their holdings under certain conditions. However, most bibliometric and scientometric scholars are not trained in parsing millions of XML-encoded records and very few have expertise in the setup and maintenance of multi-terabyte databases.

¹ This work was supported by the National Science Foundation under Grant No. IIS-0238261, IIS-0513650, IIS-0534909, and CHE-0524661 and a James S. McDonnell Foundation grant in the area Studying Complex Systems entitled "Modeling the Structure and Evolution of Scholarly Knowledge". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. We thank Kevin W. Boyack and Russell Duhon for insightful comments on an earlier version of this paper.

The Scholarly Database, online at <https://sdb.slis.indiana.edu/>, aims to improve the quality and reduce the costs of bibliometric and scientometric research and practice by providing easy access to high quality, comprehensive scholarly datasets.

Database Architecture and Access

The Scholarly Database comprises production, development, and research systems. The *production system* utilizes two Sun V490 servers with sixteen gigabytes of memory each serving two redundant, mirrored clusters of the Scholarly Database running PostgreSQL v8.1.4 under Solaris 10. Failover, load balancing, and replication are provided via a planned intermediary system service between each of these clusters. In addition, each instance of the database is isolated within its own zone under Solaris 10, providing increased security and failover for the system at-large. The *development system* is used for developing future production versions of various datasets. After post-processing, incorporation, and testing, new datasets receive their own data store and various schema templates are incorporated. When approved, these datasets are then pushed to one of the production clusters. The *research system* runs on a Sun V480 with thirty two gigabytes of memory providing a sandbox for researchers to develop and refine personal or proprietary datasets. Aside from being a completely isolated system, this cluster provides users with sufficient memory and disk storage for large-scale cyberinfrastructure projects involving multi-gigabyte and, in the future, terabyte-scale data and application requirements. All three systems are hosted at the Cyberinfrastructure for Network Science Center at Indiana University.

The general system architecture of the Scholarly Database has three major parts: The *Data Space* stores raw data, pre-processed data, metadata, and any other raw data artefacts. *Data Services* support data harvesting, data mining, pre- and post-processing, statistical analysis, and in the near future natural language processing, multi-agent data simulations, and information visualization services. Aside from backup and storage, *Data Provenance* is provided via the use of metadata associated with the raw data, internal database artefacts (e.g., schemas, tables, views, etc.), and user artefacts such as queries or views.

The database schema is too extensive to describe or depict here in anything but an abstract overview. The current implementation utilizes three schemas: public, base, and auxiliary. The *public schema* describes all post-processed raw data that has been loaded into the database. This data and associated metadata are rarely modified, except when new updates are received from a data provider. The *base schema* contains all foundational views found in the web interface used for search, display, and download of data. Most views are virtual, but some materialized tables are used for extremely large datasets. The *auxiliary schema* provides a space where schemas, tables, views, and functions created by the users of the system can be stored for re-use. Its ancillary purpose is to provide an area where one-time or proprietary non-public components (e.g., views or functions) can be explored for a dataset by an authorized user.

Access to the database is available via a web front-end at <https://sdb.slis.indiana.edu/>, a pgAdmin PostgreSQL administration and management tool, and a psql PostgreSQL interactive terminal. The front-end interface allows external users to search through all of the articles and documents in the database via author, title, or keyword for a specified year range. Results are returned showing generic fields such as journal name, title, author name, and date of publication. Each record can be further expanded to show fields specific to a given dataset. Selected datasets can be downloaded. Future services will support the extraction and download of networks such as co-author and paper-citation networks.

Dataset Acquisition, Processing and Coverage

The Scholarly Database is unique in that it provides access to diverse publication datasets and to patents and grant award datasets. Datasets are acquired from a wide variety of sources. Some are one time acquisitions. Others are updated on a continuous basis. Several have access restrictions. Table 1 provides an overview.

Medline publications provided by the National Library of Medicine (<http://www.nlm.nih.gov>), consists of two types of data: baseline files that are distributed at the end of each year and include all PubMed records that have been digitally encoded in XML for Medline; and newly added data for that particular year which is subsequently updated in future baseline releases. It is provided in XML format with a custom DTD. Update files are provided regularly.

Table 1. Datasets and their properties (* future feature).

Dataset	# Records	Years Covered	Updated	Restricted Access
<i>Medline</i>	13,149,741	1965-2005	Yes	
<i>PhysRev</i>	398,005	1893-2006		Yes
<i>PNAS</i>	16,167	1997-2002		Yes
<i>JCR</i>	59,078	1974, 1979, 1984, 1989, 1994-2004		Yes
<i>USPTO</i>	3,179,930	1976-2004	Yes*	
<i>NSF</i>	174,835	1985-2003	Yes*	
<i>NIH</i>	1,043,804	1972-2002	Yes*	
Total	18,021,560			

Physical Review papers provided by the American Physical Society (<http://aps.org>) come in 398,005 XML-encoded article files covering nine journals (A, B, C, D, E, PR, PRL, RMP, PRST, and AB) over a one hundred and ten-year time span: 1893-2006. A single DTD exists for the entire collection. It encompasses all changes made throughout the history of the digital encoding of these files that were previously available in SGML format. It is a proprietary dataset that cannot be shared or used for commercial purposes.

Proceedings of the National Academy of Sciences provided by PNAS (<http://www.pnas.org>), comprise full text documents covering the years 1997-2002 (148 issues containing some 93,000 journal pages). The dataset is also available in Microsoft Access 97 format. It was provided by PNAS for the Arthur M. Sackler Colloquium, Mapping Knowledge Domains, held May 9-11, 2003. It is available for research and educational purposes to anybody registered for that Colloquium and who signed the copyright form. It cannot be redistributed without prior permission from PNAS. It cannot be used for commercial purposes.

Journal Citation Report (JCR-Science Edition) dataset by ISI Thomson Scientific (<http://www.isinet.com>) comprises two datasets: (1) covers the years 1994-2004; and (2) contains cited and citing pairs records for 1974, 1979, 1984 and 1989 from the Science Citation Index – Expanded (SCI-E). Both are restricted use for the purpose of an NSF grant. This data cannot be used, distributed, or otherwise shared without prior written permission from Thomson Scientific.

Patents by the United States Patent and Trademark Office (USPTO) (<http://www.uspto.gov>) come as XML-encoded dumps downloadable from the USPTO website (<ftp://ftp.uspto.gov/pub/patdata/>). This is a publicly accessible dataset of about three million records organized into ca. 160,000 patent classes.

NSF Grants awarded by the National Science Foundation (NSF) (<http://www.nsf.gov>) support research and education in science and engineering to more than two thousand colleges, universities, and other research and education institutions in all parts of the United States through grants, contracts, and cooperative agreements. It is composed of raw text files as distributed by the NSF for the years listed above. It is a publicly accessible dataset.

NIH Grants data from the National Institutes of Health (NIH) (<http://www.nih.gov>) is composed of CRISP and Awards data downloaded from the main NIH web site and the CRISP on-line search engine <http://crisp.cit.nih.gov/> for the years listed above. The CRISP data includes information regarding extramural projects, grants, contracts, and so on associated with projects and research

supported by the National Institutes of Health. NIH award data is composed of principal investigator and institution NIH grant award amounts concerning projects found in the CRISP data for the years listed above.

Detailed information on these datasets, and their quality and coverage, as well as available data fields, is available online at <https://nwb.slis.indiana.edu/community/> (select ‘Datasets’). New datasets are added on a continuous basis.

Temporal Coverage. As can be seen in Fig. 1 (left), the Medline dataset has the most records per year, with about 500,000 new records each year. There are about 200,000 new USPTO patents each year, 10,000 new NSF awards, and 50,000 new NIH awards per year. The number of unique authors per year is shown in Fig. 1 (right). A concatenation of first author name and last author name was employed to identify unique authors. There is more than one ‘John Smith’ in the Medline dataset, and we know that some authors change names, but the graph provides a rough estimate of how many unique authors contribute to the growth of each dataset and the increase in the number of authors over time.

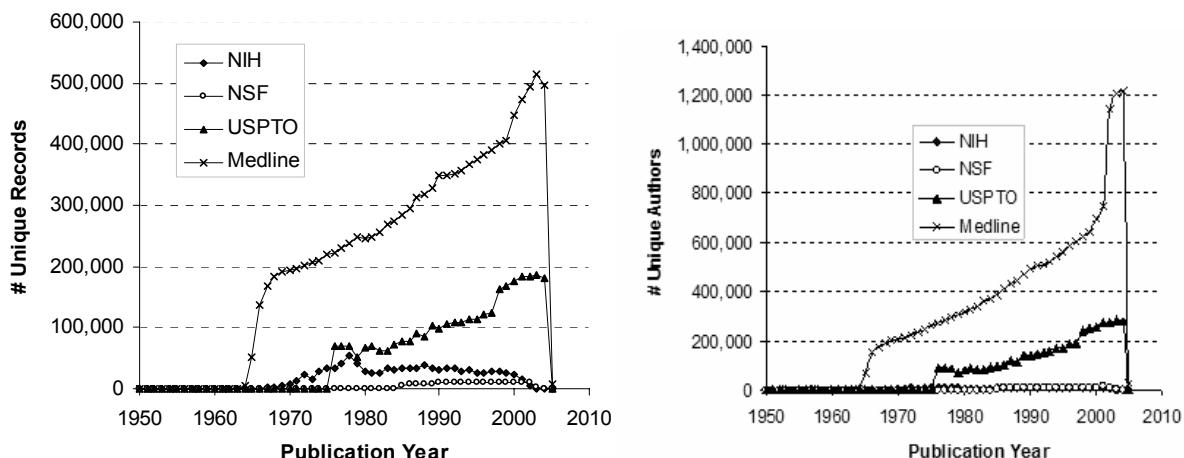


Figure 1. Number of records and unique authors per year for different datasets for 1950 to 2005.

Geographical Coverage. The geographical coverage of datasets can be examined by geolocating papers, patents, and awards based on the affiliation of their authors, inventors, and awardees. It is frequently not clear how best to divide the contributions of single authors across the team. In some datasets, affiliation data is only available for the first author. Therefore, we attribute the location of the paper, patent, or award to the first author, inventor, or awardee. For each first author, inventor, or awardee we retrieved either a zip code or a city-state pair. Zip codes were matched against zip code data provided by Novak Banda at <http://www.populardata.com> to derive latitude and longitude coordinates. When a zip code was not available, all zip codes for the city-state pair were retrieved and a geospatially central zip code was assigned and geolocated. As we did not have access to world wide geocoding services, this analysis is restricted to US. Due to page limitations, we prototypically plot the coverage of only two datasets: Medline and NIH, in Fig. 2.

Note that only 1,420,816 of the 13,149,741 Medline publications had an US affiliation. Out of those, only 1,036,865 had zip codes. There were 13,188 unique zip codes and 10,450 of those could be geolocated. As for NIH, 971,754 main awardees had city-state pairs that were used to identify 1,986 unique geolocations. Fig. 2 shows that major funding and publication patterns are concentrated in urban areas where research centers, such as universities, research labs, hospitals, etc., are more prone to exist.

Topic Coverage. Interested to see the topic coverage of different datasets, we tried to identify the number of journals that the different publication databases cover. In particular, we ran a query that matched Medline journals and JCR journals based on ISSN numbers. However, there were only 3,547

matches. This is partially due to the fact that journals can have multiple ISSN numbers and Medline and Thomson Scientific data might not use the same ones. Matching based on journal names is even more difficult, because abbreviations and omissions differ among the databases under consideration. Medline covers 6,991 unique journals and JCR has 9,227 unique journals from 1994 to 2004.

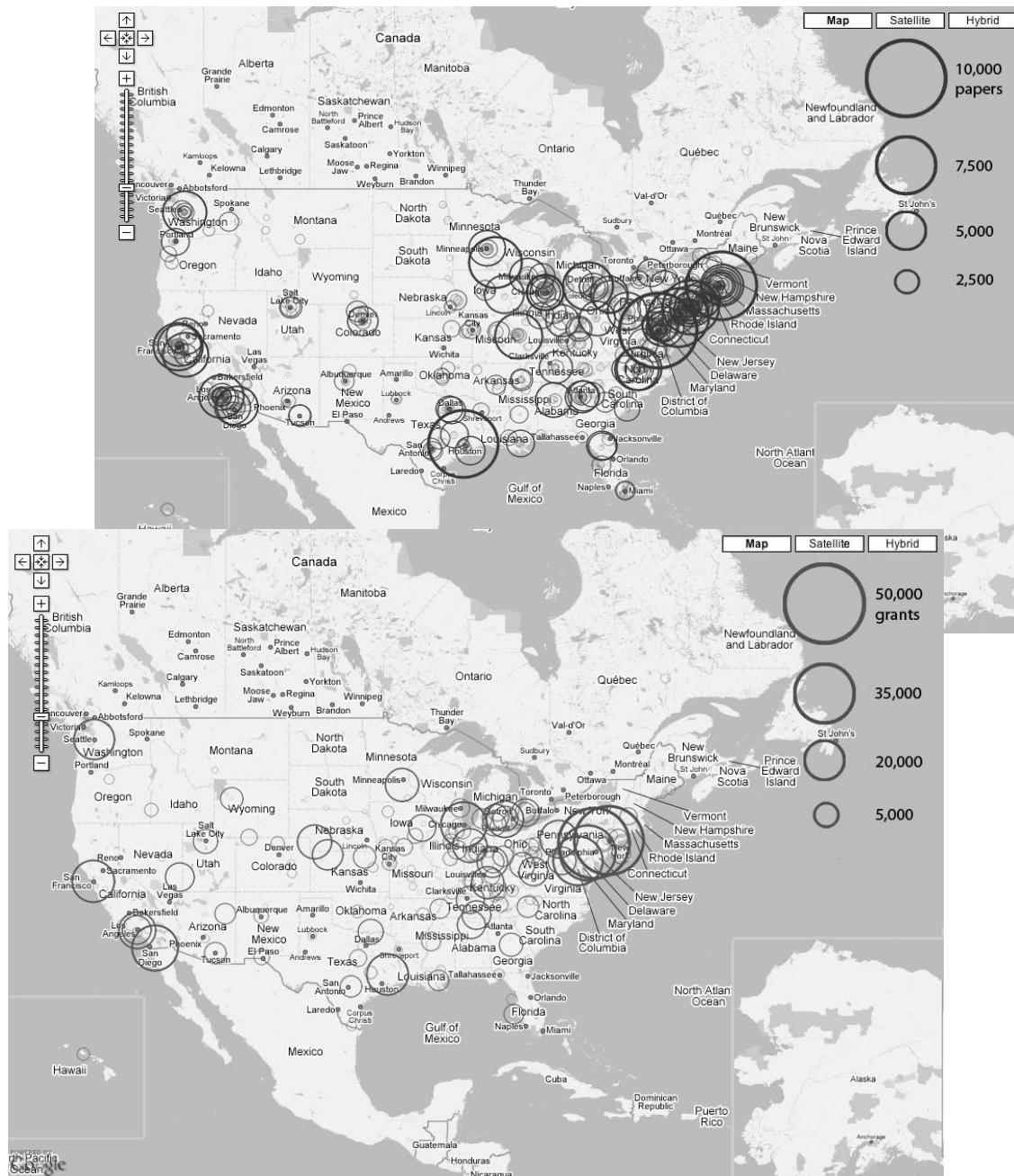


Figure 2. Number of Medline publications and NIH awards by geo-location (U.S. only)

Discussion and Future Work

The Scholarly Database addresses a central need for the large-scale study of science: access to high quality, centralized, comprehensive scholarly datasets. The value and quality of this database will depend on its adoption by the community. The more scholars and practitioners use it, the more likely it is that missing records or links will be discovered, important datasets will be integrated, the best (author/institution/country/geo-code) unification algorithms can be applied, and research studies are conducted, replicated, and verified.

The rate of adoption will greatly depend on the utility and usability of the SDB. Hence, future work aims to make the SDB easier to use and easier to extend by adding new datasets and services. Concurrent to the work being done on the Open Archives Initiative (OAI) (Bekaert & Sompel, 2006), we are working on an internal metadata framework that will encompass common relations between various scholarly datasets. This metadata framework will ease schema matching between datasets. Our solution will incorporate, where possible, any pre-existing metadata descriptions from the OAI and other standards. In order to provide reliable access to non-proprietary data, the Scholarly Database has been designed for easy mirroring in geographically distinct locations. All software used is open source and the database setup is documented in detail to ease installation.

We expect to serve ten major datasets by summer 2007 – about 20 million records. Plus, the open access parts of the database will be made available for information science research in database design, data integration, data provenance, data analysis, data mining, data visualization, and interface design. This will require close collaboration with many researchers, practitioners, and dataset providers. In return, we expect to gain access to more sophisticated data harvesting, preservation, integration, analysis, and management algorithms that are urgently needed to improve data access and management tools for scholars, practitioners, policy makers, and society at large.

References

- Atkins, D. E., Drogemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., Messian, P., Ostriker, J. P., & Wright, M. H. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Arlington, VA: National Science Foundation.
- Bekaert, J., & Sompel, H. V. d. (2006). Augmenting Interoperability Across Scholarly Repositories, *Meeting sponsored and supported by Microsoft, the Andrew W. Mellon Foundation, the Coalition for Networked Information, the Digital Library Federation, and the Joint Information Systems Committee*. New York, NY. Retrieved from <http://msc.mellon.org/Meetings/Interop/FinalReport> on 2/15/2007.
- Börner, K., Chen, C., & Boyack, K. (2003). Visualizing Knowledge Domains. In B. Cronin (Ed.), *Annual Review of Information Science & Technology* (Vol. 37, pp. 179-255). Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.
- Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the Backbone of Science. *Scientometrics*, 64(3), 351-374.
- Cronin, B., & Atkins, H. B. E. (2000). *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*: American Society for Information Science & Technology.
- Shiffrin, R. M., & Börner, K. (Eds.). (2004). *Mapping Knowledge Domains* (Vol. 101 (Suppl. 1)): PNAS.
- White, H. D., & McCain, K. W. (1989). Bibliometrics. In M. E. Williams (Ed.), *Annual Review of Information Science & Technology* (Vol. 24, pp. 119-186). Amsterdam, Netherlands: Elsevier Science Publishers.

Interdisciplinarity in Environmental Technology Applications – Examining Knowledge Interaction between Physics and Chemistry Research Teams¹

Katarina Larsen*

*larsen@infra.kth.se & klarsen@stanford.edu
KTH – The Royal Institute of Technology, 100 44 Stockholm, (Sweden)

Abstract

This paper examines interdisciplinarity in science-based environmental technology applications, including nanoscience applications in solar cell technology and sensor technology for pollution monitoring (of for example vehicle exhaust gas). Data and methods include analysis of co-authorship links, citation data combined with content analysis and interviews with researchers active in the field. Interdisciplinarity of science and technology areas has been analysed at the level of researcher affiliation, journal type and keywords. In addition to these measures of interdisciplinarity, this study acknowledges the importance to recognise the character of links. This includes both their content and the context that generates interdisciplinary links between research teams through citations or co-authorships interaction. To illustrate this, the knowledge networks of two research teams from the same university are examined and visualised using bibliographic coupling and co-authorship networks. The findings show that although there are no direct co-authorship links between the two research teams in the dataset analysed, links were identified through bibliographic coupling. In examining the context generating a shared reference between the research teams, one of the teams show citation patterns described here as “interdisciplinary outlook” compared to the “intradisciplinary magnifier” pattern of the other research team.

Keywords

science-based technology; interdisciplinary; nanotechnology; citation map; bibliographic coupling; coauthorship

Background and purpose

Interdisciplinarity of science and technology areas has previously been analysed at different levels, including researcher affiliation, journal type and keywords (Schummer, 2004; Rinia et al., 2001 and Tijssen, 1991; Meyer and Persson 1998). In addition to these measures of interdisciplinarity, this study acknowledges the importance to recognise the character of links. This includes both their content and the context that generates interdisciplinary links between research teams (through citations or co-authorships). This is also called for, given that contemporary science policy in a European context and in many nations stresses the importance of increased cross-disciplinary interaction to advance innovation and application of science in society, but often without further clarifying the scope of the interdisciplinary collaboration in different scientific fields.

This distinction is also called for, given that contemporary science policy in a European context and in many nations stresses the importance of increased cross-disciplinary interaction to advance innovation and application of science in society, but often without further clarifying different scope of interdisciplinarity in different phases of innovation processes. Studies of interdisciplinary knowledge distinguish between multidisciplinary (additive) and interdisciplinary (integrated) approaches to knowledge production (Gilbert, 1998). Although the strive for interdisciplinary science is a strong trend in contemporary science policy, there are also studies that recognise associated caveats since “many still feel the tension between the scientific promise of the interdisciplinary path and the academic prospect of the tenure track.” (Rhoten and Parker, 2004).

The specific fields analysed includes sensor technology for pollution monitoring (used for, among other things, vehicle exhaust gas) and solar cell technology. These areas of application were selected since the application of nanotechnology in environmentally sound technology is of particular interest to get a better understanding of knowledge interaction between different pockets of scientific knowledge. Previous studies have focused on of communities of nanoscience and nanotechnology

¹. This work was supported by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) and the Sweden-America Foundation.

(Meyer, 2000a-c) and also recognised the limited interaction between different nanoscience communities (Schummer, 2004). While much of previous efforts have focused on analysis at the level of interaction between communities or field evolution, there is a need for further studies of knowledge interaction within and between specific nanotechnology applications. This study of knowledge networks and citation maps of two different research groups also acknowledge recent work analysing the nano-related journals included in the chemistry cluster, as compared to other nano-related journals included in the physics cluster (Leydesdorff and Zhou, 2007).

The purpose of this study is to examine knowledge interaction between two environmental nanotechnology applications, also including analysis of knowledge interaction within each area of application. The wider rationale for undertaking this type of study is twofold. First, it raises questions about shared knowledge bases between scientific fields. This, in turn, can be used to improve our empirical understanding of interdisciplinary science. Secondly, the knowledge networks that are generated in the study are used to visualise and discuss with researchers about the extent of collaboration that this type of data can capture and the content of the links. By this it is meant that in addition to enquiring about the preconditions for collaborations with other researchers abroad and nationally, the study also takes an interest in if collaboration had to do with measurement techniques, shared equipment or skills etc. This motivates using a combination of quantitative bibliometric data and a case study approach including interviews with researchers to examine the nature of knowledge networks and content of the links between different research teams.

The approach used to analyse this further was to study two university departments that have demonstrated the use of nanotechnology in their research by application of nanostructured materials using nanocrystalline Titanium dioxide (TiO_2) and associated measurement techniques for desired properties of their technical devices. The two departments are within the same university (but located in two different parts of the city) and are working with sensor technology (physics department) and solar cell technology (chemistry department).

From a policy perspective, environmental applications of nanoscience and technology are at the heart of policy studies raising questions about impact and potential of new applications of nanotechnology alongside expectations in other areas of applications such as in biotechnology (Royal Society, 2005). A multitude of applications and expectations on potential applications for nano-scale science and technology are provided in policy studies reviewing social and economic impact of nanotechnology (ESRC, 2003; Arnall, 2003).

Data and methods

For examining knowledge interaction in different phases of innovation processes methods used include analysis of co-authorship links, citation data combined with content analysis of publications and patents. The analysis of co-authorship links and citation pattern were predominantly used for examining interdisciplinary patterns between university departments in different disciplines and to explore citation patterns of publication output and patent-to-publication citations. The main focus in this paper is on analysis of the publication data. In the next phase of the study, following the current paper, this is complemented by analysis of co-inventor patterns and non-patent references (NPR) for the fields analysed. The publication data used for the areas of environmental technology applications were retrieved from the Science citation index. In total 33 papers were included in the citation analysis covering 24 papers from department of chemistry and 9 papers from department of physics. The program BibCoul was used for visualization of the bibliographic coupling among authors in the ISI-set in terms of their shared references (Leydesdorff, 2006).

The links between advances in science in the area of chemistry and physics are examined using co-authorship and citation data. The citation data was analysed in order to reveal shared references among the researchers working at the two research departments. Hence, papers from this university were selected (using author affiliation) and the search was further narrowed to not include scientific publications from many other research departments than the two studied here. This was made by only including the papers that were authored (or co-authored) by researchers from the university that shared

a citation to a paper on nanostructured material application in solar cells (Oregan and Gratzel 1991). This procedure of retrieving data was made to ensure that the authors from the two university departments have (at least one) shared reference to the same source of knowledge, rather than analysing two fields that are not remotely related to each other.

The method of bibliographic coupling can be used to create maps or visualisations of knowledge networks. This is described as different ways by which articles can be connected referring to methods such as bibliographic coupling (BC) and co-citation (CC) and Longitudinal coupling (LC) as methods for visualizing science by citation mapping (Small, 1999, 802): “Considering the publication years of articles, there are three ways two articles can be connected by taking two steps on a citation network: 1) Bibliographic coupling (BC), which connects papers by one step back then one step forward; 2) co-citation (CC), which takes one step forward then one step back; and 3) a third form which connects older and younger papers by taking two steps in the same direction, either forward or backward. This third form has been called longitudinal coupling (LC)”. In this paper bibliographic coupling are used to visualise knowledge networks between researchers in the two research departments, also combined with co-authorship data. It is argued that co-authored papers indicate substantive research relationships though which tacit knowledge can be shared (Hicks and Katz, 1997). However all collaboration does not lead to co-authored papers and there are other outputs of collaboration than co-authored papers (Melin and Persson, 1996, p. 365). Given these previous studies, the current paper also builds on interviews with researchers in the field to get a better understanding of the content of co-authorship links and clearer picture of technical skills and complementarities in competences that spurs collaboration across departments and between organisations. Selection of interviewees was made among the research leaders and productive researchers to get a better understanding of content of links in repetitive ties between organisations and disciplines that co-authorship and citation data reveal.

Results

Having a shared source of knowledge could imply that there are more shared references between the two research teams both working with nanostructured Titanium dioxide. The publications analysed all share (at least) one reference to the Oregan and Gratzel paper (from 1991). The results show however that other shared references are scarce (when removing the links in Figure 1-2 that arise due to the shared references made to the Oregan and Gratzel paper). The two separate knowledge networks of the two university departments are shown in Figure 3. The network of the chemistry department involves a greater number of researchers at the same level co-occurrence threshold, which can be explained by the greater number of publications involving researchers from chemistry department in the dataset analysed, compared to publications from the physics department. Another remark when comparing the in the two networks in Figure 4 is that the nodes in the physics network, are connected to two or more other nodes. This is compared to the chemistry network where several nodes are only linked to one other node. The interpretation of this here is that the co-authors that constitute the core of the physics network share a strong knowledge base. This is sometimes (but not always) founded in references made in co-authored papers to other co-authored publications. This pattern can also be seen in the chemistry network, although not as protruding for the chemistry network in Figure 3.

Another level of aggregation is analysed in the co-authorship networks between different university departments, shown in Figure 4. When comparing the co-authorship networks of physics and chemistry, it becomes clear that the chemistry network has collaboration with mainly research divisions abroad. This is contrasted with the physics network connected mainly by national research divisions. Explanations to this were given in interviews with the research teams, pointing to that the collaborators to the chemistry department were mainly found abroad since it is a highly specialised field. The national collaboration in case of the physics department working with advances in sensor technology, on the other hand, relies on research infrastructure that is established in a network of centres for material science located at different university departments. Hence, the double affiliation address for publications (referring to the material sciences institute being located at the department as well as the university department itself) is frequently listed in the science citation index. In addition to this, it is common that researchers in the physics field have two (or more) affiliations at different university departments. This explains some of the connections in Figure 4.

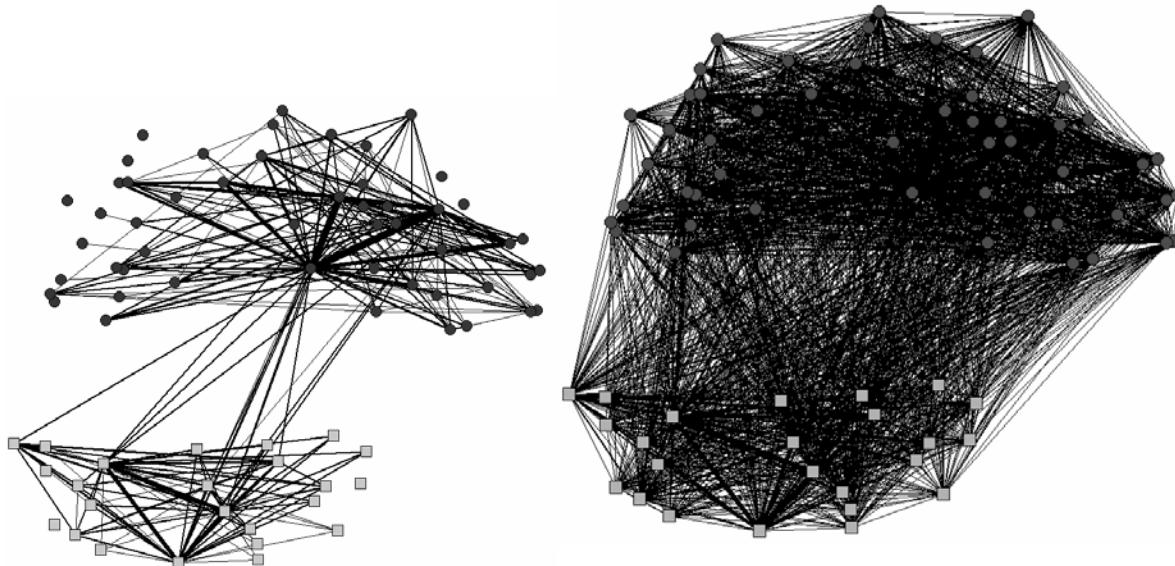


Figure 1. (left) Citation network analysed using bibliographic coupling. (Node attributes: circles in upper part of network represent authors at chemistry department (and their co-authors) and squares represent authors at physics department (and their co-authors). Co-occurrence threshold=1)

Figure 2. (right) Citation network analysed using bibliographic coupling. (Node attributes: same as in Figure 1. Co-occurrence threshold=50.)

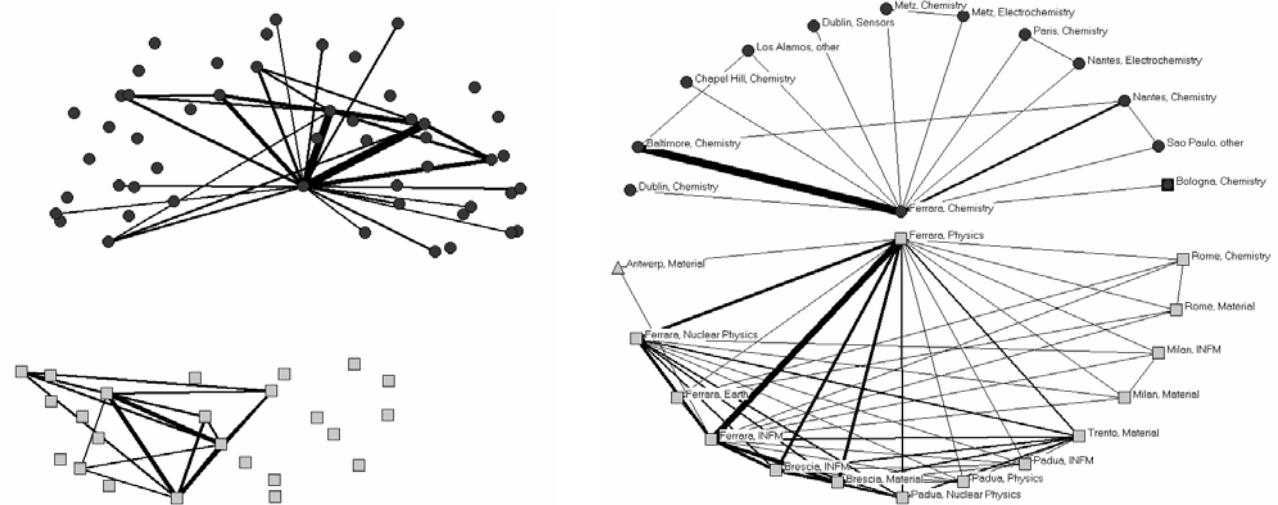


Figure 3. (left) Citation network analysed using bibliographic coupling. (Node attributes: circles in upper part of network represent authors at chemistry department (and their co-authors) and squares represent authors at physics department (and their co-authors). Co-occurrence threshold=200.)

Figure 4. (right) National and international co-authorship networks showing collaboration between different university departments. (Node attributes represent the department type distinguishing between co-authors of physics department and co-authors of chemistry department. Circles represent university departments that have collaboration with the chemistry department, predominantly international collaboration with few exceptions (here circle on a black square to the right in the upper network). Squares represent university departments (predominantly within the country) that have collaboration with the physics department, with few exceptions (here triangle to the left in the lower network).

The results, based on content analysis of publications combined with interviews with researchers and network analysis, show that the links made to the shared reference Oregan and Gratzel (from 1991 in Nature) are grounded in different types of references. In the physics research team, the citations to the seminal paper by have a character of an “interdisciplinary outlook” relating to different applications of Titanium dioxide (TiO₂) listed in citations 1-9 below. In this example, citation number 3 is made to the Nature 1991 paper. This recent publication by the Physics research teams working with material sciences and applications in gas sensors also outlines a number of other areas of application in addition to nanostructured solar cells (Guidi et al. 2003, p.120).

“Titania films possess an immense range of applications, e.g., in the field of optics, (citation 1) electrical insulation, (citation 2) photovoltaic solar cells, (citation 3,4) electrochromic displays, (citation 5) antibacterial coatings, (citation 6) photocatalitic reactors, (citation 7) high-performing anodes in ion batteries, (citation 8) and gas sensing. (citation 9) In the nanometric domain, titania revealed unexpected properties because of the increased fraction of atoms located at the surface or at the grain boundaries. The production of titania films featuring nanometric grain sizes opens up a new field of research, and currently, some applications have greatly benefited from a nanostructured phase for TiO₂. (Citation 10,11)”

Following the outline of the range of applications (citation 1-9 above) it is stated that “titania films featuring nanometric grain sizes opens up a new field of research” with reference to citations 10 and 11 published in Materials science journals (subject category Materials science, Multidisciplinary). In publications authored by the chemistry department, the citation made to the Nature 1991 paper, also citation number 3 here, is made when referring to previous advances in the use of TiO₂ but focusing on the application in photoelectrochemical cells, as shown in example below (Moss et al. 2004, p. 1784).

“Much interest has been focused on photoelectrochemical cells utilizing dye-sensitized nanocrystalline TiO₂ electrodes since early reports of high photocurrent efficiencies. These cells consist of thin films of TiO₂ derivatized by adsorbed chromophores such as [Ru(dcb)₃]²⁺ (dcb is 2,2'-bipyridine-4,4'-dicarboxylic acid) in a thin-layer arrangement of the TiO₂ anode and a platinum cathode with propylene carbonate containing I₃-I⁻ as an electron donor and electrolyte. (Citation 1-5)”

In the example of photoelectrical chemical cells (above), the references to earlier work focus on one area of application described in detail, rather than outlining the range of areas of applications. Thereby, the citations have a character of “intradisciplinary magnifier” citing advances in chemistry journals such as Journal of the American Chemical Society. The examples above are included to illustrate that although the two research teams cited the same reference (citation number 3 made to the Oregan and Gratzel 1991 paper in both examples), their knowledge bases and the context in which they cite the paper differ. The analysis of citation data using bibliographic coupling (Figure 1-3) also confirms the two distinct research team networks. Also the co-authorship data at the level of the research department, in Figure 4, confirms the differences in international contra national grounding for the advances of scientific publication in the dataset analysed. The pattern of mainly national collaboration in the physics network has the origin in double affiliation of authors, existing institutional research (and laboratory) networks and existing repetitive collaboration on measurement techniques. Furthermore it does not rule out publications in internationally recognised journals or international collaboration in other science fields of the department or in other areas.

Conclusions

The research design of analysing two related fields of nanostructured materials using bibliographic coupling and co-authorship analysis reveal two distinct knowledge networks in chemistry and physics. This distinction is also made in analysis of nano-related journals in one chemistry-cluster and one physics-cluster (Leydesdorff and Zhou, 2007). Some conclusions regarding the methods used are that a combination of relational data using co-authorship data and bibliographic coupling can be applied to a limited research area, given limits of data and visualization methods, for examining knowledge interaction in development of environmental technology applications. The study of the knowledge networks provides novel insight into scope and limits of current knowledge interaction across department boundaries, based on citation mapping and existing co-authorship relations. Complementarities in capabilities, skills and research interests in combination with everyday funding

decisions, targeted actions and structural barriers of interdisciplinary science determine arguments and counterarguments for scientific research teams applying an “interdisciplinary outlook” or “intradisciplinary magnifier” pattern. Further work, includes analysis of cross-disciplinary scope in applications of knowledge in patents by academic inventors and among technicians involved in developing demonstration technology for early applications of scientific knowledge.

References

- Arnall, HA (2003) Future technologies, today's choices: Nanotechnology, artificial intelligence and robotics; a technical, political and institutional map of emerging technologies.
- ESRC (2003). The social and economic challenges of nanotechnology, Report by S Wood, R Jones and A Geldart, ISBN 086226-294-1. The Economic and Social Research Council (ESRC), UK.
- Gilbert, LE (1998) Disciplinary breadth and interdisciplinary knowledge production, *Knowledge, Technology and Policy*, 11,4-15.
- Guidi, V, MC Carotta, M Ferroni, G Martinelli and M Sacerdoti (2003) Effect of Dopants on Grain Coalescence and Oxygen Mobility in Nanostructured Titania Anatase and Rutile, *Journal of Physical Chemistry B*, 107, 120-124.
- Hicks, D and S Katz 1997. The changing shape of British industrial research. STEEP Special Report No. 6, SPRU – Science and Technology Policy Research, University of Sussex.
- Leydesdorff, L (2006). Bibcoupl.exe used for generating networks based on bibliographic coupling, downloaded from <<http://users.fmg.uva.nl/lleydesdorff/software.htm>>, last accessed february 2007.
- Leydesdorff, L and P Zhou (2007). Nanotechnology as a field of science: its delineation in terms of journals and patents, *Scientometrics*, 70(3), 693–713.
- Melin, G and O Persson (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36(3), 363-377.
- Meyer, M (2000a). Does science push technology? Patents citing scientific literature. *Research Policy*, 29, 409-434.
- Meyer, M (2000b). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, 49(1), 93-123.
- Meyer, m (2000c). Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology. *Scientometrics*, 51(1), 163-183.
- Meyer, M and O Persson (1998). Nanotechnology – interdisciplinarity, patterns of collaboration and differences in application. *Scientometrics*, 42(2), 195-205.
- Moss, JA, JC Yang, JM Stipkala, X Wen, CA Bignozzi, GJ Meyer and TJ (2004) Sensitization and Stabilization of TiO₂ Photoanodes with Electropolymerized Overlayer Films of Ruthenium and Zinc Polypyridyl Complexes: A Stable Aqueous Photoelectrochemical Cell, *Inorganic Chemistry*, 43, 1784-1792.
- Oregan, B, and M Gratzel (1991) A low-cost, high-efficiency solar cell based on dye-sensitized TiO₂ films, *Nature*, Vol. 353, pp. 737-740.
- Rinia, EJ, TJ van Leeuwen, HG van Vuren, and AFJ van Raan, (2001) Influence of interdisciplinarity on peer-review and bibliometric evaluations in physics research. *Research Policy*, 30(3), 357-361.
- Rhoten, D and A Parker (2004) Risks and Rewards of an Interdisciplinary Research Path, *Science*, Policy forum, Vol 306, p.2046.
- Royal Society (2003). Nanoscience and Nanotechnology, <<http://www.royalsoc.ac.uk/nanotechnology>>, last accessed 4 August 2005.
- Schummer, J (2004) Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics*, 59(3), 425-465.
- Small , H (1999) Visualizing science by citation mapping, *Journal of the American Society for Information Science*, 50(9), 799-813.
- Tijssen, RJW (1991) A quantitative assessment of interdisciplinary structures in science and technology: co-classification analysis of energy research. *Research Policy*, 21, 27-44.

Strength and Weakness of National Science Systems. A Bibliometric Analysis through Cooperation Patterns

The N. van Leeuwen and Robert J.W. Tijssen

leeuwen@cwts.nl

CWTS (Center for Science & Technology Studies), Leiden University,
Wassenaarseweg 52, 2300 RB Leiden (the Netherlands)

Abstract

Scientific cooperation is becoming more and more important in nowadays scientific organization, and while scientific cooperation, and especially international collaboration is supposed to generate high scientific impact, we will show in this study that lately, research publications resulting from international cooperation are losing ground in terms of scientific impact. In this study, scientific cooperation is analyzed over the last 25 years. The worldwide scientific output as indexed in the citation indexes and that of a selected set of countries is analyzed through distinguishing various types of scientific output, and impact measures are determined. In the study we focus on a number of country's for which we have observed strong changes in the patterns of their scientific output. The study shows the importance of the analysis of the composition of the scientific output of a country, and the role single address and first-authored publications play in applications of bibliometric data in research assessment processes on the national level.

Keywords

scientific cooperation; co-authored research publications; first-authorship; impact assessment

Introduction

Scientific collaboration has become more and more important in modern day scientific activity. With the appearance of large virtual network, not only in physics around the large physics-related instrumentation, like CERN and synchrotron facilities, to name a few, but also in medicine, with initiatives around the unraveling of the human genome, and large supra-national comparative study groups. In the field of bibliometrics, scientific cooperation, and in particular international cooperation has been focal point of research over a long period. While the issue of international cooperation was researched on the level of countries (Moed et al, 1991, Glänzel & de Lange, 2001) and institutional scientific cooperation (Bordons et al, 1996, Zitt et al, 2000), more recently the focus has shifted towards network structures (Glänzel & Schubert, 2004, Calero & Moed, 2006). In the Dutch Observatory of Science & Technology, the focus was in 2003 on scientific cooperation as can be measured through publications published in international refereed journals, showing a huge increase in the growth of the share of international cooperation, especially an increasing growth of intra-EU cooperation (NOWT 2000). Recent research showed the growing intra EU cooperation, discussing an 'Europeization' of the research conducted within the boundaries of the EU (Mattsson et al., forthcoming). A growing awareness amongst policy makers in various countries of the possibilities of bibliometric data on country level created an urge for making comparisons, not only of countries but also of institutes. However, the comparison and ranking of these various entities within science systems also demanded more insight in the nature and composition of national outputs, and in particular the role of international scientific cooperation as an impact-generating mechanism in a country's impact score. In this study we will focus on the composition of a country's output in terms of the share of national and international scientific cooperation, and more in particular the role of single address papers and first authorships within the national output of countries.

Methodology

In this study we used data collected form the Web of Science as supplied to CWTS by Thomson Scientific. This database covers the period 1981 up until 2006, although the last year used for this analysis was 2005. The analysis is based on database years, that is, the moment that publications were processed for the Web of Science. The dataset used for the study contains 17.427.506 publications, and is limited to articles, letters, notes, and reviews. From those publications, we collected all addresses, and used the country names attached to the publications. Those publications resulting from

large international collaborations (like virtual institutes, networks, and consortia), and for which we could not distinguish addresses, these publications are excluded from the analysis.

Indicators for scientific collaboration are based on an analysis of all addresses in papers in the database. Each paper is classified in one of three categories. First, we identified all papers authored by scientists sharing the same one-single address, i.e., from the same research unit or institute, and thus from one country. These papers are classified as '*single address*' or *SA*, as they involve no collaboration or only 'local' (i.e., within the research group) collaboration. A paper is classified as '*international collaboration*' or *IC* if two or more different country names appear in the address list of a publication. The remaining papers are classified as '*national collaboration*' or *NC* when there is one single address from the respective university in combination with one or more different addresses but from all from the same country. For example, if a paper is the result of collaboration with both a Dutch institution and an institute outside the Netherlands, it is marked as '*international collaboration*'.

To define 'first authorship', we started from the hypothesis that the first author of a publication and the first mentioned address (and thus the country name) are connected. And thus the first mentioned country name on a scientific publication in the database is the country from which the first author originates. While for the international collaboration publications, every publication is added to a country's set of publications if the country's name appears on the address list, we distinguished in this dataset between those publications in which a country is mentioned on the first position of the address list, and all the publications for which we do not find this. This set is indicated as '*international collaboration-first authorship*' or *IC-FA*.

As these three main classes (*SA*, *NC*, and *IC*) are mutually exclusive, we can produce country shares per class over a period of time. In this study we determined annual changes over the period 1981-2005, and within this period we focused on the last ten years. Next, in this study we determined impact scores. These impact scores are the country's actual impact (*CPP*), compared with the mean field impact (*FCSm*, which stands for the mean Field Citation Score) of that country, as determined by the output of the country itself. This *FCSm* is the mean (world-wide) citation rate of the fields in which the country has published, taking into account the type of paper (e.g., normal article, review) as well as the years in which the country's papers were published. Our definition of sub-fields is based on a classification of scientific journals into *categories* developed by Thomson Scientific. Although not perfect, it is at present the only classification that can be automated consistently in our data-system, and that fits the multidisciplinary nature of the journals processed for the Web of Science. A country is always active in more than one field of science, and thus we calculate a weighed average value, the weights being determined by the total number of papers the country has published in each field.

So we compare the average number of citations to a country's output (*CPP*) to the relevant field mean citation scores (*FCSm*), by calculating the ratio for both. If the ratio *CPP/FCSm* is above 1.0, the country's output is cited more frequently than an 'average' publication in the field(s) in which the country is active. About 80 percent of all indexed papers are authored by scientists from the United States, Canada, Western Europe, Australia and Japan. Therefore, the 'world' average is dominated by the Western world.

In this study we applied a citation window of three years: the year of publication, and two years more. Although this is considered somewhat short, due to the field normalization procedure this problem is solved, as all publications in a field suffer from the same 'handicap'. Furthermore, the choice for a three year window is influenced by the wish to generate as actual as possible citation scores, and 2003 is now the most recent year for which a three year window can be applied.

The study will focus on global science, and a number of countries. For the countries in this study, we selected a number of Anglo-Saxon countries (Great Britain, USA, Canada, and Australia), and a number of European countries (two well-known 'high-impact' countries, Switzerland and the Netherlands, and two upcoming European nations, Finland and Spain).

Results

In this section we will present results of our analyses. First we will show how the three main classes we defined above are distributed among the worldwide output in all fields of science, including the social sciences and humanities. This is presented in Figure 1. In this figure we see the shift in types of scientific activity over the last 25 years, in general a shift from single address publication output towards stronger scientific cooperation. We observe a decrease of the share of single address publications of 70% in 1981 to 43% in 2005, and consequently a change of cooperation-based research output of 30% in 1981 to nearly 60% in 2005 (combining both national and international cooperation).

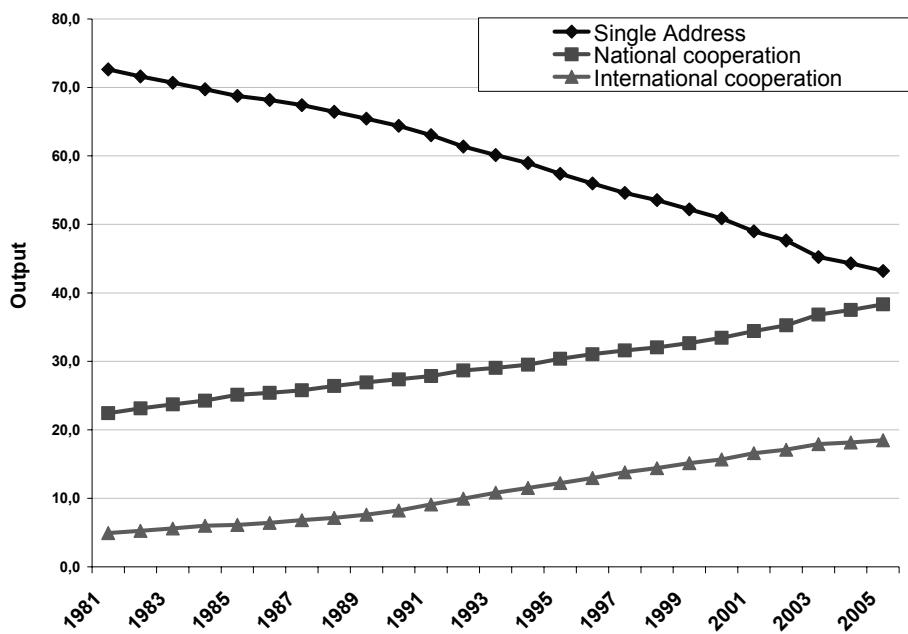


Figure 1. Shares of various output types during 1981-2005

Given this huge change in output patterns over the last 25 years, where international cooperation as a share of the global research output more than tripled, and national cooperation over the last 25 years nearly doubled, the focus is then on impact scores in the global science system. The results of that analysis are displayed in Figure 2. Here we present field-normalized impact scores (CPP/FCSm) over the period 1981-2003 (due to the choice for a three year citation window, no impact scores can be generated for the years 2004 and 2005, since these years do not have a fully filled citation window at the moment this study takes place). We observe the highest impact scores for publications that result from scientific cooperation, with the highest impact scores for those publications that result from international cooperation. This is consistent with the findings we encountered during numerous studies, where the highest impact scores are generated by international cooperation. Publications that carry only one single address tend to have low(er) impact scores, as compared to the two other types. However, we also observe a decreasing trend in the impact scores resulting from scientific cooperation, in particular on the output that results from national cooperation.

If we focus on the last ten years of the total period analyzed (Figure 3), we find for the output trends strong tendencies for all three classes of scientific output. The SA output decreases very strongly, from nearly 56% to 43% ($R^2=0.99$), the NC-output increases strongly, from 30% to 38% ($R^2=0.98$), and the IC-output nearly increases from 13% to 19% ($R^2=0.98$). So the changes observed over the full period 1981-2005 are also very well visible within the period of ten years 1996-2005.

As we focus on the impact scores of the last eight years (1996-2003) of the total period, we find for the impact trends strong, but varying tendencies for all three classes of scientific output, albeit somewhat less strong as we observed for the output shares. In Figure 4, we find that the SA impact increases, albeit it very slowly, from nearly 0.82 to 0.86 ($R^2=0.86$), the NC-impact decreases strongly,

from 1.17 to 1.05 ($R^2=0.98$), and the IC impact decreases as well, from 1.32 to 1.24 ($R^2=0.98$). So the changes observed over the full period 1981-2005 are very well visible within the period of eight years 1996-2003.

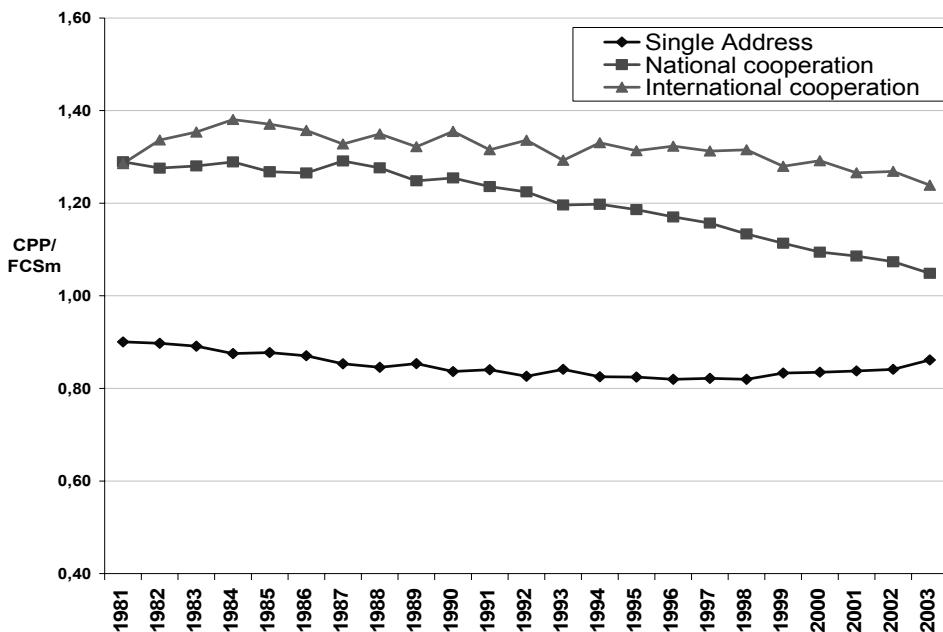


Figure 2. Field-normalized impact scores of various output types during 1981-2003

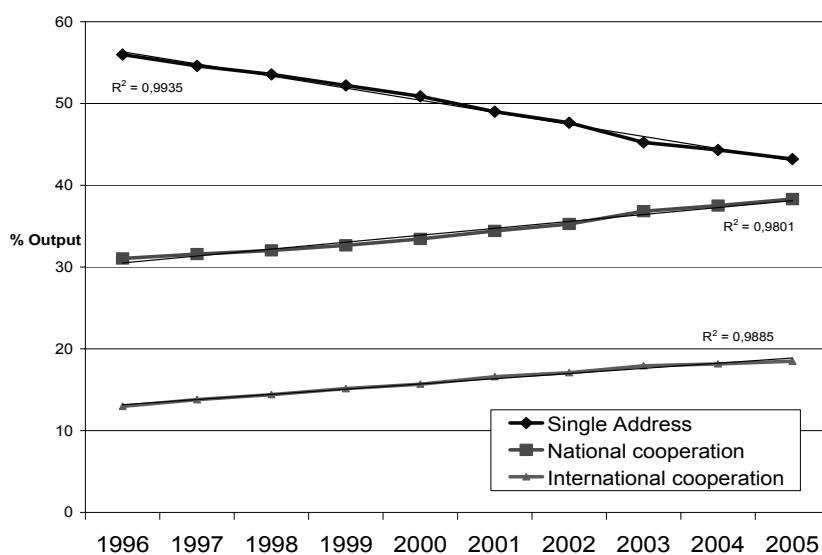


Figure 3. Development of the shares of various output types during 1996-2005

So here we find a decrease in impact scores, while scientific cooperation is in general supposed to be rewarding, as is commonly assumed and underlined by numerous studies conducted over the last years. So the question arises what is causing this development in the global science system, especially in the light of the change of output shares across the three main types of output distinguished in this study.

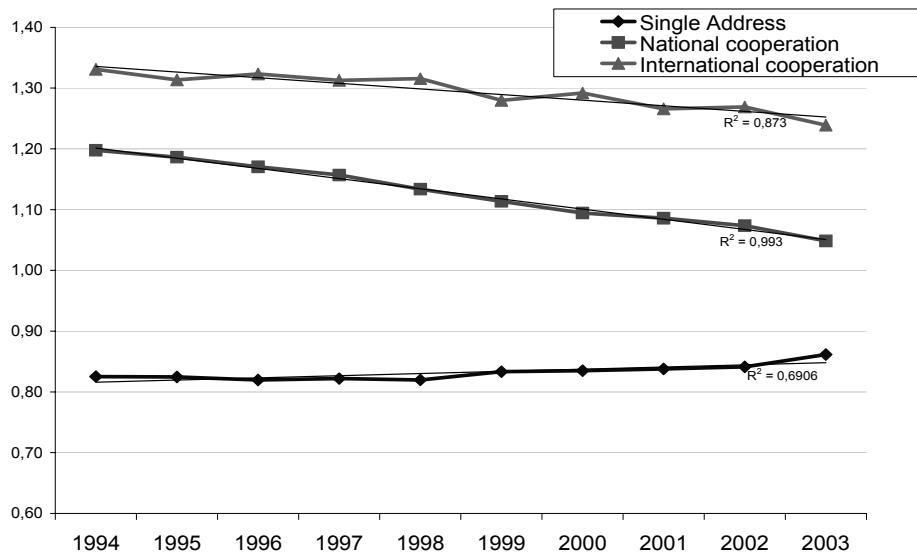


Figure 4. Development of various impact types during 1996-2003

In Figure 5, we show the development of the output of a number of countries, some Anglo-Saxon countries (Great Britain, USA, Canada, and Australia), and some European countries (the Netherlands and Switzerland, as two countries with high impact scores, and Finland and Spain, as two runner-up countries within the European Research Area). The output numbers over the period 1981 – 2005 are presented, on a log scale to be able to show the development of both the USA as well as that of the other seven countries. One clearly observes the well-discussed ‘stagnation’ of the output growth of the USA, just like that of Great Britain and Canada. For the other countries we clearly observe a growth of the output, with perhaps the strongest growth in output found for Spain (within this sample of countries, Spain rises from the second-last position in terms of output to the fourth position, within a broader context Spain is among the top-ten ranking countries in terms of national output in internationally refereed journals as indexed in the WoS). In the legend of Figure 5 we find the R^2 -values of the trend in output numbers, and we find only for Canada a relative low R^2 -value, due to a significant decrease in output numbers between 1996 and 2003 for Canada.

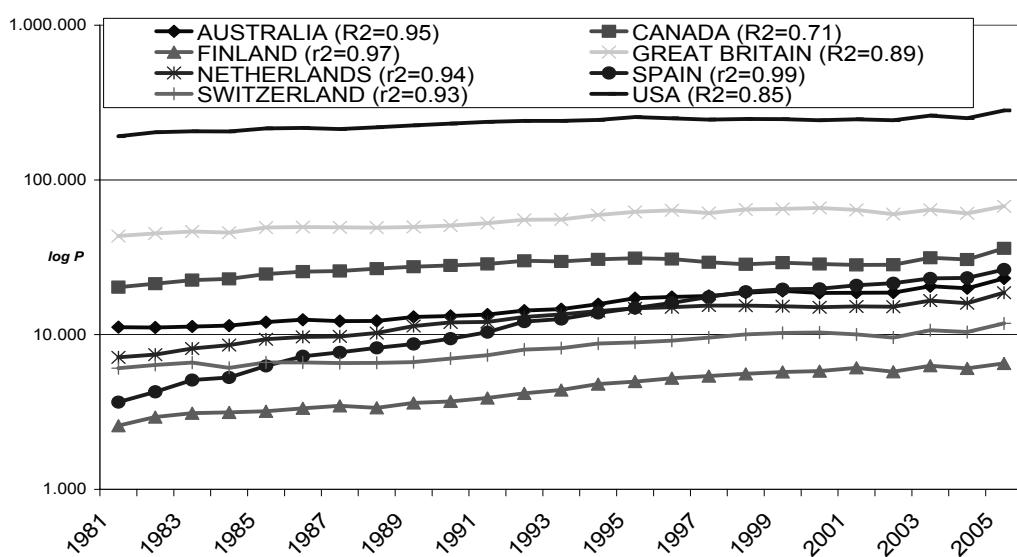


Figure 5. Output development for eight countries, 1981-2005.

In Figure 6, the impact trends related to the output data shown in Figure 5, are presented. Roughly, we can see a decrease of the impact of the USA and to a lesser extent, Switzerland. We see a strong increase of the national impact of Spain, and a somewhat more modest increase for Finland. Both Great Britain and the Netherlands seem to be rather stable, while Canada and Australia show a recovery of the impact scores at the beginning of this millennium, after a decrease in the nineteen nineties (Butler, 2002). So in general we find a decrease of the impact for two countries for which we normally find (very) high impact scores.

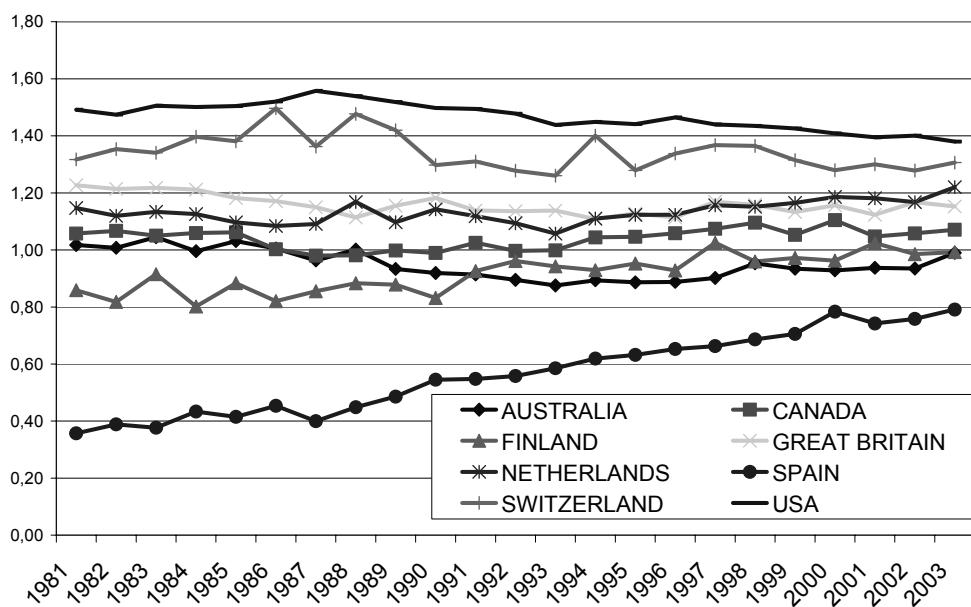


Figure 6, Impact development (mean 3-year citation impact data, based on field-normalized impact scores), 1981-2003/05

In Figure 7, ‘first authorship’ is introduced in the analysis. This aspect of scientific publishing is commonly recognized as an indication of scientific leadership. Having a high share of first authorships as a group, or institute as compared to the total output is an indication of the extent to which the entity succeeds in getting the findings of their research published in which that entity plays a strong role themselves. In particular when we focus on first authorships on publications which are the result of international cooperation, we can identify to which extent countries are taking the lead in publishing of scientific results, as produced by international collaboration.

We then can compare the four types of scientific output as defined above (SA, NC, IC, and IC-FA), keeping in mind that IC-FA is a sub-set of IC, on the mean impact scores generated by these various output types over the period 1981-2003/05. The impact scores presented in Figure 7 show average field-normalized impact scores, calculated over a three year period. So the mean field-normalized impact for Finland on its IC-output is nearly 1.60, and for Switzerland slightly above 1.80.

If one compares the impact scores per type, it becomes clear the USA, Switzerland, Great Britain, the Netherlands, and Canada do have the strongest impact position in both IC and IC-FA, with none of the countries having an impact score below worldwide average level, except for Spain in first-authored international cooperation publications. However, if we focus on both NC-output and SA-output, we get a different picture. On these types of scientific output, we find more low impact scores, and remarkable is the low impact score of Canada on SA-output as compared to their IC and IC-FA output. Actually, only Switzerland and the USA have impact on SA-output that is significantly above worldwide average level, while both the Netherlands and Great Britain perform at that worldwide average level on this type of scientific output.

Another remarkable outcome is the comparison of IC and IC-FA related impact scores for both Switzerland and the USA. While the impact score of the USA on IC-output is the second highest (after Switzerland), and closely followed by Finland, the Netherlands, Great Britain, and Canada, the impact of the USA on their IC-FA output is the only score that exceeds the impact level of the IC-output based impact score.

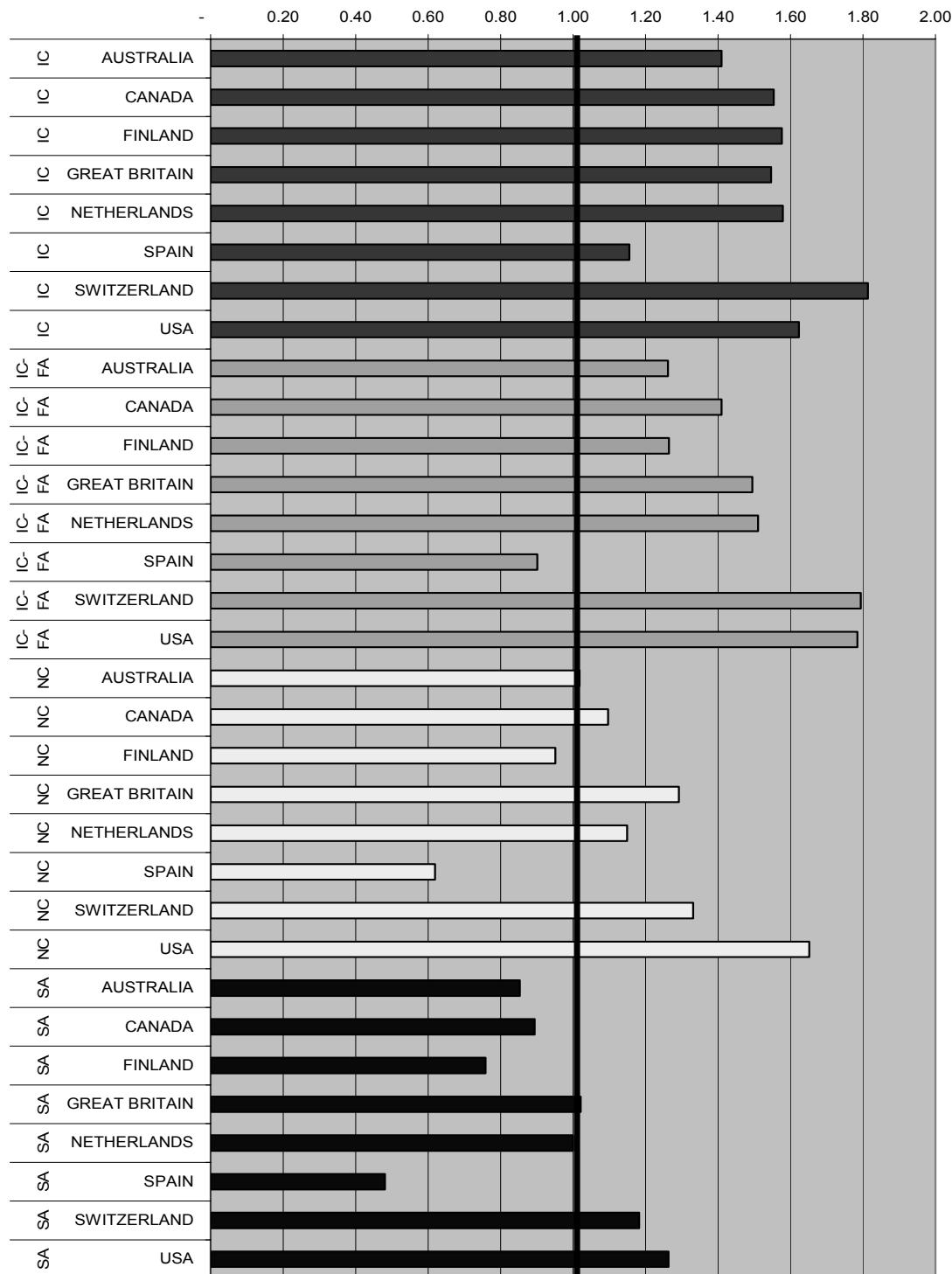


Figure 7. Mean field-normalized impact scores of various types of scientific output, 1981-2005

Two major findings of this comparison are in the first place that SA-output related impact of 1.00 (which stands for worldwide average impact level) or higher is an indication of the strength of the

science system, as we find that some countries have high impact scores on output resulting from international cooperation, but relatively low impact scores on output resulting from single address publications. Yet another finding relates to the application of first authorship in this type of analyses. As the impact situation of the USA on IC and IC-FA publications indicates, first authorship can be bibliometrically used as an indication of the leadership of a country.

In Table 1, we present the mean impact scores of SA, IC, and IC-FA output, for both the full period 1981-2003/05 and 1994-2003/05. Per period we calculated the differences between the impact scores generated by the three types we focus our analysis on. This difference is expressing the relative position of one impact score as compared to one another. For example, the relative difference between the USA impact level of IC and IC-FA output is 16%, which expresses the higher impact level of USA IC-FA output (1.78) as compared to the impact of the country's IC-output (1.62). For Finland we find the difference between IC and IC-FA impact levels to be the highest, namely -31%, which means that IC-FA impact is 31% lower than the Finnish IC-output (1.58 versus 1.26, respectively). The differences between two types (e.g., SA-IC), show in the comparison between two countries interesting outcomes: although the actual impact levels for SA and IC output for both countries differ strongly, the relative difference between SA and IC output between Spain and Switzerland is roughly at the same level. The differences among the three types of scientific output are graphically shown in Figures 9a and 9b, comparing the two periods analyzed in Table 1.

Table 1. Impact scores on SA, IC, and IC-FA output ('81-'03/05 and '94-'03/05)

	SA	IC	IC-FA	Diff IC/IC-FA	Diff SA-IC	Diff SA/IC-FA
1981-2003/05						
AUSTRALIA	0.85	1.41	1.26	-15%	56%	41%
CANADA	0.89	1.55	1.41	-14%	66%	52%
FINLAND	0.76	1.58	1.26	-31%	82%	51%
GREAT BRITAIN	1.02	1.55	1.49	-5%	52%	47%
NETHERLANDS	1.00	1.58	1.51	-7%	58%	51%
SPAIN	0.48	1.16	0.90	-25%	67%	42%
SWITZERLAND	1.18	1.81	1.79	-2%	63%	61%
USA	1.26	1.62	1.78	16%	36%	52%
1994-2003/05						
AUSTRALIA	0.81	1.43	1.21	-22%	62%	40%
CANADA	0.88	1.65	1.44	-21%	76%	55%
FINLAND	0.81	1.62	1.21	-41%	81%	40%
GREAT BRITAIN	0.98	1.54	1.46	-8%	56%	48%
NETHERLANDS	1.01	1.57	1.45	-12%	56%	44%
SPAIN	0.62	1.27	0.94	-33%	65%	33%
SWITZERLAND	1.15	1.74	1.70	-5%	60%	55%
USA	1.19	1.59	1.77	18%	41%	58%

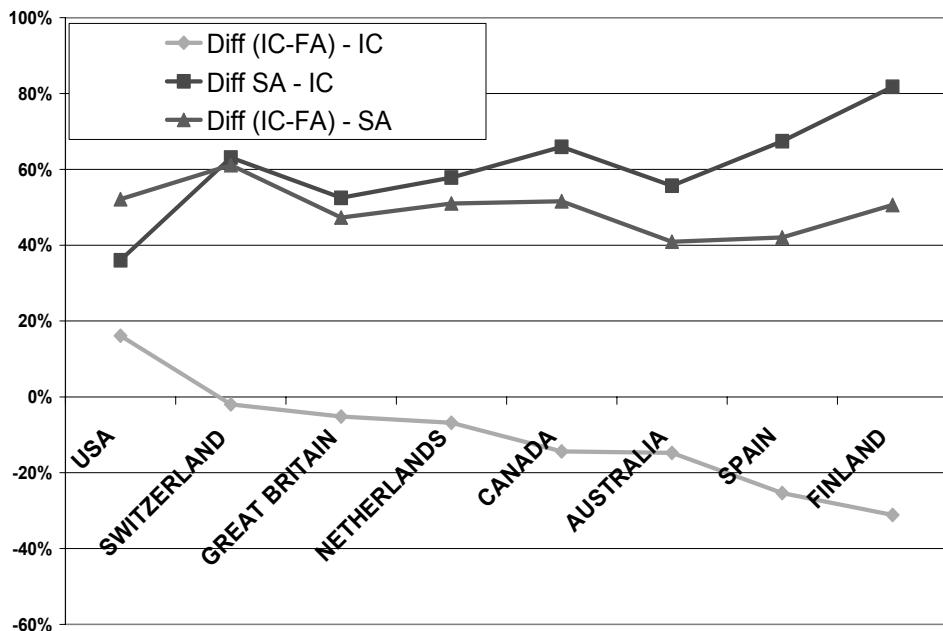


Figure 8a. Relative differences in impact between types of scientific output, 1981-2003/05

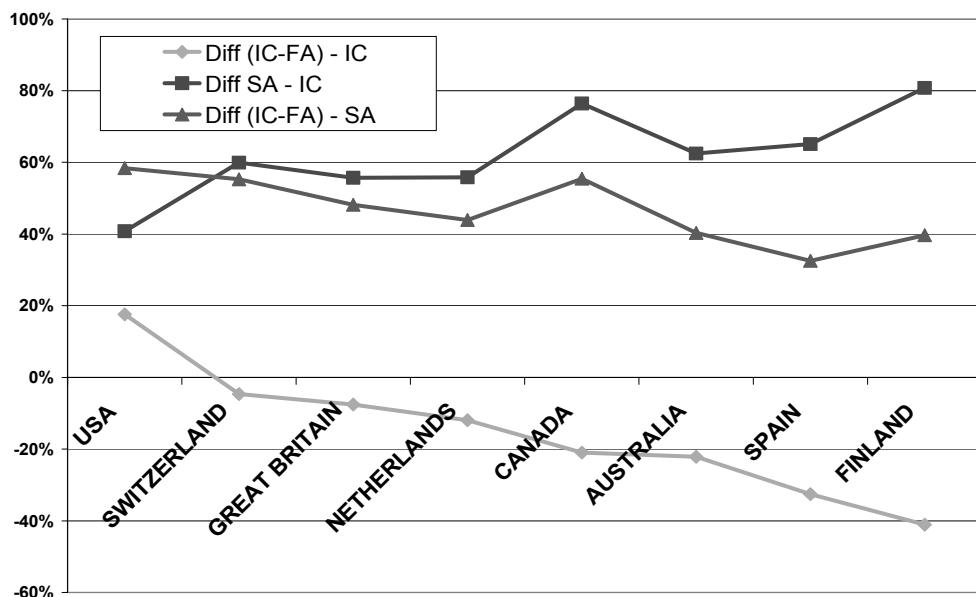


Figure 8b. Relative differences in impact between types of scientific output, 1994-2003/05

Figures 8a and 8b clearly show the relative differences in impact level between the various types of scientific output. Both figures are organized by the descending difference between the impact of IC-output and IC-FA output. In Figure 8a, the difference in impact between SA-output and IC-output is diverging from that between IC and IC-FA output (with an exceptional situation for Canada, for which we find relative strong difference between SA and IC output related impact scores). In the more recent period (1994-2003/05), displayed in Figure 8b, we find the same divergence between differences in (IC-FA) - IC on the one hand and SA - IC-output on the other hand, while we also observe a decreasing trend of the difference in impact between SA and IC-FA output, indicating a relation between both types of scientific output (Pearson rank correlation increases from 0.42 to 0.77).

However, we find for Canada again the same remarkable position on the comparison of impact scores on various types of scientific output.

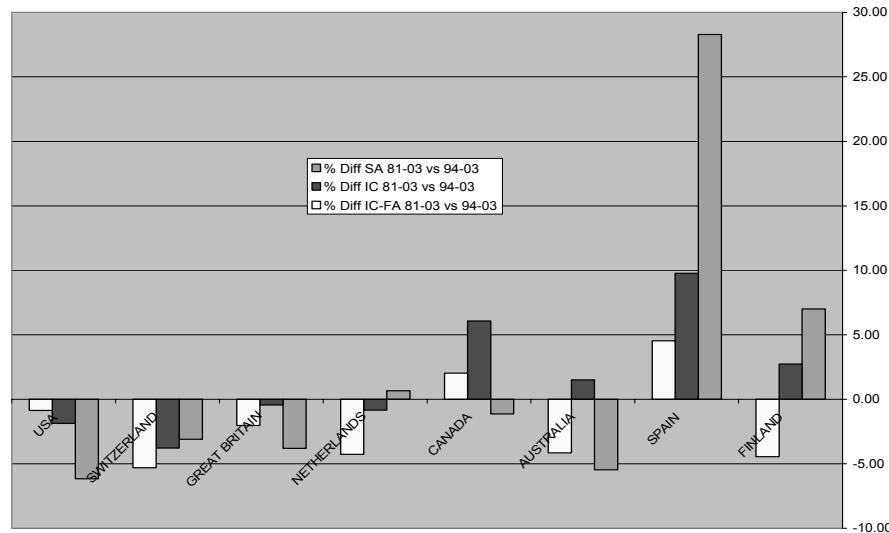


Figure 9. Changes in impact level on three types of output

Figure 9 displays the differences in impact levels of the various types of scientific output for the selected countries, when comparing 1981-2003/05 with 1994-2003/05. We find decreasing impact levels of all three types of scientific output for the USA (>5% decrease on the impact of SA output), Switzerland (5% decrease of the impact level of IC-FA output), and Great Britain (relative strongest decrease in impact found for SA output). Then, for the Netherlands we observe an increase of the impact related to SA-output, while we observe a decrease for Canada on this type of output, next to an increase of the IC-FA and particularly the IC-FA output. For Australia we observe decreasing impact scores for both SA (>5%) and IC-FA output. Finally, for Spain and Finland we find mainly increasing impact scores, with an exception for IC-FA output of Finland. Spain shows increasing impact levels of all three types, in particular of the SA-output of the total Spanish research output (increase of 28%).

Conclusions and Discussion

In this study we have focused on long term developments of various types of scientific publishing, and the field-normalized impact generated by these various types. The types of scientific output distinguished are output resulting from international cooperation, national cooperation, and single address publications, in which no apparent cooperation is found. A fourth type is distinguished by focusing on first authorship, especially in international cooperation. Changes in especially the share of a country's output from first-authored international cooperation and the share of single address publications can be regarded as indicators of strength and/or weakness of a science system.

Over the full period 1981-2005, we find a strong change in the patterns of scientific output on a global scale. Scientific cooperation has taken a much more significant place in the worldwide science system. While (international) cooperation is regarded to generate high(er) impact scores in comparison with publications that result from one institute (and thus one country) only, we find in this study that publications resulting from international and national cooperation seem to have lower impact scores. This finding asked for a more in-depth analysis on the country level, and eight countries were selected. For these eight selected countries, we found increases in output in the period 1981-2005 (albeit it somewhat slower for the USA, Great Britain, especially, Canada). In terms of impact, we found various patterns: the USA and Switzerland show decreasing national impact scores, while especially Spain and Finland show increasing scores. When we focus on the composition of the national output in the various types of scientific output, we find that the various countries follow their own pattern of development, both in output shares as well as in impact generated by this output. However, it becomes

clear that traditional leadership in science by Switzerland, the USA, the Netherlands and Great Britain is under pressure, but still intact, we also notice some runners-up. While the difference in impact level generated by single address output and first-authored international cooperation publications tends to decreasing, thereby indicating an increase of the impact of single address publications up to international cooperation impact level, we also find increasing impact scores for Spain and Finland on nearly all types of scientific output. A remarkable finding remains the fact that the USA, despite its huge annual output, is still leading as can be concluded by the fact that their single address publications have the highest impact, and the fact that their first-authored international cooperation publications have a higher impact than the impact score generated by all international cooperation publications in which the USA is involved. A final observation relates to the position of Switzerland as the second ranking country in terms of impact. While we previously found Switzerland as the number one ranking country, we now find the USA on this position. This change in the position of Switzerland is due to the composition of the Swiss output in this analysis. While we used in previous studies the CD-Rom version of the citation indexes, we have now switched to the WoS. As the WoS includes more journals, and thus more publications in other languages than English, the Swiss impact position is under pressure due to an increased share of their output in German language journals. As these journals are only moderately cited, and thus contribute hardly to the total number of citation received by Swiss research output, the impact of the total country decreases as their output increases relatively stronger(see van Leeuwen et al, 1999), notwithstanding the presence of such an institute as CERN in Switzerland.

An interesting issue is the question how these findings relate to the current global ranking hype ? Although the ranking analyses are mainly focusing on the institutional level, the discrepancy of high impact scores for institutes and decreasing global impact scores, especially for cooperation publications, put these analyses in a new perspective if one does not take into consideration the various aspects of scientific collaboration, and the position a country takes in its various types of scientific output.

Finally, future research will go deeper into the underlying disciplinary profiles of the countries in the study, to see which disciplines and fields are ‘responsible’ for the changes described in this study.

References

- Bordons, M, I. Gomez, M.T. Fernandez, M.A. Zulueta, and A. Mendez, (1996) Local, domestic and international scientific collaboration in biomedical research, *Scientometrics* 37, 279-295
- Butler, L., A (2002) list of published papers is no measure of value - The present system rewards quantity, not quality - but hasty changes could be as bad. *Nature* 419, 877-877.
- Calero Medina, C and H.F. Moed, (2006) Depicting the landscape of research universities, presentation at the 9th International Conference on Science & Technology Indicators, 7-9 September 2006, Leuven, Belgium
- Glänzel, W, and C. de Lange, (2001) A distributional approach to multinationality measures of international scientific collaboration, *Scientometrics* 54, 75-89
- Glänzel, W, and A. Schubert, (2004) Analysing scientific networks through co-authorship, chapter 11 in Moed H.F., Glanzel W. and Schmoch, U. (eds.), *Handbook of Quantitative Science and Technology Research*, page2 57-276, Springer
- Mattsson, P., P. Laget, A. Nilsson, and C.J. Sundberg, Intra-European vs. extra-European scientific co-publication, submitted to *Scientometrics*.
- Moed, H.F., R.E. De Bruin, A.J. Nederhof, and R.J.W. Tijssen, (1991) International scientific co-operation and awareness within the European Community: Problems and perspectives, *Scientometrics* 21, 291-311
- NOWT 2000, Report of the Dutch Observatory of Science & Technology 2000, page 64 (see www.nowt.nl, NOWT 2000 report).
- Van Leeuwen, Th.N., H.F Moed, R.J.W. Tijssen, M.S. Visser, and A.F.J. van Raan, (1999) Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance, *Scientometrics*, 51, 335-346.
- Zitt, M., Bassecoulard, E. and Okubo, Y., (2000) Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics* 47, 627-657.

Science in Brazil: Contribution of Male and Female Scientists¹

Jacqueline Leta and Flávio Martins Teixeira

jleta@bioqmed.ufrj.br, fmteixeira@bioqmed.ufrj.br

Universidade Federal do Rio de Janeiro, Instituto de Bioquímica Médica, Prédio do CCS, Bloco B – sala 39,
Cidade Universitária, CEP 21.941-590, Rio de Janeiro (Brazil)

Abstract

The establishment of science is a very recent event in most of the countries located in the south hemisphere. Training human resources for scientific and technological activities seems to be a means to reduce the scientific and technological gap between the two hemispheres. In this process, human diversity, including gender, may play an important role. This study presents data on the scientific publication output of male and female scientists from Brazil. Our aim is to identify some publishing trends, which may contribute to the debate on consequences that emerge from the imbalance between males and females participation in the scientific community, including administrative positions in academia.

Keywords

scientific publications; gender; Brazil; research output; policy-making.

Introduction

Different from European and North American countries, the establishment of modern science is a very recent event in most of the countries in the south hemisphere. In Latin American countries, the delay in having science institutionalized is easily noticed, for example, when we consider the foundation of their main science funding agencies, which occurred only after the 1950s. Although the consequences of scientific and technological output have contributed to changing life worldwide, there is an imbalance in this scenario. On the one hand, a reduced number of countries, developed ones, are responsible for most of the scientific and technological enterprise; on the other hand, the majority of countries, mostly developing ones, are consumers of products and knowledge produced by the former group. Such an imbalance was recently addressed by Annan (2003) in one of the most prestigious scientific journals. He argues that “The idea of two worlds of science is anathema to the scientific spirit”, and suggests that science and scientists can help the world in changing this scenario.

One of the challenges posed by this dilemma is the increase in the training of human resources for the scientific and technological enterprise. In this process, human diversity, including different ethnic groups, social status, and gender, merit special attention. Thus, we started studying the contribution of male and female scientists to Brazilian science. Our focus is not on feminism issues; our discussion centers around the actual contribution of females to the scientific enterprise, historically seen as a male activity and where the hierarchical structure of authority has traditionally been male dominated (Schiebinger, 1999).

The scientific community under study comes from Brazil. The country experienced a remarkable growth in the fraction of female scientists in the 1970s as a consequence of a higher entrance of females to universities (Pena, 2005). Data from the National Statistical Agency (IBGE, 2004) indicate that there were 121 females enrolled in PhD programs in the country (20.5% of the total enrollments) in 1972; after three decades, females were responsible for almost 6,000 (or 50%) of the total number of PhD fellowships awarded by the main funding agency of the country, the National Council for Scientific and Technological Development (CNPq, a). This growth contributed to changing the profile of the country’s scientists in terms of gender: in 2004, females represented 47% of the scientists registered in the CNPq database, and in 1995, they represented 39% of the whole community.

Despite these changes, the female scientists of Brazil (as well as worldwide) still have to face some challenges. Among these, their misrepresentation in scientific committees, as full professors, in

¹ This study was supported by CNPq (Proc. Nr. 401850/2004-8).

administrative positions in academia and as recipients of the “productivity in research”² fellowship a particular type of fellowship awarded by CNPq. There is a vast international literature describing and discussing the reasons behind female’s misrepresentation, among other reasons, the scientific productivity of researchers (e.g. Rossi, 1965; Long, 1973; Creamer, 1999). A common feature is that male scientists usually publish more than their female peers. Hence, with a higher scientific capital (Bourdieu, 2003), male scientists can get more prestige and progress faster in their academic career.

The lack of information about the contribution of Brazilian female scientists makes it hard to tell whether or not this picture is also true for its particular community. We present data on the scientific publication of male and female scientists from Brazil. Our aim is to identify some publishing trends considering gender. Our results are part of a large quasi-qualitative study we are carrying out in order to provide evidence of the extent to which gender and academic performance can be correlated in this particular community

Data set

One of the major difficulties to study the participation of males and females in the scientific literature is to identify the gender of the authors once most of the journals record only their last names and initials. This is probably the main reason why most of the studies focusing on gender in science are scarce for large scientific communities. Two examples, however, should be highlighted, namely the study with Iceland scientists developed by Lewison (2001) and the one in Poland by Webster (2001). In both cases, cultural characteristics for naming people in these countries made it possible to identify the gender of the authors. In order to overcome such limitation, all the quantitative analysis presented in this paper were extracted from CNPq database. This database covers information of more than 90% of all the human resources engaged in the Brazilian science, including graduate and undergraduate students, post docs, technicians, and researchers from all fields, with or without a permanent position in a research institution.. Our sub-set database comprises general information, such as the gender of only 51,233 junior and senior scientists with a PhD degree. Also, this particular database includes the number of publications by these scientists in domestic and international journals, during the 1997-2005 period.

Sample Characteristics

Among the 51,233 junior and senior PhD scientists of our sample, 30,725 are males and 20,480 are females (Figure 1A). This distribution is, however, a little different from the data obtained directly from the CNPq website, which provides the fraction of total male and female scientists and do not take into account academic degrees (40% female versus 47% male). In both cases, it is clear that scientists are predominantly male in Brazil. This trend may be associated with the recent entrance of females to scholarly research. In fact, time (in years) in academia since the PhD degree for males and females (Figure 1B) corroborates this statement. Scientists are predominantly female among those that have earned the degree more recently and represent very low fractions among those more “experienced” scientists.

Results and discussion

The number of papers published in the 1997- 2004 period by female and male scientists from our sample is shown in figure 2. As can be seen, male scientists (square) tend to publish more than female ones (diamond) in both analyses: they have more papers published in domestic and international journals (figure 2A) and in international journals alone (figure 2B).

This difference in the publishing rate is also presented in table 1. The fraction of females with no publications is higher, especially among those in international journals. Among the most productive scientists (those with 17 or more publications in the period), females represent a very low fraction; females’ share in international publications is less than half of that observed for males.

² The “productivity in research” is a fellowship awarded by CNPq to a very tiny fraction of the country’s researchers. It is renewed every year through a peer review process and, among the criteria used, it takes into account the researcher’s publication outputs.

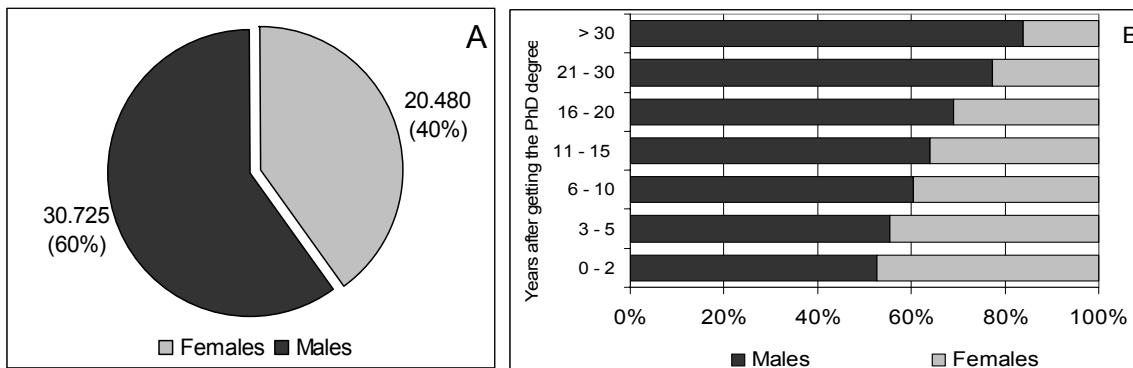


Figure 1. Share of Brazilian scientists according to gender (A) and years since PhD degree (B).

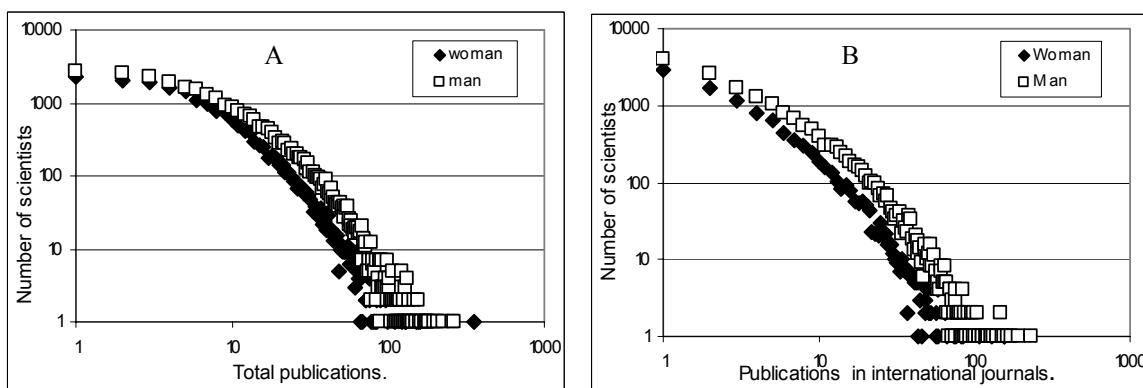


Figure 2. Distribution of Brazilian scientists according to the number of papers published in domestic plus international journals (A) only in international journals (B) in the 1997-2004 period.

Thus, the imbalance between male and female authors in terms of papers published in international journals is in accordance with previous studies. However, different from European countries and the U.S., most of the female researchers from our sample may be considered newly independent (figure 1B) who will need some time to establish their carriers. In fact, 55% of them earned the PhD degree five or so years ago.

Other factors other than career stage deserve attention in the assessment of the publication output of scientists. One of these factors is their field of research. It is well known that depending on the field, scientific productivity may vary considerably. As for gender, the international literature reports that female scientists are still concentrated in some particular fields, especially in the “soft sciences”, whereas they are still scarce in technological fields. This “territorial segregation” (Schiebinger, 2001) is a worldwide phenomenon that has recently been discussed by Hermann & Cyrot-Lackmann (2002) in a paper on French science.

The general trend presented in table 1 does not change significantly when fields, scientific productivity, and gender are considered for analysis. The definition of the six fields follows CNPq classification (CNPq, b).

Some interesting features should be pointed out. In engineering, a traditional male-dominated field, there is a balance between females’ and males’ fractions of international publications. On the other hand, in the social sciences, a typical female-dominated field, the number of publications by female authors is not higher than that by male ones. Similarly, in biology, where scientists are predominantly female (3,803 versus 3002), they are not as much productive as their male peers.

Table 1. Fraction of Brazilian male and female scientists in the total number of publications and in publications in international journals, 1997 - 2004.

Number	PUBLICATION			
	Publication in international plus national journals		Publications in international journals	
	Female	Male	Female	Male
0	13,5%	14,1%	51,9%	45,2%
1 - 4	38,1%	31,4%	31,6%	30,8%
5 - 8	21,1%	18,3%	8,4%	9,8%
9 - 12	10,8%	10,8%	3,5%	4,8%
13 - 16	5,8%	7,1%	1,7%	3,0%
17 - 20	3,3%	4,6%	1,1%	1,9%
21 - 24	2,2%	3,3%	0,5%	1,2%
> 24	5,1%	10,5%	1,2%	3,2%
No scientists	20480	30725	20480	30725

Table 2. Fraction (%) of male and female authors of international publications in different fields, 1997– 2004.

Nr of Pubs	Agriculture		Biology		Medicine		Chem &Phys		Engineering		Social Sc.	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
0	58,5	59,7	26,8	22,7	53,3	43,4	30,9	25,9	46,0	45,9	90,4	87,5
1 - 4	32,4	29,2	40,9	34,1	30,9	30,9	39,9	32,8	37,3	35,2	9,2	10,7
5 - 8	5,4	6,3	15,8	16,1	8,1	10,9	13,9	14,1	7,7	9,4	0,3	1,0
9 - 12	1,9	2,3	6,9	9,8	3,5	5,2	6,4	7,9	4,1	4,1	0,1	0,4
13 - 16	0,8	1,0	4,0	5,0	1,9	3,5	2,9	5,8	1,1	2,0	0,0	0,2
17 - 20	0,6	0,6	2,6	3,3	1,0	2,2	1,8	3,6	1,5	1,3	0,0	0,2
21 - 24	0,1	0,2	1,1	2,3	0,5	1,4	0,9	2,7	0,9	0,6	0,0	0,0
> 24	0,3	0,6	1,9	6,7	0,9	2,5	3,3	7,2	1,4	1,5	0,0	0,0
Nr. Of scientists	1851	4162	3803	3002	3146	3821	2857	7070	1355	5283	1809	3207

In terms of fields, is there any difference between the number of papers between female and male authors from our sample? Do male and female scientists have different efficiency in publishing? To address this question we carried out a *per capita* analysis (figure 3). In fields of more internationally-oriented issues, like, biology, medicine, and hard sciences (physics &chemistry), male authors, in average, publish more than female ones. However, this trend changes for fields with more nationally-oriented interests namely, agriculture, and social sciences. In the case of engineering, similar profiles in productivity among males and females may represent compensation means form females to over passing the high competition that do occur in this field. But, an overall conclusion from this picture is that instead of crossing gender boundaries, some female authors seek to gain visibility in a less competitive arena. Other analysis, such as, *per capita* by years since PhD degree and *per capita* by the condition of being or not a research group leader are under way and will be presented at conference.

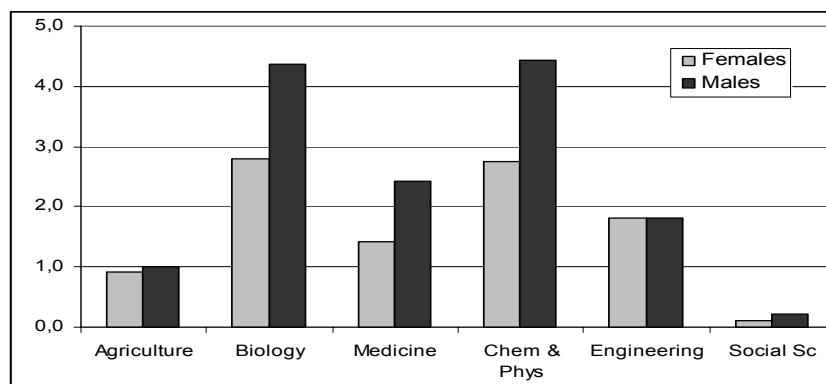


Figure 3. Publications *per capita* in international journals among Brazilian scientists, 2001-2004.

Conclusion

Although our data are consistent with the international literature on gender in science, we believe that there are other factors, other than gender, underlying the scientific productivity of the researchers included in this study. Our sample, which comes from Brazil, is likely to be affected by huge regional and institutional differences, whose importance should not be underestimated.

In order to reduce the possibilities of making a biased approach, we are focusing our analysis on the publishing trends of male and female scientists working in similar graduate programs. At present, the country has more than one thousand graduate programs. Every three years they are submitted to a national evaluation process and ranked from 1 to 7, with the highest score (7) achieved by those with the best research performance, especially in terms of publishing. We will now look at male and female productivity in programs with high and medium scores, and we expect this analysis may provide evidence of the extent to which gender can be correlated to academic performance in this particular community.

References

- Annan K (2003) A Challenge to the World's Scientists, *Science* 299 (7), pg 1485.
 Bourdieu P. Os usos sociais da ciência. Por uma sociologia clínica do campo científico. São Paulo: Editora UNESP, 2003.
 CNPq, National Council for Scientific and Technological Development. Available at: http://dgp.cnpq.br/censo2004/series_historicas/index_basicas.htm and <http://www.cnpq.br/areasconhecimento/index.htm>
 Creamer, EG (1995) The scholarly productivity of women academics. *Initiatives* 57(1): 1-9.
 Hermann C, Cyrot-Lackmann F (2002) Women in Science in France, *Science in Context* 15(4), 529–556
 IBGE (2004): Statistics of century XX – Table 5.1.5.2.4, Retrieved from: www.ibge.gov.br
 Lewison G (2001) The quantity and quality of female researchers: a bibliometric study of Iceland *Scientometrics* 52(1): 29-43
 Long, J S (1993) Measures of sex differences in scientific productivity. *The Scientist* 7(4): 3-14
 Pena MVJ, Correia MC, Van Bronkhorst B, Oliveira IR (2005) Questão de gênero no Brasil. Banco Mundial & CEPID: Brasília, Brasil.
 Rossi, A.S. (1965) Women in science: why so few? Social and psychological influences restrict women's choice and pursuit of careers in science. *Science*, n. 148, p. 1196-1202.
 Schiebinger L (1999) *Has Feminism Changed Science?* Cambridge: Harvard University Press. Foreign Translations: Portuguese (Editora da Universidade do Sagrado Coração, 2001)
 Webster, BM (2001) Polish women in science: a bibliometric analysis of Polish science and its publications, 1980-1999 *Research Evaluation* 10 (3): 185-194.

Atypical Citation Patterns in the Twenty Most Highly Cited Documents in Library and Information Science

Jonathan Levitt and Mike Thelwall

j.m.levitt@wlv.ac.uk, m.thelwall@wlv.ac.uk

Statistical Cybermetrics Research Group, School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB (UK)

Abstract

We investigate the citations to the twenty most highly cited documents in the Web of Science subject category of ‘Information Science & Library Science’, finding that not only are most of these papers quite old but, surprisingly, for most their citation count is continually increasing, and for the rest it is undulating but not steadily decreasing.

Keywords

citation analysis; highly cited documents

Introduction

One of the underlying assumptions in research policy is that the level of citation of an article is an indication of its quality. It is important to quantify the extent to which other factors besides perceived quality could affect levels of citation, however. Patterns of annual citation may shed light on this issue, and we examine annual citation patterns of articles perceived to have had particular influence, namely the most highly cited.

Two recognised anomalous citation patterns that could create highly cited documents are ‘delayed recognition’ and ‘sleeping beauty’. In informetrics a document has delayed recognition if its level of citation rises substantially many years after publication, and a delayed recognition document is a sleeping beauty if the extent of the delayed recognition is particularly marked. Garfield (1980), Glanzel, Schlemmer and Thijs (2003), Glanzel and Garfield (2005) examine delayed recognition and Van Raan (2004), Van Dalen and Henkens (2005) and Burrell (2005) examine sleeping beauties. Nevertheless, the typical citation pattern is for the annual number of citations of an article to reach a peak about five years after the article is published and then to gradually decline. This pilot study examines two questions.

- Do the citation patterns of very highly cited documents differ from typical citations patterns?
- Do very highly cited documents display similar citations patterns to each other?

Methods

We used the Thompson Scientific (formerly: Institute for Scientific Information, ISI) citation database for the raw data, and the Web of Science (WoS) interface. We used the ISI Information Science & Library Science (IS&LS) category as our scope. We completed the following steps to obtain the complete set of all ISI publications that contain one or more IS&LS documents.

Since the WoS contains over 36 million documents and allows at most 100,000 documents to be processed in a single search, it was not possible to check all documents at once. Hence our search had to be split into many different searches and the theoretical minimum number of search sequences to accomplish the task exceeded 360. We conducted the searches as follows. The publications with documents in IS&LS were identified from the Social Science or Arts and Humanities Citation Indexes (113 searches) and from the Science Citation Index (542 searches). Each search was a wildcard document search, restricted to IS&LS, excluding journals already identified, and restricted to journals starting with a given letter of the alphabet (e.g., A*). The results were then scanned for new journal titles. The total number of search sequences was over 650, and the sequences on average processed 55,000 documents. Journals starting with numbers were also checked.

We used only four search sequences to identify the 20 most highly cited document in IS&LS. The reason for this small number of search sequences the Advanced Search ability to search a Boolean combination of up to fifty publications in a single search. We selected these Boolean combinations in such that the combinations collectively covered every publication in IS&LS and no combination contained more than 100,000 documents. For each of these four searches the most highly cited documents were identified by using the facility to rank results by 'Times cited', and details of the highest ranking documents pasted into Word. Finally, these results were collated and the twenty documents with the highest number of citations identified.

Results

Table 1 lists our sample, some of which are more computer science than information science.

Table 1. The twenty most highly cited documents in IS&LS (24 Nov. 2006).

Title	Citations
<i>Visual pattern recognition by moment invariants</i>	871
<i>Perceived usefulness, perceived ease of use, and user acceptance of information technology</i>	744
<i>Rough sets</i>	740
<i>A translation approach to portable ontology specifications</i>	570
<i>Indexing by latent semantic analysis</i>	560
<i>An algorithm for suffix stripping</i>	508
<i>Term-weighting approaches in automatic text retrieval</i>	437
<i>Quantizing for minimum distortion</i>	436
<i>Relevance weighting of search terms</i>	366
<i>Cocitation in scientific literature - new measure of relationship between 2 documents</i>	325
<i>Low-density parity-check codes</i>	263
<i>Understanding information technology usage - a test of competing models</i>	259
<i>The measurement of end-user computing satisfaction</i>	254
<i>Information needs and uses</i>	248
<i>Ask for information-retrieval: Part 1: Background and theory</i>	248
<i>General theory of bibliometric and other cumulative advantage processes</i>	241
<i>Question-negotiation and information seeking in libraries</i>	240
<i>Improving retrieval performance by relevance feedback</i>	239
<i>Relevance - review of and a framework for thinking on the notion in information science</i>	239
<i>Perceived usefulness, ease of use, and usage of information technology - a replication</i>	225

Table 2 presents the pattern of annual citations of the ten most highly cited documents in IS&LS. The documents are numbered in the same order as in Table 1, and 'Average to 1995' denotes the total number of citations to 1995 divided by the number of full years between publication and the end of 1995.

Table 2 answers Question 1 for the 10 most highly cited: The patterns of citation of 9 of the 10 documents differ from the typical pattern for citations; all apart from document 8 have not demonstrated any decline in the level of citation as late as 26 years after their average year of publication. Document 3 is an example of delayed recognition; its average annual number of citations for 2003 to 2005 exceeds its total number of citations from publication (in 1982) to 1995.

Table 3 answers Question 1 for the 10 next most highly cited. The patterns of citation of the documents differ from the typical pattern for citations in two ways: The trend for documents 11 and 12

are strongly upwards, and the trend for the remaining documents is undulating with periods of both rise and fall. Document 11 could be regarded as a sleeping beauty, in that it was cited 128 times during 2003 to 2005, but cited only once during 1962 to 1989.

Table 2. Pattern of annual citations of the ten most highly cited documents.

Document	1	2	3	4	5	6	7	8	9	10
<i>Year published</i>	1962	1989	1982	1993	1990	1980	1960	1988	1976	1973
2005	69	127	136	116	97	97	83	5	25	16
2004	58	105	87	92	81	67	55	9	26	15
2003	53	74	86	66	71	83	52	9	25	13
2002	50	50	61	40	53	40	38	5	23	12
2001	33	49	43	33	26	24	26	9	13	17
2000	45	27	42	28	38	31	24	8	16	12
1999	41	38	37	34	22	14	13	10	13	14
1998	46	29	23	31	29	11	11	11	10	7
1997	35	14	35	19	13	8	12	15	11	4
1996	40	16	20	10	9	11	12	10	13	9
Average to 1995	11	13	7	9	6	4	10	8	9	9

Table 3. Pattern of annual citations of the ten next most highly cited documents.

Document	11	12	13	14	15	16	17	18	19	20
<i>Year published</i>	1962	1995	1988	1986	1982	1976	1968	1990	1975	1992
2005	56	53	23	14	21	27	12	22	12	30
2004	38	27	19	18	14	11	13	23	8	17
2003	34	44	12	15	13	11	17	28	7	21
2002	18	20	18	13	7	7	3	20	18	26
2001	24	17	16	9	14	8	9	20	12	21
2000	16	14	23	5	9	5	9	15	7	18
1999	11	12	16	26	16	4	12	11	6	17
1998	3	8	14	15	17	4	10	11	14	15
1997	3	6	19	13	10	3	12	14	14	16
1996	4	3	21	9	7	7	7	14	12	12
Average to 1995	0	N/A	8	11	8	7	4	9	6	5

Why do some of the articles seem to have an enduring value for research? For example, “An algorithm for suffix stripping” describes a technique for removing the endings of words to aid information retrieval. Its high continuing citation count reflects two factors; (1) no better algorithm has yet been found, and (2) due to the increase in computing, information retrieval and natural language processing research, increasingly many papers are being written that describe systems using suffix-stripping. Similarly, the continuing citations accruing for the seminal “Cocitation in scientific literature” reflect that the technique introduced in this paper is used increasingly today.

We find it surprising that the levels of citation of 11 of the 20 most highly cited have risen sharply in the last few years. However, that these 11 are within the top 12 is probably spurious, as the 11th highest cited has only 17% more citations than the 20th highest cited.

Conclusion

Atypical citation patterns are important as they may add to our understanding of what citation connotes. This investigation found two atypical citation patters: A strong upward trend in number of citations (present in eleven of the twelve most highly cited) and an undulating citation pattern (present in documents thirteen to twenty).

Hirsch (2005) proposed the h-index as a measure of citation level; the h-index for a set of documents can be defined as the largest integer (h) for which the h^{th} most cited document is the set is cited at least h times. We evaluated the h-index for IS&IL (h_1) as 107; and found that in our sample the marked upward trend in citation level was only present in documents cited at least 2.4 h_1 times. We will be conducting other investigations to establish whether this finding applies to other subject areas.

These findings are based on the study of only 20 documents in one subject category, and may not be typical for other categories. The increasing use of computing has strongly influenced the subject, probably more significantly than many others. The results for fields like physics, history and sociology might show different patterns, perhaps including many articles that were parts of ‘trendy’ research fields that are no longer current, or for scientists that have fallen somewhat out of favour (e.g., Marx). Nevertheless, it is clear that the study of the citation patterns of highly cited articles is potentially fruitful for increasing understanding of the reasons for citation and for setting citation windows in citation analysis.

It is interesting that the methods used in this study here are made possible for casual researchers by the Web of Science interface, which is another illustration of the extra power to understand science that can be delivered by modern computing power. The method described above for identifying all documents in a subject category can in principle be used to identify the records in any subject category on Web of Science. Identifying all records in a subject category is useful not only in studies such as this, but also for obtaining norms for subject categories.

References

- Burrell, Q.L. (2005). Are "sleeping beauties" to be expected? *Scientometrics* 65(3), 381-389.
- Garfield, E. (1980). Premature discovery or delayed recognition - Why? *Current Contents* 26, 5-10.
- Glanzel, W. & Garfield, E. (2005). The myth of delayed recognition. *The Scientist* 18(11), 8.
- Glanzel, W., Schlemmer, B. & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, 58(3), 571-586.
- Hirsch J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America* 102(46), 16569-16572.
- Van Dalen, H.P. & Henkens, K. (2005). Signals in science - On the importance of signaling in gaining attention in science. *Scientometrics* 64(2), 209-233.
- Van Raan, A.F.J. (2004). Sleeping Beauties in science. *Scientometrics* 59(3), 467-472.

The References on UK Cancer Clinical Guidelines¹

Grant Lewison*,**

*glewisonxx@aol.com

Evaluametrics Ltd, 50 Marksbury Avenue, Kew, Richmond, Surrey, TW9 4JF (England)

**grant.lewison@ucl.ac.uk

School of Library, Archive & Information Studies, University College London, Gower St, London WC1E 6BT, (England)

Abstract

One way in which biomedical research can be put into practice is through clinical guidelines, which are increasingly used to help doctors choose the most effective treatment for their patients. There has been a substantially increased interest in them recently, and in the UK there are three series. In cancer, there are 43 guidelines published to date, each of which has an evidence base in the form of references, many of which are papers in peer-reviewed journals. These have all been identified and analysed to determine their geographical provenance and type of research, in comparison with oncology research overall published in the peak years of guideline references (1999-2001). UK papers, as expected, were over-cited by a factor of nearly three: the cited papers acknowledged much more explicit funding from all sectors than did UK cancer research papers at the same research level. The references can also be used to identify cities in other countries whose research has influenced British cancer care.

Keywords

cancer; guidelines; treatment; evidence; geography; funding

Introduction: clinical guidelines and their evidence base

It is increasingly being recognised that the quantitative evaluation of biomedical research cannot depend only on the counting of citations in the serial literature. They may measure academic influence, but the funders of such research are usually more concerned to see if it has had a practical benefit, especially to patients. One of the ways in which research can influence practice is through its contribution to the evidence base supporting clinical guidelines (Heffner, 1998; Gralla *et al.*, 1999; Connis *et al.*, 2000; Van Wersch and Eccles, 2001; Aldrich *et al.*, 2003). These are increasingly being used to inform physicians and surgeons on which treatments are, and which are not, effective for the treatment of a particular condition. Most of them are published by national professional medical associations (e.g., Rizzo *et al.*, 2002; Atwood *et al.*, 2004; Makuuchi and Kokudo, 2006), but some are developed by governmental bodies (e.g., Pogach *et al.*, 2004).

It is normal for such guidelines to have lists of references that comprise their evidence base. However the quality of the evidence is sometimes doubtful (Ackman *et al.*, 2000; Watine, 2002; Burgers and van Everdingen, 2004) and schemes have been devised to grade the quality of the clinical trials which form a large part of the evidence base (e.g., Psaty *et al.*, 2000; Liberati *et al.*, 2001; Michaels and Booth, 2001; Hess, 2003; Guyatt *et al.*, 2006). Even when the guidelines have been published they are sometimes criticised as inadequate (Jacobson, 1998; Norheim, 1999; Walker, 2001), insufficient (Toman *et al.*, 2001) or they may become outdated (Shekelle *et al.*, 2001). There is also the question of whether the guidelines will actually be followed in clinical practice (Grol, 2001; Butzlaff *et al.*, 2002; Bonetti *et al.*, 2003; Bloom *et al.*, 2004).

A further cause of disagreement is the question of cost: a new drug may be clinically effective, and better than existing drugs or a placebo, but so costly that an equivalent or greater health gain may be achievable by other means, e.g., better screening to detect the disease at an early stage. This can cause

¹ This work was supported by Cancer Research UK; the Chief Scientist Office, Scottish Government; the Medical Research Council; and the Wellcome Trust. The MS Excel macros used to perform the geographical analysis of the papers were written by Dr Philip Roe. The identification of the individual papers cited on the clinical guidelines was carried out by Isla Rippon and Vicky Friedlander

considerable dissension and lead to lawsuits to make the drug available for particularly articulate patients (Dyer, 2006a), or from companies and patients' advocacy groups which sometimes receive their subsidies (Dyer, 2006b). Lobbying of the UK National Institute for Health and Clinical Excellence (NICE) by pharmaceutical firms is now rife (Ferner and McDowell, 2006) and a US politician has adopted bully-boy tactics in his efforts to subvert evidence-based medicine (Kmietowicz, 2006).

Despite all these criticisms, clinical guidelines are nevertheless gaining increasing recognition as the way forward. It does therefore seem worthwhile to treat them as an outcome indicator, albeit a partial one, of the clinical impact of the research they cite. Several studies have analysed the evidence base of selected clinical guidelines (Grant, 1999; Grant *et al.*, 2000; Lewison and Wilcox-Jay, 2003). They have established that the papers cited are very clinical (when positioned on a scale from clinical observation to basic research); that UK guidelines over-cite UK research papers; and that the cited papers are quite recent, with a temporal distribution comparable to that of the papers cited on biomedical research papers. Research from other European countries seems to be cited about as much as would be expected on UK clinical guidelines, but that from Japan and from most developing countries is almost totally ignored.

In this study, we examined three sets of UK guidelines on a single subject, cancer, and the references on 43 different guidelines, almost all concerned with treatment rather than with prevention. The bibliographic details of the references were assembled in a file and compared with those of cancer research publications in the three peak years (1999-2001). The objective was to answer several policy-related questions:

- how do countries' relative presences among the cited references compare with their presences in cancer research?
- how many of the cited references are actually classifiable as cancer research?
- what is the research level distribution of these cited references compared with that of cancer research papers?
- are the cited references published in journals of high citation impact?
- how does the funding of the cited papers compare with that of cancer research overall?

The latter two questions need to take account of the finding that the references on clinical guidelines are much more clinical than other biomedical research..

Methodology: UK cancer guidelines and the analysis of their references

There are three sets of clinical guidelines in common use in the UK:

- Published by the British Medical Association in *Clinical Evidence*. This takes the form of a book which is revised and extended every six months, but is also accessible on the Web (to people in the UK);
- Developed by the National Institute for Health and Clinical Excellence (NICE) for the National Health Service (NHS) in England and Wales, based on Health Technology Assessments. Most of these are available on the Web, but not all (though it is intended by NICE that they should be);
- Developed by the Scottish Intercollegiate Guidelines Network (SIGN) for use by the NHS in Scotland. All these are freely available on the Web.

Only a minority of these guidelines are applicable to cancer. The numbers are, respectively, 15, 18 and 10. Each of these 43 guidelines has a set of references, most of which are articles in peer-reviewed journals. A total of 3217 references were found and their details downloaded to file. Their addresses were parsed by means of a special macro so that the integer and fractional counts of each country were listed for each paper. The research level of each paper was determined using the new system developed by Lewison and Paraje (2004), in which each journal is assigned a research level based on the presence of "clinical" and "basic" words in the titles of papers it has published on a scale from 1 = clinical to 4 = basic. In addition, the research level of groups of individual cited papers could be calculated with reference to their individual titles, and the presence of "clinical" or "basic" words

within them. The potential citation impact of each cited paper was also determined with reference to a file of Journal Expected Citation Rates provided by Thomson Scientific. This gave the mean number of citations for papers published in a journal in a given year and cited in the year of publication and the four subsequent years.

Funding data for virtually all the UK papers (790 out of 796) were obtained from inspection of the acknowledgements to their funding sources in the British Library. Many of the papers had previously been looked up for the Research Outputs Database (Webster *et al.*, 2003) or for other projects, and only 151 needed to be sought anew. The main comparator used to normalise the results of the analysis of the cited references was a file of world oncology research papers (Cambrosio *et al.*, 2006). For the years 1999-2001, there were over 100,000 such papers, and their characteristics were used to see how the cited references compared with them, with due account being taken of the differences expected in mean research levels (the cited references being more clinical than oncology papers overall).

Results

Figure 1 shows the distribution of the 3217 cited references by publication date. There is a clear peak in the year 2000, and 31% of all the references were published in the three years, 1999-2001, so this was the time period used for many of the comparisons with world oncology research.

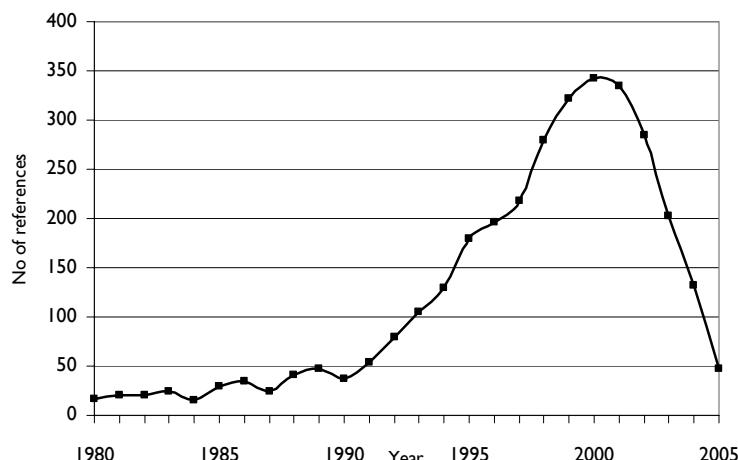


Figure 1. Temporal distribution of the 3217 references on UK cancer clinical guidelines

Of the references classed as “articles” or “reviews”, 88% were within the sub-field of oncology as defined by Cancer Research UK (Cambrosio *et al.*, 2006). This percentage remained sensibly constant over the period, 1994-2004. However, the references were in much more clinical journals than world oncology papers for the year 2000, the peak year for the numbers of references, see Figure 2. This result was obtained earlier (Grant, 2000; Lewison & Wilcox-Jay, 2003) but with a much simplified (and less accurate) method of categorisation of journals by research level. Of the 3217 papers, 2747 titles (86%) had either a “clinical” or a “basic” keyword, and the mean research level was 1.07, which is very close to the lower end of the scale (RL = 1.0), and much below the mean RL based on all the papers in the individual journals (RL = 1.43). This shows that the references were being published in journals that were relatively more basic than the papers themselves, and reinforces the message that the papers were therefore almost entirely clinical observation.

A geographical analysis of the presence of 20 leading countries in oncology research for 2000 and in the references from the clinical guidelines gave the results shown in Table 1, where the data have been shown on a fractional count basis. Figure 3 presents the ratio between a country’s presence in the guideline references and its presence in oncology research, i.e., the values shown in the last column of Table 1. As would be expected, the UK oncology research is over-cited, by a factor of almost 3, but several other European countries’ work is also relatively over-cited, notably that of Denmark, Ireland and Sweden. Although Italy, which is strong in clinical trials, shows to advantage, Germany is

relatively much under-cited compared with its presence in cancer research in recent years. Japanese work is almost ignored, but it is likely that the Science Citation Index, where most of the references were found, does not cover Japanese clinical journals. This, however, is only a small part of the reason for the paucity of Japanese references.

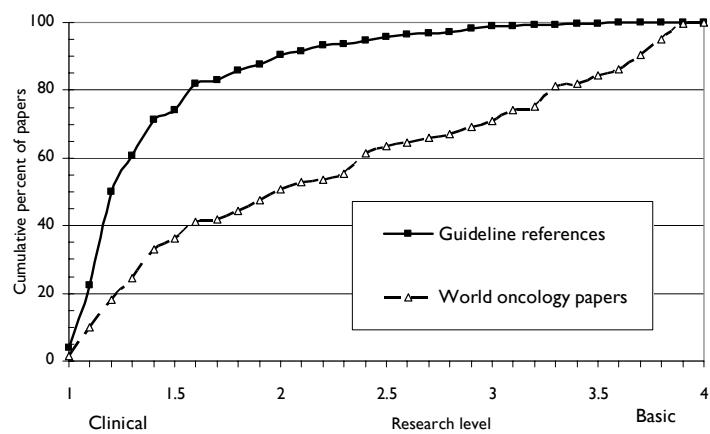


Figure 2. Research level distributions (cumulative percentages) for references on cancer clinical guidelines (solid squares) and for oncology research in 2000 (open triangles)

Table 1. The fractional count outputs and percentage presences of 20 countries in oncology research in 2000, their presence in the references on 43 UK cancer clinical guidelines, and the ratio of the two percentages.

Country	ISO	ONCOL, fr	ONCOL fr %	G refs, fr	G refs, %	Ratio
Australia	AU	552.2	1.54	94.0	2.97	1.93
Austria	AT	401.7	1.12	24.6	0.78	0.69
Belgium	BE	353.4	0.99	47.2	1.49	1.51
Canada	CA	1056.3	2.95	143.2	4.52	1.53
Switzerland	CH	409.8	1.14	32.7	1.03	0.90
Germany	DE	2735.6	7.64	133.2	4.21	0.55
Denmark	DK	256.3	0.72	45.2	1.43	1.99
Spain	ES	646.4	1.80	45.6	1.44	0.80
Finland	FI	316.7	0.88	25.4	0.80	0.91
France	FR	1749.1	4.88	197.9	6.25	1.28
Greece	GR	270.2	0.75	24.8	0.78	1.04
Ireland	IE	69.9	0.20	10.8	0.34	1.75
Italy	IT	1939.5	5.41	259.3	8.19	1.51
Japan	JP	4600.6	12.84	66.9	2.11	0.16
Netherlands	NL	953.2	2.66	105.9	3.35	1.26
Norway	NO	188.2	0.53	19.5	0.62	1.17
Portugal	PT	42.3	0.12	1.9	0.06	0.51
Sweden	SE	627.0	1.75	90.1	2.85	1.63
United Kingdom	UK	2331.8	6.51	604.6	19.1	2.93
United States	US	12428.1	34.70	1067.8	33.7	0.97

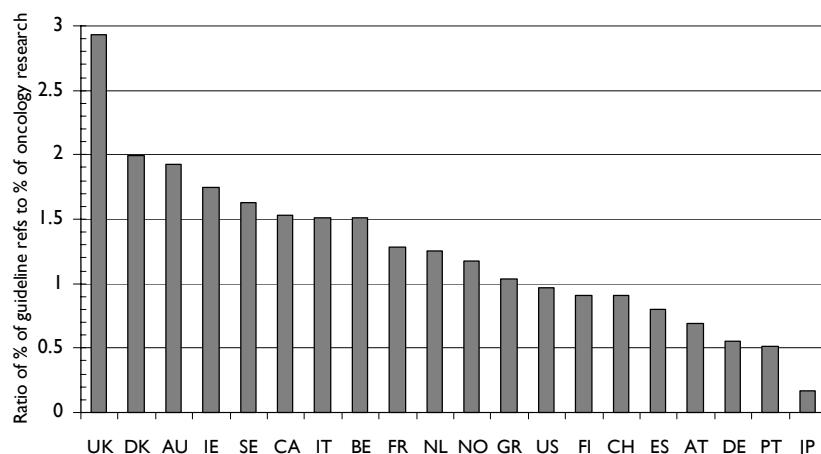


Figure 3. Ratio of countries' presence among UK cancer clinical guideline references and their presence in world oncology research, 2000: fractional counts

Table 1 and Figure 3 show overall values, but an analysis can also be made of subsets of papers for groups of two or three years, chosen so that the four periods each have about 20% of the total cited references, see Table 2. For nearly all the countries, there are close similarities between the time trends, which suggest that the guidelines are rather consistent in the geography of their citing behaviour. Thus Australia, Canada, Sweden, the UK and the USA have all shown a reducing presence in oncology research, and a reducing presence in the guideline references; Germany, on the other hand, has increased its presence in both (but is still much under-cited). France and Japan increased their presence in both sets of papers but it went down slightly during the latest period.

Table 2. Variation in time of the percentage presences of 10 leading countries in both UK guideline references and world oncology research; fractional counts.

Period	1995-7	1998-9	2000-1	2002-5	1995-7	1998-9	2000-1	2002-5
	Guideline references				World oncology research			
AU	3.2	3.2	2.4	2.5	1.7	1.6	1.6	1.6
CA	5.3	4.7	4.5	4.0	3.0	3.0	2.9	2.8
DE	4.1	4.8	4.4	5.0	6.8	7.3	7.5	7.5
FR	5.9	7.4	6.6	6.0	5.1	5.2	4.7	4.6
IT	8.2	9.2	10.0	7.8	5.6	5.3	5.6	5.6
JP	1.7	2.6	2.7	2.5	11.2	12.6	12.4	11.6
NL	3.5	3.2	4.2	3.7	2.8	2.6	2.6	2.6
SE	3.4	2.3	3.1	3.1	2.1	1.8	1.8	1.6
UK	22.0	15.5	17.5	17.8	7.4	6.7	6.4	6.0
US	31.5	31.7	27.9	29.9	36.6	34.9	34.7	34.8

The references cited tend to be published in high-impact journals. Table 3 shows that in each research level grouping, the guideline references are published in journals with a higher mean citation score (the potential citation impact, PCI, of the papers) than world oncology papers from the year 2000.

The overall mean is higher, too, at 19.9 cites in five years compared with 13.4. The “superior performance” of the guideline references occurs because a large number of them are published in the high-impact general journals, *Lancet* (138 of them), *New England Journal of Medicine* (133), *BMJ* (78) and *Journal of the American Medical Association* (50).

Table 3. Mean potential citation impact (PCI = expected cites in 5 year window) for guideline references (g. refs) and world oncology papers for 2000 (oncology).

	RL	N g. refs	PCI g. refs	N oncology	PCI oncology
<i>Clinical</i>	1 - 1.5	2316	21.5	12465	9.6
	1.5 - 2	511	14.3	4958	10.2
	2 - 2.5	217	12.1	4747	10.0
	2.5 - 3	114	23.5	2941	14.6
	3 - 3.5	38	24.8	4976	18.9
<i>Basic</i>	3.5 - 4	12	51.9	5944	21.6

We turn now to the analysis of funding of the UK cited references. Of the 796 UK papers, all but 6 were found and inspected in order to determine their funding sources. These were taken both from the addresses (as for some organizations this is an indication of funding) and from the formal acknowledgements. For the purposes of this analysis, funding sources were grouped into five main sectors:

- UK government, both departments and agencies;
- UK private-non-profit, including collecting charities, endowed foundations, hospital trustees, mixed (academic), and other non-profit. A subset of this sector is Cancer Research UK, and its two predecessors, the Cancer Research Campaign and the Imperial Cancer Research Fund;
- pharmaceutical industry, both domestic and foreign (it is often difficult to distinguish as some subsidiaries have considerable autonomy in the use of research funds), and including biotech companies;
- non-pharma industry;
- no funding acknowledged.

The remaining funding organizations are foreign governmental and private-non-profit sources, and international organizations, such as the European Commission (EC) and the World Health Organization (WHO).

The funding sources vary with the research level (RL) of the papers: the more clinical papers have fewer sources and the more basic papers have more. Table 4 shows the analysis for UK papers in oncology in 1999-2001 and Table 5 shows the results for the UK papers cited on cancer clinical guidelines. For each RL group, an estimate has been made of the funding that would have been expected had they been typical of UK cancer research, and in the last row there are given the ratios of observed to expected numbers of papers (integer counts) on the assumption that the cancer clinical guideline citations are typical of oncology, but with due allowance for the different RL distributions.

For example, the UK oncology papers in the first group (RL from 1.0 to 1.5) have UK government funding on 11.1% of them, so it might be expected that there would be $0.111 \times 544 = 60.4$ government-funded papers among the corresponding group cited on cancer clinical guidelines. In fact there were 149 such papers, showing that many more are government-funded than might have been expected. When the totals for each of the six groups are added, it can be seen that the observed number of UK government-funded papers is almost twice the predicted number. The observed total is still higher ($\times 2.5$) for the pharma-industry funded papers, and a little lower for Cancer Research UK papers ($\times 1.8$), for non-pharma industry papers ($\times 1.6$) and UK private-non-profit papers ($\times 1.3$). Not surprisingly, there are many fewer “unfunded” papers, the ratio of observed to expected numbers of papers being only just over half.

Discussion

UK cancer clinical guidelines are sufficient in number and variety to provide a fair window on the practical influence of cancer research, not only in the UK but in other leading countries, particularly in western Europe. We have seen that almost all the references (88%) are to papers that are within the sub-field of cancer research. Since about one third of the research supported by Cancer Research UK,

Table 4. Funding of UK oncology research papers in 1999-2001, grouped by RL (integer counts); mean annual totals. A status = inspected papers. GOV = UK government; PNP = UK private-non-profit; CRUK = Cancer Research UK.

RL: ONCOL	N(A)	% of A	GOV	GOV, %	PNP	PNP, %	CRUK	CRUK%
1 - 1.5	880	31.9	97.7	11.1	208.0	23.6	118.3	13.4
1.5 - 2	426	15.4	52.3	12.3	133.7	31.4	62.3	14.6
2 - 2.5	443	16.1	81.7	18.4	250.7	56.6	146.7	33.1
2.5 - 3	225	8.2	39.7	17.7	123.7	55.0	54.7	24.3
3 - 3.5	330	12.0	76.7	23.2	189.0	57.3	99.3	30.1
3.5 - 4	452	16.4	163.0	36.1	300.3	66.4	162.7	36.0
<i>Total</i>	2756	100.0	511.0	18.5	1205.3	43.7	644.0	23.4
RL: ONCOL	N(A)	% of A	Pharm	Ph'm, %	Ind'y	Ind'y, %	None	None %
1 - 1.5	880	31.9	53.3	6.1	25.3	2.9	526.7	59.8
1.5 - 2	426	15.4	39.0	9.2	17.0	4.0	199.7	46.9
2 - 2.5	443	16.1	71.0	16.0	19.7	4.4	95.7	21.6
2.5 - 3	225	8.2	22.3	9.9	7.0	3.1	48.3	21.5
3 - 3.5	330	12.0	43.3	13.1	16.7	5.1	42.7	12.9
3.5 - 4	452	16.4	65.3	14.5	20.0	4.4	37.0	8.2
<i>Total</i>	2756	100.0	294.3	10.7	105.7	3.8	950.0	34.5

Table 5. Funding of UK papers cited by cancer clinical guidelines, grouped by RL (integer counts). O = observed number of papers; C = calculated on basis of ONCOL papers.

RL: G refs	G refs	%	GOV-O	GOV-C	PNP-O	PNP-C	CRUK-O	CRUK-C
1 - 1.5	544	69.2	149	60.4	198	128.5	142	73.1
1.5 - 2	127	16.2	26	15.6	49	39.9	39	18.6
2 - 2.5	83	10.6	13	15.3	46	47.0	33	27.5
2.5 - 3	19	2.4	4	3.4	13	10.5	9	4.6
3 - 3.5	13	1.7	2	3.0	5	7.4	3	3.9
3.5 - 4	1	0.1	1	0.4	1	0.7	0	0.4
<i>Total</i>	787	100.1	195	98.0	312	233.9	226	128.1
<i>Obs/Calc</i>			1.99		1.33		1.76	
RL: G refs	G refs	%	Pharm O	Pharm-C	Indy-O	Indy-C	None-O	None-C
1 - 1.5	544	69.2	116	33.0	25	15.7	156	325.5
1.5 - 2	127	16.2	19	11.6	8	5.1	40	59.6
2 - 2.5	83	10.6	18	13.3	8	3.7	21	17.9
2.5 - 3	19	2.4	0	1.9	1	0.6	4	4.1
3 - 3.5	13	1.7	3	1.7	0	0.7	3	1.7
3.5 - 4	1	0.1	0	0.1	0	0.0	0	0.1
<i>Total</i>	787	100.1	156	61.6	42	25.7	224	408.8
<i>Obs/Calc</i>			2.53		1.63		0.55	

in common with that of other medical research charities working in a particular disease area, is outwith this sub-field [most of this would comprise basic biology], it follows that little of this work can be expected to influence clinical guidelines – hardly a surprising conclusion, but nevertheless one that is worth stating. Many of the guideline references are to papers in the US and UK general medical journals – *JAMA*, *N Engl J Med*, *BMJ* and *Lancet*. This is one reason, but by no means the only one, for the guideline references as a whole to be in high-impact, and therefore well known, journals. It appears that if researchers want their work, particularly clinical trials, to be part of the evidence base for clinical guidelines, then it is desirable for them to publish in highly visible journals. Disproportionately many of these papers will have been funded by government or the pharmaceutical industry, with charities also playing an enhanced role compared with cancer research overall.

Although this study was confined to the references on cancer guidelines from the UK, there are sufficient data to allow some conclusions to be drawn about cancer research in other countries. For example, in the three leading continental European countries, Germany, France and Italy, it is possible to compare the percentage presence of the leading cities within each country's output of oncology research (the year 2000 has been used) and its percentage presence on the guideline references to show which cities are making the greatest relative contribution to cancer treatment. Table 6 shows the data. In Germany, Munich and Essen show to advantage; in France, Villejuif and Lille, and in Italy, Milan and Florence. This may well reflect the type of research carried out in these cities.

Table 6. Leading cities in Germany, France and Italy with their presence (fractional counts) in oncology research, 2000, and in UK cancer guideline references.

	<i>Oncology</i>	<i>% oncol.</i>	<i>G refs</i>	<i>% g. refs</i>	<i>Ratio</i>
Germany	2736		127.1		
<i>Munich</i>	215	7.9	15.8	12.4	1.58
<i>Berlin</i>	266	9.7	14.0	11.0	1.13
<i>Essen</i>	83	3.0	8.7	6.8	2.25
<i>Tubingen</i>	102	3.7	7.0	5.5	1.48
France	1748		197.9		
<i>Paris</i>	390	22.3	39.1	19.8	0.88
<i>Villejuif</i>	132	7.5	20.1	10.2	1.35
<i>Lyon</i>	164	9.4	16.5	8.3	0.89
<i>Lille</i>	64	3.7	9.2	4.6	1.27
Italy	1940		259.3		
<i>Milan</i>	316	16.3	86.3	33.3	2.05
<i>Rome</i>	257	13.3	23.3	9.0	0.68
<i>Florence</i>	69	3.5	15.7	6.1	1.71
<i>Genoa</i>	126	6.5	14.1	5.4	0.84
<i>Naples</i>	128	6.6	13.1	5.1	0.77

Such conclusions are necessarily tentative, but would be reinforced if they were confirmed by parallel analyses of cancer clinical guidelines from other countries. It would be desirable for such work to be carried out in the relevant countries but for the results to be made more widely available so that research evaluators could have an increasingly rich resource for their work. It would also be helpful if there were agreement on the format of the databases that would be created, so that data from different countries could form a seamless whole. Such an activity could very well be co-ordinated by the ISSI, with all data contributors having also the right to gain access to the data provided by workers in other countries.

References

- Ackman ML, Druteika D and Tsuyuki RT (2000) Levels of evidence in cardiovascular clinical-practice guidelines. *Canadian Journal of Cardiology*, 16 (10), 1249-1254

- Aldrich R, Kemp L, Williams JS *et al.*, (2003) Using socioeconomic evidence in clinical-practice guidelines. *BMJ*, 327, 1283-1285.
- Atwood CW, McCrory D, Garcia JGN *et al.*, (2004) Pulmonary-artery hypertension and sleep-disordered breathing – ACCP evidence-based clinical-practice guidelines. *Chest*, 126 (1), S72-S77.
- Bloom BS, de Pouvorville N, Chhatre S *et al.*, (2004) Breast cancer treatment in clinical practice compared to best evidence and practice guidelines. *British Journal of Cancer*, 90 (1), 26-30.
- Bonetti D, Johnston M, Pitts NB *et al.*, (2003) Can psychological models bridge the gap between clinical guidelines and clinician behaviour – a randomized controlled-trial of an intervention to influence dentists' intention to implement evidence-based practice. *British Dental Journal*, 195 (7), 403-407.
- Burgers JS and van Everdingen JJE (2004) Beyond the evidence in clinical guidelines. *The Lancet*, 364, 392-393.
- Butzlaff M, Floer B, Koneczny N *et al.*, (2002) Assessment and utilization of clinical guidelines by primary care physicians and internists. *Zeitschrift für ärztliche Fortbildung und Qualitätssicherung*, 96 (2), 127-133.
- Cambrosio A, Keating P, Mercier S *et al.* (2006) Mapping the emergence and development of translational cancer research. *European Journal of Cancer*, 42, 3140-3148.
- Connis RT, Nickinovich DG, Caplan RA and Arens JF (2000) The development of evidence-based clinical practice guidelines – integrating medical science and practice. *International Journal of Technology Assessment in Health Care*, 16 (4), 1003-1012.
- Dyer C (2006a) Patient is to appeal High court ruling on breast cancer drug. *BMJ*, 332, 443.
- Dyer C (2006b) NICE faces legal challenge over restriction on dementia drug *BMJ*, 333, 1085.
- Ferner RE and McDowell SE (2006) How NICE may be outflanked. *BMJ*, 332, 1268-1271.
- Gralla RJ, Osoba D, Kris MG *et al.*, (1999) Recommendations for the use of anti-emetics: evidence-based, clinical practice guidelines. *Journal of Clinical Oncology*, 17 (9), 2971-2994.
- Grant J (1999) Evaluating the outcomes of biomedical research on healthcare. *Research Evaluation*, 8 (1), 33-38.
- Grant J, Cottrell R, Cluzeau F and Fawcett G (2000) Evaluating “payback” on biomedical research from papers cited in clinical guidelines – applied bibliometric study. *BMJ*, 320, 1107-1111.
- Grol R (2001) Successes and failures in the implementation of evidence-based guidelines for clinical practice. *Medical Care*, 39 (8), 1146-1154.
- Guyatt G, Guterman D, Baumann MH *et al.*, (2006) Grading strength of recommendations and quality of evidence in clinical guidelines. *Chest*, 129, 174-181.
- Heffner JE (1998) Does evidence-based medicine help the development of clinical practice guidelines? *Chest*, 113 (3), S172-S178.
- Hess DR (2003) Evidence-based clinical practice guidelines: where's the evidence and what do I do with it? *Respiratory Care*, 48 (9), 838-839.
- Jacobson JJ (1998) Evidence-based clinical practice guidelines: friend or foe? *Oral Surgery Oral Medicine Oral pathology Oral Radiology and Endodontics*, 86 (2), p 137.
- Kmietowicz Z (2006) Experts defend NICE against attack by US politician. *BMJ*, 333, p 1087.
- Lewison G and Wilcox-Jay K (2003) Getting biomedical research into practice: the citations from UK clinical guidelines. *Proceedings of the 9th International Conference on Scientometrics and Informetrics*, Beijing, China; 152-160.
- Lewison G and Paraje G (2004) The classification of biomedical journals by research level. *Scientometrics*, 60 (2), 145-157.
- Liberati A, Buszetti R, Grilli R *et al.*, (2001) Which guidelines can we trust: assessing strength of evidence behind recommendations for clinical practice. *Western Journal of Medicine*, 174 (4), 262-265.
- Michaels J and Booth A (2001) Pragmatic system for the grading of evidence and recommendations in clinical guidelines. *Journal of Clinical Excellence*, 3 (3), 139-145.
- Makuuchi M and Kokudo N (2006) Clinical practice guidelines for hepatocellular carcinoma: the first evidence-based guidelines from Japan. *World Journal of Gastroenterology*, 12 (5), 828-829. See English translation of guidelines at: www.jsh.or.jp
- Norheim OF (1999) Health-care rationing: are additional criteria needed for assessing evidence-based clinical practice guidelines? *BMJ*, 319, 1426-1429.
- Pogach LM, Brietzke SA, Cowan CL *et al.*, (2004) Development of evidence-based clinical practice guidelines for diabetes: the Department of Veterans Affairs/Department of Defense guidelines initiative. *Diabetes Care*, 27, B82-B89.
- Psaty BM, Furberg CD, Pahor M *et al.*, (2000) National guidelines, clinical trials and quality of evidence. *Archives of Internal Medicine*, 160 (19), 2577-2580.
- Rizzo JD, Lichtin AE, Woolf SH *et al.*, (2002) Use of Epoetin in patients with cancer: evidence-based clinical practice guidelines of the American Society of Clinical Oncology and the American Society of Hematology. *Journal of Clinical Oncology*, 20 (19), 4083-4107.

- Shekelle PG, Ortiz E, Rhodes S *et al.*, (2001) Validity of the Agency for Health-Care Research and Quality clinical practice guidelines: how quickly do guidelines become outdated? *Journal of the American Medical Association*, 286 (12), 1461-1467.
- Toman C, Harrison MB and Logan J (2001) Clinical practice guidelines: necessary but not sufficient for evidence-based patient education and counseling. *Patient Education and Counseling*, 43 (3), 279-287.
- Van Wersch A and Eccles M (2001) Involvement of consumers in the development of evidence-based clinical guidelines: practical experiences from the North of England Evidence-Based Guideline Development Programme. *Quality in Health Care*, 10 (1), 10-16.
- Walker T (2001) Evidence-based clinical guidelines: are they effective? *New Zealand Medical Journal*, 114, 20.
- Watine J (2002) Are laboratory investigations recommended in current medical practice guidelines supported by available evidence? *Clinical Chemistry and Laboratory Medicine*, 40 (3), 252-255.
- Webster BM, Lewison G and Rowlands I (2003) *Mapping the Landscape II: Biomedical Research in the UK, 1989-2002*. The City University, London, Department of Information Science: available at <http://www.ucl.ac.uk/ciber/MappingtheLandscape.php>

Is the United States Losing Ground in Science? A Global Perspective on the World Science System in 2005

Loet Leydesdorff* and Caroline Wagner**

*loet@leydesdorff.net

Amsterdam School of Communications Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, (The Netherlands)

**cswagner@gwu.edu,

SRI International, 1100 Wilson Boulevard, Arlington, VA, 22209 and George Washington University (USA)

Abstract

Based on the *Science Citation Index-Expanded* web-version, the USA is still by far the strongest nation in terms of scientific performance. Its relative decline in percentage share of publications is largely due to the emergence of China and other Asian nations. In terms of citations, the competitive advantage of the American “domestic market” is diminished, while the European Union (EU) is profiting more from the enlargement of the database over time than the US. However, the USA is still outperforming all other countries in terms of highly cited papers and citation/publication ratios, and it is more successful than the EU in coordinating its research efforts in strategic priority areas like nanotechnology. In this field, the People’s Republic of China (PRC) has become the second largest nation in 2005 in both numbers of papers published and citations behind the USA.

Keywords

national science; bibliometrics indicators; nanotechnology

Introduction

The last decade has witnessed significant changes in locations where science is conducted. Data show exponential growth in the share of the People’s Republic of China (PRC) for almost all science and technology indicators (Jin & Rousseau, 2004; Zhou & Leydesdorff, 2006). Kostoff (2004) argued that a number of indicators show the PRC outperforming the USA in 2004 in strategic areas like nanotechnology. Using a number of these indicators, Shelton & Holdridge (2004) issued a warning that the USA is losing to competition from an expanding European Union. In the meantime, serious concern has been voiced about whether the USA is declining not only in terms of relative shares (because of the zero-sum game) but also in terms of absolute numbers, for example, on the crucial indicator of performance in terms of scientific publications (NSB, 2006; Shelton, 2006).

In this study, we present recent data that are relevant to this debate. By placing the data for 2005 in the context of developments over the last ten years (King, 2004), we are able to show that the USA is still by far the leading nation in the world of science. The numeric lead of the EU-25, which is larger in size and population than the USA, cannot hide the endogenous problems of the EU science system. The rise of the PRC and other smaller Asian nations seems to continue almost predictably along exponential and linear curves, respectively, expanding scientific output rather than pushing out other performers.

Methods and data

Data are drawn from the expanded version of the *Science Citation Index* available on the Internet as part of the “Web of Science” (<http://portal.isiknowledge.com>). The publication counts are based on the three scientific document types which can be cited: research articles, reviews, and letters. We used the field tag “cu=” for countries and the delineations of years as provided by the database. Percentages of world share are based on attributing one full point to each country in the case of internationally co-authored papers. For this reason, world shares may add up to more than 100% (Braun *et al.*, 1991; Martin, 1991). The EU-25 series is corrected for “within-Europe” co-authorship.

The subset of nanoscience and technology (NSE) is based on a detailed study of the journal structure in this area (Leydesdorff & Zhou, forthcoming). Using measures from social network analysis

(Freeman, 1977, 1978/9; Leydesdorff, 2006), ten journals can be sampled as more specifically focused on nanoscience than their environments (Table 1). This analysis is based on the aggregated journal-journal citation data contained in the *Journal Citation Reports* of the *Science Citation Index* 2004. We use this set as a representation of the development of nanoscience and technology for comparative reasons, and add citation data. However, the specialty is not yet stabilized in a core set of journals; papers can be published in a large number of journals in different specialties (Braun *et al.*, 2007).

Table 1. Ten journals as representing nanoscience and nanotechnology as a specialty, on the basis of the *Journal Citation Reports* 2004.

Journal Title
<i>Advanced Materials</i>
<i>Chemical Physics Letters</i>
<i>Chemistry of Materials</i>
<i>Fullerenes, Nanotubes, and Carbon Nanostructures</i>
<i>Journal of Materials Chemistry</i>
<i>Journal of Nanoparticle Research</i>
<i>Journal of Nanoscience And Nanotechnology</i>
<i>Journal of Physical Chemistry B</i>
<i>Nano Letters</i>
<i>Nanotechnology</i>

Results

Figure 1 shows the percentage of world shares of scientific publications for the six major countries and the EU-25 contributing to science.

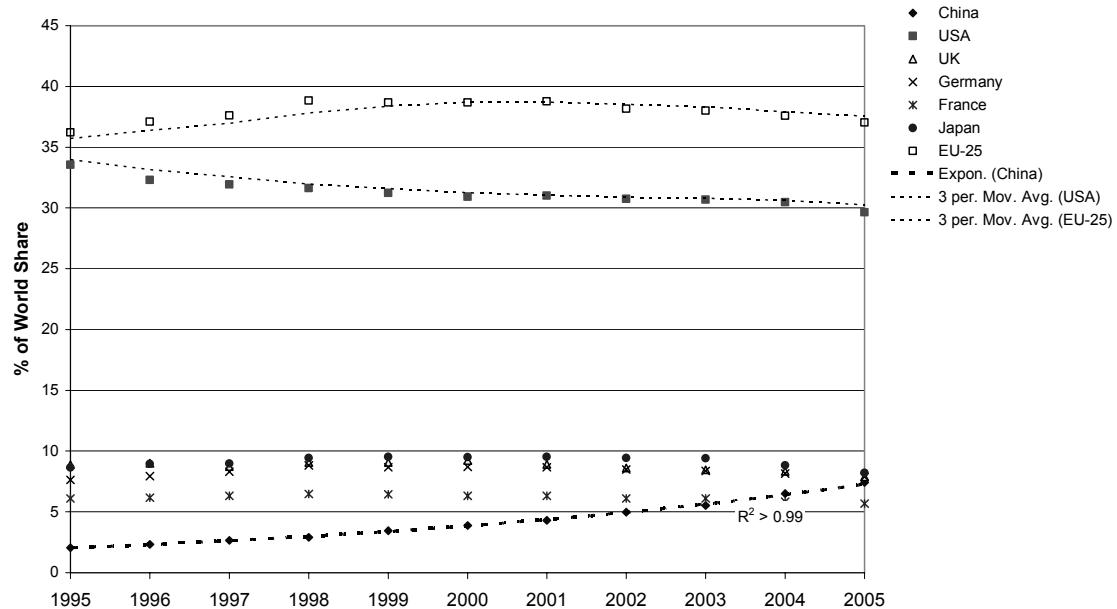


Figure 1. Percentages of world shares of publications held by six leading countries and the EU-25, 1995-2005.

The EU expanded from fifteen member states to twenty-five member-states in May 2004, thus the data are reconstructed from the perspective of hindsight. However, Zhou & Leydesdorff (2006) provided data for both the EU-15 and EU-25 until 2004. They conclude that the extension has changed the size of the EU, but not the trends in scientific productivity. Our data show that all the established countries are declining in relative shares. This decline is largely due to the spectacular increase of the percentage

share held by China that continues to follow an exponential curve ($r^2 > 0.99$). The three-year moving averages added to the lines for both the EU and USA show a similar pattern since 2000.

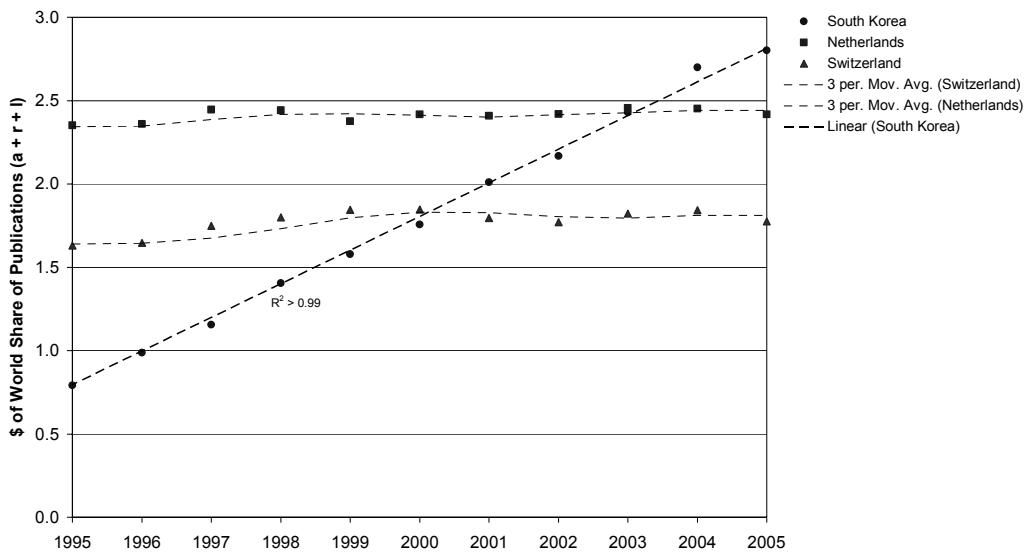


Figure 2. Percentages of world shares of publications held by smaller countries, 1995-2005.

Figure 2 extends the analysis to three smaller countries as examples. Unlike the larger countries, nations like Switzerland and the Netherlands have been able to stabilize their shares. This means that they keep pace with the increased competition, but are no longer able to improve their marginal return (as they did previously). South Korea, however, shows a steady increase. South Korea shares this linear pattern ($r^2 > 0.99$) with other “Asian Tigers” like Taiwan and Singapore (Zhou & Leydesdorff, 2006).

Note that China showed exponential growth (in Figure 1). This spectacular and hitherto sustained pattern of growth may be due to the increasing availability of human capital at Chinese universities and research institutions for publishing in ISI-listed journals, as well as to incentives within China to publish in refereed journals (Zhou & Leydesdorff, forthcoming).

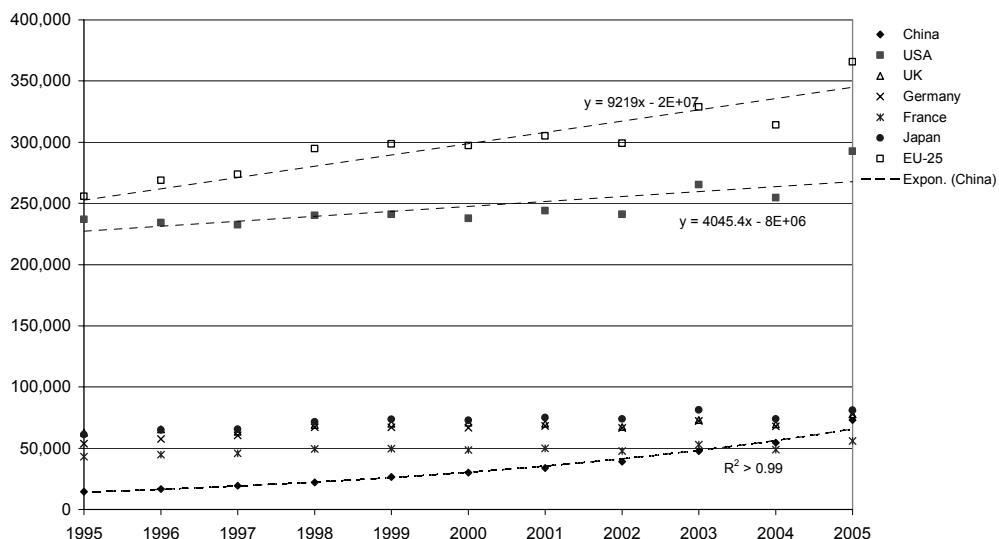


Figure 3. World shares of publications held by leading countries and the EU-25, 1995-2005.

Figure 3 shows absolute numbers of publications for the six leading countries and the EU-25. The size of the database itself varies from year to year, and this is reflected in the relative shares of major countries. The Chinese contribution again shows an exponential pattern in absolute numbers. The ten-year average of the increase for the EU-25 is more than twice as high as for the USA. The database itself is continuously expanding by an average of more than 20,000 citable items per year; the USA participates in this increase with approximately 19% and the EU with 43%, but this varies among years (Figure 4).

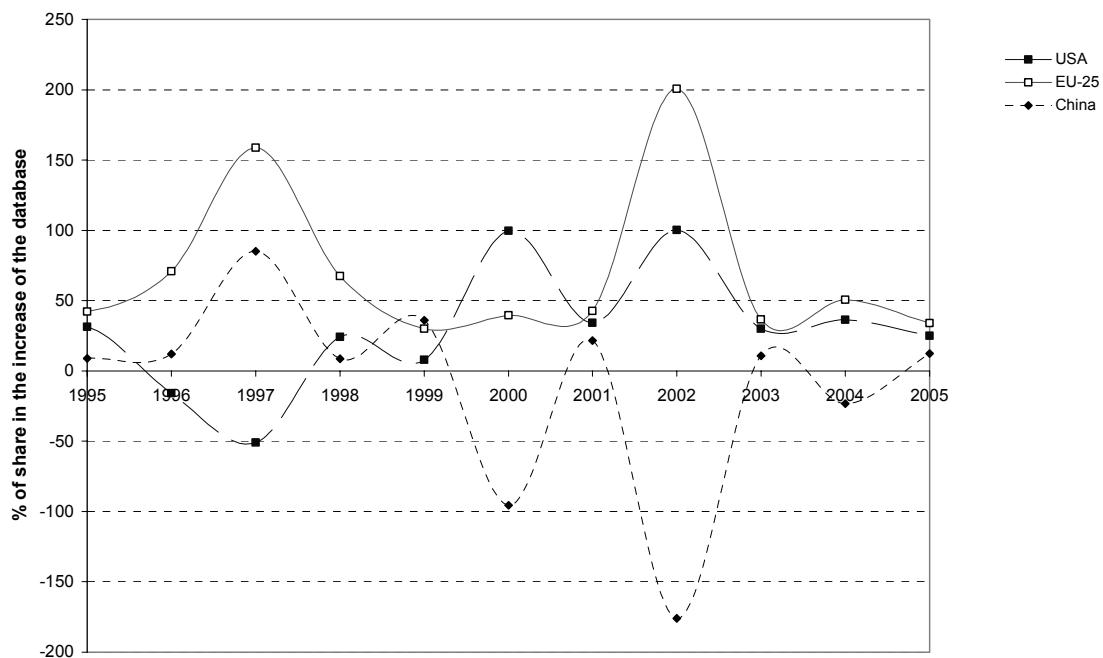


Figure 4. Percentage share of the EU-25, the USA, and the PRC in the increases of the *Science Citation Index-Expanded*, 1995-2005.

Figure 4 shows the participation of the EU-25, the USA, and the PRC in the expansion of the database as a relative percentage, for each two consecutive years. In almost all years, the EU-25 has improved its share of publications more than the USA. The pattern for China is more erratic, showing that mechanisms other than participation in the increase of the size of the database are driving China's performance figures. In other words, China does not profit from the extension of the database, while the EU does.

It is not easy to standardize the measurement of citations using the database online, because citations accumulate with time and the results are therefore dependent on the date of the data collection. However, on the basis of extensive research (Evidence, 2003), King (2004, at p. 312) provided a table of the 1% most highly cited papers in two periods, which allows us to compare the various nations at these two periods of times, that is, 1997-2001 and 1993-1997 (Figure 5).

The USA ranks second largest on this indicator during both periods (behind Switzerland), and like Switzerland it is nevertheless still able to improve its performance. The EU-15 is available in this data and it is highlighted on the map hovering close to Australia. In summary, the USA has been outperforming the EU on this quality indicator by a factor of almost two.

The *Science and Engineering Indicators 2006* of the National Science Board of the USA (NSB, 2006) are based on an analytical version of the ISI-data which has been maintained since 1988 by ipIQ, Inc. (formerly CHI Research, Inc.). In 2003 for example, this data, covers a selection of 5,315 journals from both the *Science Citation Index* and *Social Science Citation Index*. (The equivalent in the *SCI-*

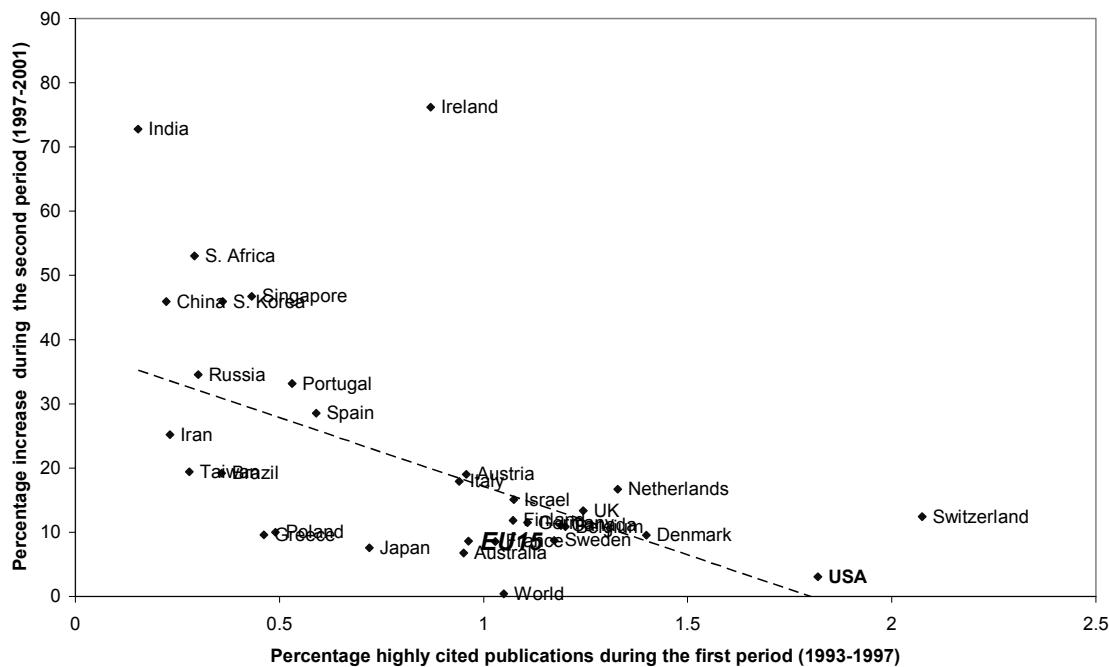


Figure 5. Percentage increase of most highly cited publications based on King (2004, at p. 312).
Source: Leydesdorff (2005, at p. 412; the linear regression line added for the orientation of the reader, n.s.).

Expanded would be 7,323 journals.)¹ Furthermore, the year 2003 is the last one available in this database. However, combining some of the tables from the report enables us to construct the following figure:

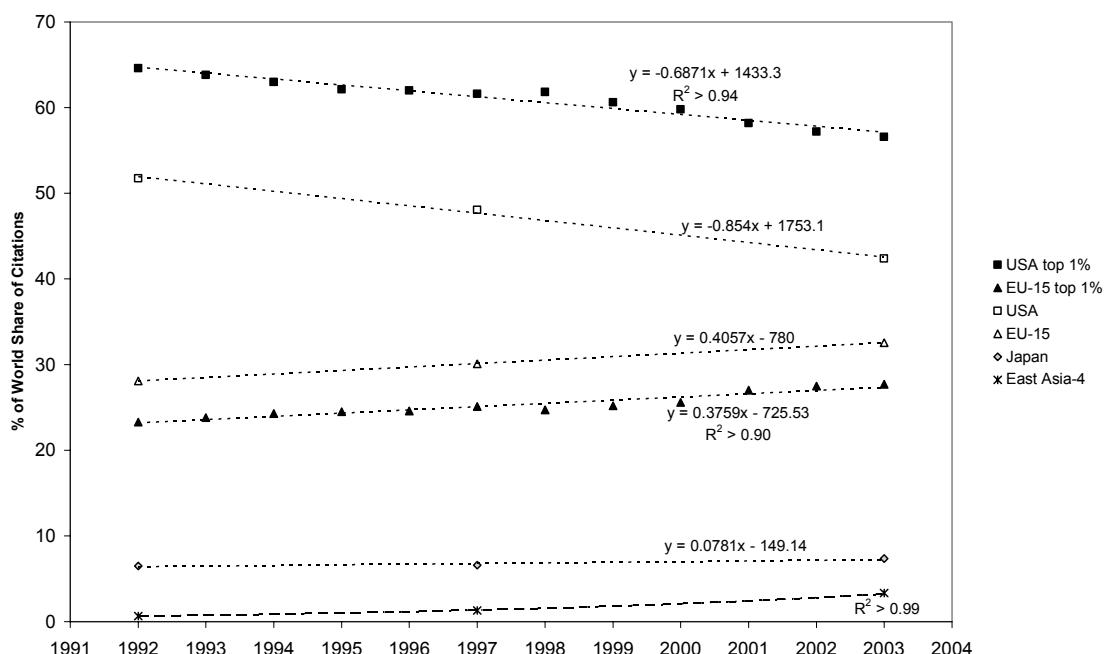


Figure 6. Development of the percentage of world shares of citations of the USA, the EU-15, Japan, and the East Asia-4 (China, South Korea, Singapore, and Taiwan). Source: NSB (2006, Tables 5-61 and 5-63).

¹ The *Science Citation Index* covered 5,714 journals in 2003, and the *Social Science Citation Index* 1,708. A subset of 99 journals is covered by both databases.

Citation data are available in this report only for 1992, 1997, and 2003, but for the top 1% of most highly cited papers, the in-between years are also provided. The curves show how the values for the EU-15 and the USA have grown together during the decade. The curves for the percentage of world share of 1% most highly cited papers are farther apart than the citation curves, and the slopes are reduced. Thus, the top segment is less affected by these changes than the average ones.

In a first approximation, one could assume that a large part of the 0.4% increase/year of the EU-15 is due to the expansion of the database with mainly European journals, but part of the decrease on the American side would still remain unexplained. Thus, the USA is losing ground in terms of number of citations, mainly because the average American paper is increasingly similar in this respect to the average European one. Note that the values for the East Asia-4 (China, South Korea, Singapore, and Taiwan) fit an exponential curve again with an r^2 larger than 0.99.

Nanoscience and nanotechnology

The delineation of an emerging field like nanoscience and nanotechnology in terms of a relevant journal set is not a *sine cure* given the interdisciplinarity (between chemistry and applied physics) of this subject area (Braun *et al.*, 1997; Meyer & Persson, 1998; Zitt & Bassecoulard, 2006; Mohrman & Wagner 2006; Braun *et al.*, 2007; Mogoutov & Kahane, 2007; Porter *et al.*, in preparation). The U.S. Patent and Trade Office (USPTO) decided in 2004 to introduce a new category (Class 977) into its classification scheme devoted to “nanotechnology.” Patents issued before this date have been reclassified. Similar efforts have been under way in the European and Japanese Patent Offices, and in international classification schemes (Scheu *et al.*, 2006).

In another context (Leydesdorff & Zhou, forthcoming), one of us analyzed this patent data in more detail. As could be expected, American inventors and assignees are overrepresented in the USPTO-database, while European ones are similarly overrepresented in the EPO-database. However, the USPTO database can also be considered as a window on the remainder of the world (Granstrand, 1999; Jaffe & Trajtenberg, 2002). From this perspective, the virtual absence of European patent holders in the “nano” category of this database is remarkable. Among the 1,027 (co-)inventors of the 336 patents classified as “nano” in 2005, 152 came from Japan, but only 33 from Germany. Other Asian nations are represented to a larger extent than European nations. A similar, but even more pronounced pattern can be made visible for the national distribution of the assignees. China is less active in patenting than Japan, Taiwan or South Korea.

Using “betweenness centrality” as a measure in the relevant networks of aggregated journal-journal citations (Freeman, 1977; Leydesdorff, 2006), we are able to follow the major players in the field of nanoscience as represented in a set of ten core journals (see Table 1 above). Using the same limitations on document types, Figure 7 can be constructed. We limited the period to 2002-2005 because major initiatives for establishing priority programs in this area were launched in 2000 and 2001. For example, under President Bill Clinton, the U.S. Government launched an initiative in 2000 to promote nanotechnology entitled the *National Nanotechnology Initiative: Leading to the Next Industrial Revolution*. The EU countries, China, Japan, and South Korea, have all adopted nanotechnology as an S&T policy priority. The Chinese government declared nanotechnology a critical R&D priority in their *Guidance for National Development* in 2001. The launch of new journals followed upon the increased funding (Leydesdorff *et al.*, 1994).

Within this set, the EU-25 is losing more than one percent of its world share of publications per year. The percentage of contributions with an American address has increased since 2003. Similarly, the PRC is gaining ground at the expense of Japan and the leading European nations (including Russia) (Mohrman & Wagner 2006). In 2005, the percentage of world share for China within this set is 13.1, outperforming Japan (12.4%) for the first time. However, the growth of China in this domain is not exponential, and the US growth is at least as strong. The real worry is the decline in the contribution of the EU nations. It seems to conform what one calls “the European paradox.” the EU is less able to use its research potentials strategically when compared with the USA (Dosi *et al.*, 2006).

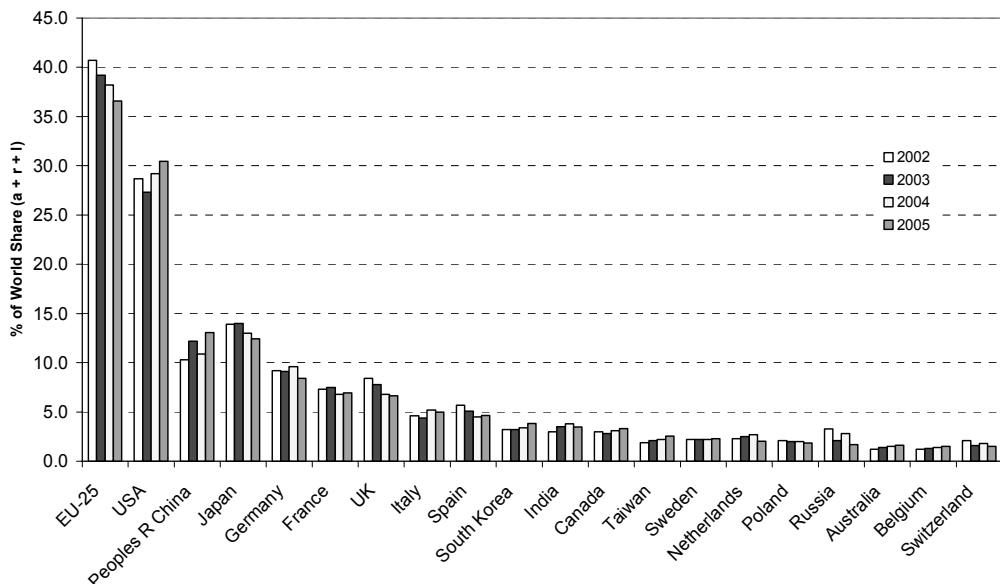


Figure 7. Percentage of world share of publications for leading countries and the EU-25 in nanoscience, 2002-2005.

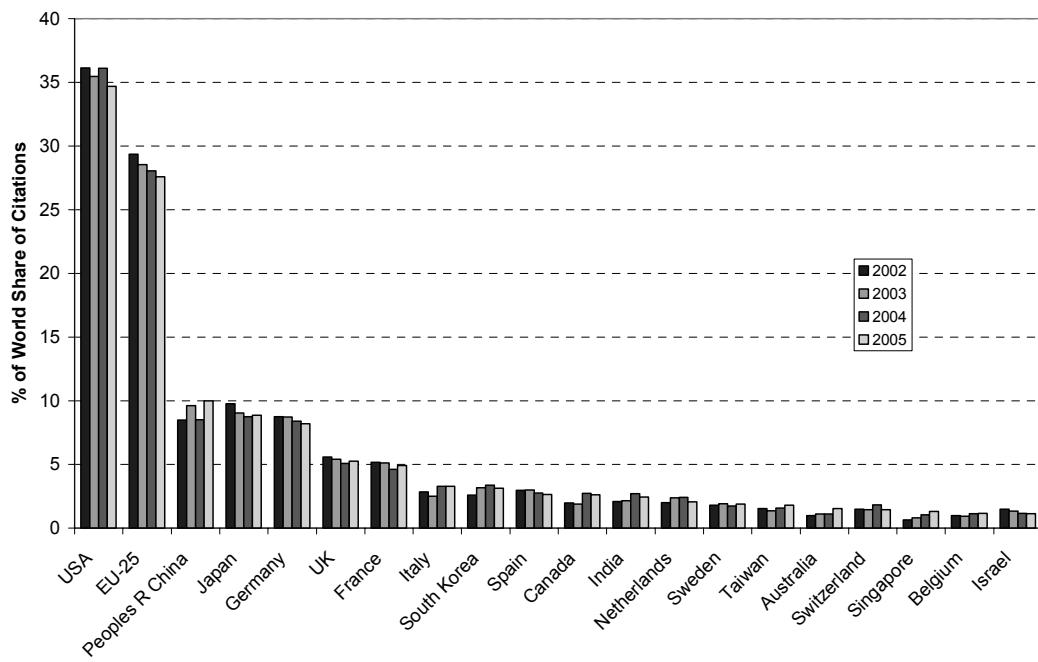


Figure 8. Percentage of world share of citations for leading countries and the EU-25 in the set of nanoscience journals, 2002-2004.

Figure 8 is based on measuring online the citation rates of all articles, reviews, and letters published in these ten journals during the years 2002-2005.² All measurement was done on October 1, 2006. The figure shows that in terms of citations the order between the EU-25 and the USA is reversed with respect to the number of publications. Unlike the general trend (Figure 6), the EU-25 is not improving its citation performance in this specialty area, but rather losing ground more rapidly than the USA to the East Asia-4.

Conclusions

The contribution of United States scientific institutions to refereed publications indexed by the Institute for Scientific Information has continued to rise in sheer numbers over the past decade, as have most countries. At the same time that the USA and other scientifically advanced countries have maintained slow growth, some countries that are newly developing their own scientific systems are making spectacular gains in numbers of publications contributed to refereed journals. As this happens, the USA and the EU drop as a percentage share of all publications. The drop in percentage share is not an absolute loss of ground.

In terms of citations, papers from the USA and the EU are becoming more equal on average, but the two major players are still far apart in the top segment of the 1% most highly cited papers. The USA is much more successful than the EU in coordinating its research efforts in strategic priority areas like nanotechnology. On all indicators, that is, absolute and relative, China shows exponential growth, while South Korea, Singapore, and Taiwan follow with sustained (mostly linear) growth during the past decade. In nanotechnology, China now ranks as the second nation behind the USA both in terms of number of publications and in its world share of citations.

Our conclusions accord with the conclusions of Braun & Dióspatoni (2005) that in terms of gatekeepers like editorial positions the dominance of the USA is unchallenged (Braun *et al.*, 2007). However, the data shows that the scientific system as a whole is growing, and new members are contributing to the pool of knowledge. As they do, the system as a whole benefits. Science is codified and networked at the global level, so it would be difficult to argue that any nationally defined contribution can “lose” in relation to any other part through the addition of new knowledge (Wagner & Leydesdorff, 2005). Far from losing ground in science to new entrants, the USA and other scientifically-advanced countries are gaining new colleagues and partners as well as access to new resources as other countries develop their scientific capacities.

References

- Bornmann, L., Leydesdorff, L., & Marx, W. (forthcoming). Citation Environment of *Angewandte Chemie. CHIMIA* (In print).
- Braun, T., Gläzel, W., & Schubert, A. (1991) The bibliometric assessment of UK scientific performance—some comments on Martin’s reply. *Scientometrics*, 20, 359-362.
- Braun, T., Schubert, A., Zsindely, S. (1997). Nanoscience and nanotechnology on the balance. *Scientometrics*, 38, 321-325.
- Braun, T. & Dióspatonyi, I. (2005). The counting of core journal gatekeepers as science indicators really counts. The scientific scope of action and strength of nations. *Scientometrics*, 52(3), 297-319.
- Braun, T., Zsindely, S., Dióspatonyi, I., & Zádor, E. (2007). Gatekeeping patterns in nano-titled journals. *Scientometrics*, 71(3), 651.
- Dosi, G., Llerena, P., Labini, M. S. (2006). The relationships between science, technologies and their industrial exploitation: An illustration through the myths and realities of the so-called ‘European Paradox’. *Research Policy*, 35(10), 1450-1464.
- Evidence, 2003. *PSA Target Metrics for the UK Research Base*. UK Office of Science and Technology, London, October 2003. Available at <http://www.dti.gov.uk/files/file14499.pdf> (last visited on 2 February 2007).

² The numbers are 5,807, 6,215, 6,788, and 9,013 documents for the four respective years. These 27,829 documents were cited 280,755 times in total until October 1, 2006. The numbers are 5,807, 6,215, 6,788, and 9,013 documents for the four respective years. These 27,829 documents were cited 280,755 times in total until October 1, 2006. These results should be considered as statistical because the citation counts in the field “times cited” are machine-generated. However, one can expect that the error thus induced is auto-correlated for consecutive years and therefore less affecting the general trend (Bornmann *et al.*, forthcoming).

- Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1), 35-41.
- Freeman, L. C., 1978/1979. Centrality in social networks: Conceptual clarification. *Social Networks*, 1, 215-239.
- Granstrand, O. (1999). *The Economics and Management of Intellectual Property: Towards Intellectual Capitalism*. Edward Elgar, Cheltenham, UK.
- Jaffe, A. B., Trajtenberg, M. (2002). *Patents, Citations, and Innovations: A Window on the Knowledge Economy*. MIT Press, Cambridge, MA/London.
- Jin, B., Rousseau, R. (2004). Evaluation of research performance and scientometric indicators in China. In Moed, H. F., Glänzel, W., Schmoch, U. (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 497-514) Kluwer Academic Publishers, Dordrecht, etc.
- King, D. A., (2004). The scientific impact of nations. *Nature*, 430(15 July 2004), 311-316.
- Kostoff, R., (2004). The (scientific) wealth of nations. *The Scientist*, 18(18), 10.
- Leydesdorff, L., (2005). The scientific impact of China. *Scientometrics*, 63(2), 411-412.
- Leydesdorff, L. (2006). "Betweenness centrality" as an indicator of the interdisciplinarity of scientific journals. Paper presented at the 9th International Science Technology Indicators Conference, Leuven, Belgium, 7-9 September 2006.
- Leydesdorff, L., Cozzens, S. E., Van den Besselaar, P. (1994). Tracking areas of strategic importance using scientometric journal mappings. *Research Policy*, 23, 217-229.
- Leydesdorff, L., Zhou, P. (2005). Are the contributions of China and Korea upsetting the world system of science? *Scientometrics*, 63(3), 617-630.
- Leydesdorff, L., Zhou, P. (2007). Nanotechnology as a field of science: its delineation in terms of journals and patents. *Scientometrics* 70(3), forthcoming.
- Martin, B. R. 1991. The bibliometric assessment of UK scientific performance—A reply to Braun, Glänzel and Schubert. *Scientometrics*, 20, 333-357.
- Meyer, M., Persson, O., 1998. Nanotechnology-interdisciplinarity, patterns of collaboration and differences in application. *Scientometrics*, 42(2), 195-205.
- Mogoutov, A., & Kahane, B. 2007. Data Search Strategy for Science and Technology Emergence: A Scalable and Evolutionary Query for Nanotechnology Tracking. *Research Policy*, 36 (In print).
- Mohrman, S., Wagner C. S., 2006. *The Dynamics off Knowledge Creation: A Baseline for the Assessment of the Role and Contribution of the Department of Energy's Nanoscale Science Research Centers*, University of Southern California, Marshall School of Business, Center of Effective Organizations, Los Angeles.
- Narin, F., Hamilton, K. S., Olivastro, D., 1997. The increasing link between U.S. technology and public science. *Research Policy*, 26(3), 317-330.
- National Science Board, 2006. *Science and Engineering Indicators*. NSF, Washington, DC.
- Porter, A., Youtie, J., Shapira, P., 2006. Refining Search Terms for Nanotechnology. in preparation.
- Shelton, R. D., 2006. Relations between national research investments inputs and publication outputs: application to the American Paradox. Paper presented at the 9th International Science Technology Indicators Conference, Leuven, Belgium, 7-9 September 2006.
- Shelton, R. D., Holdridge, G. M., 2004. The US-EU race for leadership of science and technology. *Scientometrics*, 60(3), 353-363.
- Sheu, M., Veeckind, V., Verbandt, Y., Galan, E. M., Absalom, R., Förster, W., 2006. Mapping nanotechnology patents: The EPO approach. *World Patent Information*, 28, 204-211.
- Wagner, C. S., Leydesdorff, L., 2005. Mapping the network of global science: comparing international co-authorships from 1990 to 2000. *International Journal of Technology and Globalization*, 1(2), 185-208.
- Zhou, P., Leydesdorff, L., 2006. The emergence of China as a leading nation in science. *Research Policy*, 35(1), 83-104.
- Zhou, P., Leydesdorff, L., 2007. A Comparison between the China Scientific and Technical Papers and Citations Database and the Science Citation Index in terms of journal hierarchies and inter-journal citation relations. *Journal of the American Society for Information Science and Technology*, 58(2), 223-236.
- Zitt, M., Bassecoulard, E., 2006. Delineating complex scientific fields by an hybrid lexical-citation method: A application to nanosciences. *Information Processing and Management*, 42, 1513-1531.

Revealed Similarities between the Journals *Nature* and *Science*: Using a New Cluster of Rhythm Indicators¹

Liming Liang

pllm@public.xxptt.ha.cn

Institute for Science Technology and Society, Henan Normal University, Xinxiang, 453007 (China),
University of Antwerp (UA), IBW, Universiteitsplein 1, B-2600, Wilrijk (Belgium)

Abstract

Creating a mean p-c matrix and changing the viewpoint from cited to citing year greatly expands the indicator system for describing the evolutionary rhythm of science. In this study the R_a -cluster indicators are defined, including the R_a , R_a' , r_a and r_a' indicators. A first application of the R_a -cluster indicators produces the R_a -cluster sequences of the journals *Nature* and *Science* for the recent half century, reflecting the rhythms of the two journals from different points of view – cited or citing year. Comparison of the two journals' various rhythm sequences shows us that according to the R_a -cluster indicators the two journals' rhythm sequences are very similar to each other. Comparing some specific time series reveals the main reason of this similarity.

Keywords

rhythm indicator; R_a -cluster; similarities; *Nature*; *Science*

Introduction

Two years ago a new indicator, called r-indicator, and the corresponding r-sequence were designed to describe the rhythm of scientific evolution. Two case studies were published using the data of the journals *Nature* and *Science* (liang, 2005; liang et al., 2006). the r-indicator and r-sequence were defined based on a general publication-citation matrix (here after p-c matrix for short), focusing on the “cited year”, i.e. the publication year. the methods were divided into two types in terms of the shape of the citation window used to calculate the r-sequences: the triangle method and the parallelogram method. the r-sequence is obtained by the triangle method, while the r' -sequence is obtained by the parallelogram method.

In the articles cited above a general p-c matrix was used, and the authors focused on just the cited year. from this a new idea emerged: why not try to define an r-indicator based on different types of p-c matrix, such as a mean p-c matrix, and expand our viewpoint from cited year to citing year? In this paper we will explore these ideas. the result will be the construction of a new cluster of rhythm indicators, the ra -cluster, covering indicators denoted as ra , ra' , r_a and r_a' . the methodological significance of the ra -cluster indicators is similar to that of the r-indicators: expanding the observation-based impact factor to an expectation-based impact factor and relative impact factor (liang et al., 2006).

Table 1. ra -cluster indicators

indicator	p-c matrix	point of view	citation window
R_a	mean	cited year	triangle
R_a'	mean	cited year	parallelogram
r_a	mean	citing year	triangle
r_a'	mean	citing year	parallelogram

The first application of the R_a -cluster indicators is to compare the evolutionary rhythms of the journals *Nature* and *Science*, two multidisciplinary journals with high impact factors. In the past some articles regarding *Nature* and *Science* and the comparison between them have been published (Kaneiwa et al.,

¹. This work was supported by the National Natural Science Foundation of China (70673019).

1988; Braun et al., 1989; Arkhipov, 1999). Our study takes a new approach, focusing on the journals' evolutionary rhythms.

Data

ISI's Web of Knowledge is used to obtain publication and citation data of *Nature* and *Science*. The query is: DocType = Article; Database(s) = SCI-EXPANDED, SSCI, A&HCI; Timespan=1955-2003; SO= (Nature) (or SO=(Science)). A total of 65,687 articles were found for *Nature*. Collection and classification of citation data were time-consuming, as we checked each of the 65,687 articles one by one. Citations of a certain cited article were classified by citing year. Finally, 4,951,022 citations were collected. For *Science*, 45,415 articles and 4,128,565 citations were retrieved.

Methodology

In this section we explain how to create a mean p-c matrix, and how to define the R_a -, $R_{a'}$ -, r_a - and $r_{a'}$ -indicators based on a mean p-c matrix.

Construction of a mean p-c matrix

A mean p-c matrix is derived from a general p-c matrix with P_i denoting the number of articles published in year i , C_{ij} the citations received in year j by the articles published in year i (Liang, 2005). We denote by A_{ij} the (arithmetic) average number of citations received in the year j by items published in the year i , i.e., $A_{ij} = C_{ij}/P_i$. The A_{ij} are the entries of the mean p-c matrix, see Table 2. The meaning of the grey part will be explained in the following section.

Table 2. Mean p-c matrix

	year	P_i	citing year j and mean citation A_{ij}							
			1	2	3	4	5	6	7	8
publication year i and number of publications P_i	1	1	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}
	2	1		A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}	A_{28}
	3	1			A_{33}	A_{34}	A_{35}	A_{36}	A_{37}	A_{38}
	4	1				A_{44}	A_{45}	A_{46}	A_{47}	A_{48}
	5	1					A_{55}	A_{56}	A_{57}	A_{58}
	6	1						A_{66}	A_{67}	A_{68}
	7	1							A_{77}	A_{78}
	8	1								A_{88}

Definition of the R_a - and $R_{a'}$ -indicators based on a mean p-c matrix from the viewpoint of cited year

The R_a -indicator is defined based on a mean p-c matrix as a time series of ratios $R_{ai} = O_i/E_i$, $i = 1, \dots, n$.

$$\text{Here, } O_i = \sum_{j=i}^n A_{ij}, \quad E_i = \sum_{k=1}^{n-i+1} A_k, \quad \text{with } A_k = \frac{\sum_{i=1}^{n-k+1} A_{i,i+k-1}}{n-k+1} \quad (*)$$

A_k is the key measure of the indicator R_a . The value $k = 1$ refers to the publication year. It can be proved that $\sum_{i=1}^n O_i = \sum_{i=1}^n E_i$. This equality implies that the calculation of the expected values does not increase or decrease the total number of citations. It is just a redistribution or rearrangement of actual citations over cited years. We notice that the used citation window is triangular.

The $R_{a'}$ -indicator is defined, using the parallelogram method, as a time series of ratios $R_{ai'} = O_i'/E_i'$. $R_{ai'}$ is only calculated for $i = 1$ to $n-w+1$. Here, $k_{max} = w < n$ determines the length of the citation window. The grey part in Table 1 shows such a citation window (as a parallelogram with $w = 4$). O_i' and E_i' are defined as follows.

$$O_i' = \sum_{j=i}^{i+w-1} A_{ij} , \quad E_i' = \sum_{k=1}^w A_k' , \quad \text{with} \quad A_k' = \frac{\sum_{i=1}^{n-w+1} A_{i,i+k-1}}{n-w+1}$$

It can be seen that E_i' is actually a constant, independent of i . Again, it is easy to prove that $\sum O_i' = \sum E_i'$.

Definition of the r_a' - and r_a -indicators based on a mean p-c matrix from the viewpoint of citing year

The r_a -indicator is also defined as a time series of ratios, $r_{aj} = o_j/e_j, j = 1, \dots, n$. Here,

$$o_j = \sum_{i=1}^j A_{ij} , \quad e_j = \sum_{i=1}^j A_{j-i+1} ,$$

In the formula of e_j the definition of A_k is the same as in formula (*). Again, we can prove the

$$\text{equality } \sum_{j=1}^n o_j = \sum_{j=1}^n e_j .$$

Suppose the citation window has a limitation $k_{max} = w (< n)$. Consequently, r_{aj}' is only calculated for $j = w$ to n . In this case the used C_{ij} values form a parallelogram, covering only $n - w + 1$ columns and w diagonals, as shown in the grey part of the matrix in Table 3. If this restriction is used we refer to the corresponding approach as the parallelogram method.

Table 3. Parallelogram citation window for the calculation of r_a' with $w=4$

		citing year j and mean citation A_{ij}								
publication year i and number of publications P_i	year	P_i	1	2	3	4	5	6	7	8
	1	1	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	A_{17}	A_{18}
	2	1		A_{22}	A_{23}	A_{24}	A_{25}	A_{26}	A_{27}	A_{28}
	3	1			A_{33}	A_{34}	A_{35}	A_{36}	A_{37}	A_{38}
	4	1				A_{44}	A_{45}	A_{46}	A_{47}	A_{48}
	5	1					A_{55}	A_{56}	A_{57}	A_{58}
	6	1						A_{66}	A_{67}	A_{68}
	7	1							A_{77}	A_{78}
	8	1								A_{88}

For the parallelogram method the r_a -indicator is only calculated for $j = w$ to n . We denote the indicator as $r_{aj}' = o_j'/e_j', j = w, \dots, n$. Using the parallelogram method we obtain the following formulae:

$$o_j' = \sum_{i=j-w+1}^j A_{ij} , \quad e_j' = \sum_{i=j-w+1}^j A_{j-i+1}' , \quad \text{with} \quad A_k' = \frac{\sum_{j=1}^{n-w+1} A_{j+w-k,j+w-1}}{n-w+1}$$

Obviously, e_j' is a constant when w is fixed, independent of j . Again: $\sum_{j=w}^n o_j' = \sum_{j=w}^n e_j'$.

Findings

The R_a -cluster sequences of the journals Nature and Science

Based on the data contained in *Nature*'s mean p-c matrix we calculated *Nature*'s R_a -, R_a' -, r_a - and r_a' -sequences. Similarly, we calculated *Science*'s R_a -, R_a' -, r_a - and r_a' -sequences based on *Science*'s mean p-c matrix. The calculations of the R_a - and r_a -sequences all use parallelogram methods with $w=10$. For the cases of *Nature* and *Science* $i=1$ and $j=1$ point to the year 1955.

Comparison of the R_a -cluster sequences between *Nature* and *Science*

In order to explore the similarities and dissimilarities between *Nature*'s rhythm sequences and the corresponding sequences of *Science* two methods are adopted. A statistical one: calculating correlation coefficients (CC for short) and t values for a t -test; and a visual one: showing the corresponding rhythm curves in the same figure and making a visual comparison of corresponding shapes.

Table 4 lists the CC's of the corresponding sequences of *Nature* and *Science*. For example, the CC of *Nature*'s R_a -sequence and *Science*'s R_a -sequence is 0.978. Generally speaking, these CCs are very high. Clearly, all the CCs are statistical significant at the 5% level. The t values for a t -test (H_0 : no linear correlation) are listed in Table 4. They are all much higher than the critical value at the 5% level.

Table 4. *Nature* and *Science*: CC and t value (1955-2003)

R_a-cluster sequence $n=49$	CC	t
R_a	0.969	26.889
R_a'	0.978	32.141
r_a	0.998	108.235
r_a'	0.995	68.299

Figure 1 contains four sub-figures. There is no doubt that the two journals' curves belonging to the same indicator are very similar. The two r_a curves almost coincide. We note that these similarities remain hidden when using the R - and R' -indicators, as well as when using other bibliometrics indicators, such as the journal impact factor.

In Figure 1 (a) we observe one high peak in *Nature*'s R_a curve in 1970. This is caused by an extremely highly cited paper in *Nature*, published by Laemmli U.K. (Laemmli, 1970). This particular article has been cited about 200,000 times and, by its own force, changed the rhythm of *Nature*. Indeed, highly cited papers play an important role in the progress of science. Several highly cited papers published in year i , or sometimes only one extremely highly cited paper, will greatly increase the C_{ij} , for j , in a decade or more. Such an article makes R_{ai} larger than for other years. Highly cited papers always attract peers' attention, therefore they have a special significance in the evolution of science.

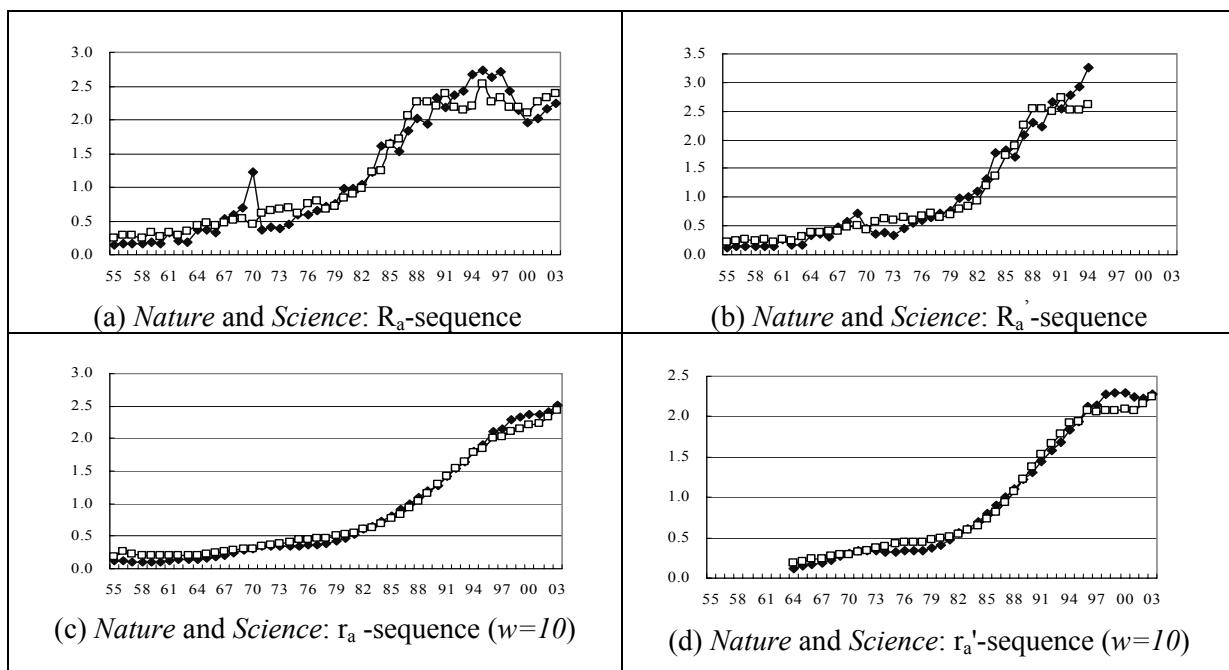
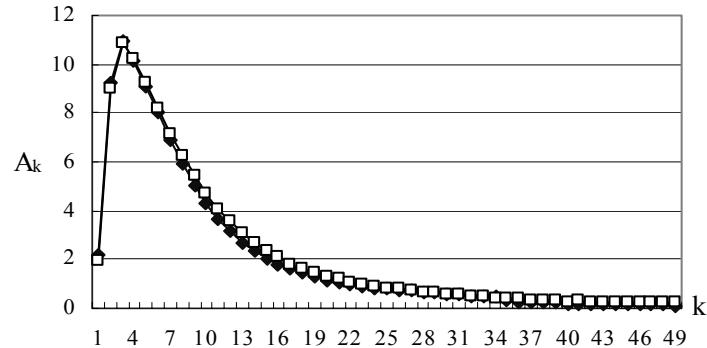


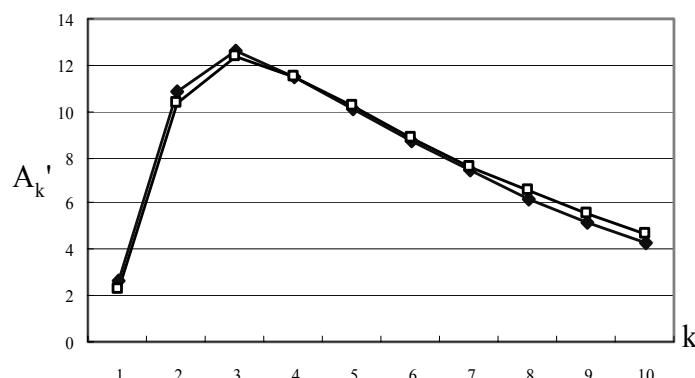
Figure 1. *Nature* and *Science*: the curve couples of the R_a -cluster sequences. (*Nature*: black diamonds; *Science*: white squares)

Comparison of the A_k (A'_k) series of Nature and Science

To find, at least partly, the reason why the two journals' rhythm sequences are so similar, the key measures A_k and A'_k are compared. Figure 2 shows two of the three A_k and A'_k curves of *Nature* and *Science*. The curve of A'_k for R_a' is omitted because of page limitations. *Nature*'s A_k and A'_k curves almost coincide with *Science*'s. Recalling the definitions of these indicators we consider this to be the best explanation for the observed similarities of the sequences of the R_a -cluster.



(a) *Nature* and *Science*: A_k for R_a and r_a



(b) *Nature* and *Science*: A'_k for r_a'

Figure 2. *Nature* and *Science*: the curve couples of the A_k and A'_k . (*Nature*: black diamonds; *Science*: white squares)

Conclusion and discussion

Creating a mean p-c matrix and changing the viewpoint from cited year to citing year greatly expands the indicator system for describing the evolutionary rhythm of science. In this study the R_a -cluster indicators are defined. The first application of the R_a -cluster indicators produced the R_a -cluster sequences of the journals *Nature* and *Science* for the recent half century, which reflects the evolutionary rhythms of the two journals from different points of view – cited or citing year. Comparison of the two journals' various rhythm sequences shows that the R_a -cluster indicators of the two journals are very similar to each other. A comparison of the two journals' A_k and A'_k series reveals the main reason why their rhythm sequences are similar.

In scientometric studies the term “rhythm” is seldom used, hence we would like to explain why we use this term. According to the Merriam-Webster online dictionary the word ‘rhythm’ has several meanings. One is “the aspect of music comprising all the elements (as accent, meter, and tempo) that relate to forward movement”. Another one is “the effect created by the elements in a play, movie, or novel that relate to the temporal development of the action”. We imagine the evolution of science as a

play presented at the stage of history with climaxes and dramatic changes. Like music the evolution of science has its own accent, meter and tempo, hence forming its own “rhythm”.

Another problem is worth discussing. What is the difference between the studies on “aging” (van Raan, 2000), as well as impact factors, and our rhythm studies? In one sentence the answer is that the former use only observed data, while the latter use the ratio of observed and expected values. In our view rhythm indicators are valid elements for exploring a purely theoretical point of view on indicator studies.

References

- Arkhipov D.B. (1999). Scientometric analysis of *Nature*, the journal. *Scientometrics*, 46 (1), 51-72.
- Braun,T., Glänzel, W. & Schubert, A. (1989). National publication patterns and citation impact in the multidisciplinary journals *Nature* and *Science*. *Scientometrics*, 17 (1-2), 11-14.
- Kaneiwa, K., Adachi, J., Aoki, M., Masuda, T., Midorikawa, N., Tanimura, A. and Yamazaki, S. (1988). A comparison between the journals *Nature* and *Science*. *Scientometrics*, 13(3-4), 125-133.
- Laemmli, U.K. (1970). Cleavage of structural proteins during assembly of head of bacteriophage-t4. *Nature*, 227 (5259), 680-
- Liang, L. (2005). The R-sequence: a relative indicator for the rhythm of science. *Journal of the American Society for Information Science and Technology*, 56, 1045-1049.
- Liang, L., Rousseau, R. and Fei S. (2006), A rhythm indicator for science and the rhythm of *Science*. *Scientometrics*, 68 (3), 535-544.
- Merriam-Webster online dictionary, <http://www.m-w.com/dictionary/rhythm>
- van Raan, A.F.J. (2000).On growth, ageing, and fractal differentiation of science. *Scientometrics*, 47(2), 347-362.

Hirsch-Type Indices and Library Management: The Case of Tongji University Library¹

Yuxian Liu* and Ronald Rousseau **, ***, ****

* *yxl@lib.tongji.edu.cn*

Library of Tongji University Siping Street 1239, 200092 Shanghai, (P.R. China)

, *, **** *ronald.rousseau@khbo.be*

KHBO (Association K.U.Leuven), Department of Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende (Belgium)

University of Antwerp (UA), IBW, Universiteitsplein 1, B-2610 Wilrijk (Belgium)
Hasselt University (UHasselt), Agoralaan, Building D, B-3590 Diepenbeek (Belgium)

Abstract

Hirsch-type indices are applied in a library management context. In this article quantitative, statistical approaches as well as a qualitative discussion are used to study the case of Tongji University Library. A comparison is made between the properties of different Hirsch-type indices. It is further shown that Hirsch-type indices can illuminate the reading interests of readers as shown by their use of a library's collection, hence expanding the field of application of such indicators.

Keyword:

Hirsch Index; g-index; H⁽²⁾-index; inequality measurement; library management; reading interests

Introduction

The h-index, recently proposed by J. Hirsch (2005) has attracted a lot of attention among scientometricians. Although introduced in a publication-citation context, it can easily be applied in many other source-items settings. In this article we intend to show that the Hirsch index, the g-index and the H⁽²⁾-index (definitions are provided in the next section) can be used as indicators in a library management setting. As a case study we chose Tongji University Library as one of us works there and has access to the relevant data.

In this article the original h-index, the g-index and the H⁽²⁾-index will be referred to as Hirsch-type indices. It is the purpose of this contribution to show how these indices are useful tools for library management.

The Hirsch index

Let us consider the list of publications [co]-authored by scientist S, ranked according to the number of citations each of these publications has received. Publications with the same number of citations are given different rankings (the exact order does not matter). Then S' Hirsch index is h if the first h publications received each at least h citations, while the publication ranked h+1 received strictly less than h+1 citations (Hirsch, 2005; Rousseau, 2006b). It is further noted that if the last article in the list occupies rank R and if it receives C > R citations then this scientist's h-index is set equal to R.

Although the h-index is a relatively simple indicator it attracted a lot of attention (Ball, 2005; Bar-Ilan, 2006; Bornmann & Daniel, 2005; Cronin & Meho, 2006; Glänzel, 2006a,b; Glänzel & Persson, 2006; Jin, 2006; Rousseau, 2006a; van Raan, 2006). Being indeed a relatively simple indicator a number of

¹ Acknowledgements

The authors thank Chief Librarian Dr. Shen Jinhua, not only for her help in the project, but also for encouraging librarians to explore new ways of developing library management. Liu gives her thanks to Chen Xin, Nie Yumei and all colleagues in the department of automation in the Library of Tongji University, who made it possible to access the library loan data used in this article. Liu further thanks Zhang Huibo, Tian Zhongzheng, Qian Liping, Qiu Xuejun and Zhang Dapeng for giving her the information necessary to understand the data. The authors specially thank Ms. Xu Wenyi for help with the organization of the article and the interpretation of the data. This work is sponsored by the National Natural Science Foundation of China (Project 70373055).

advantages and disadvantages are quite obvious. Following (Glänzel, 2006a; Hirsch, 2005) we note the following ones.

Advantages

- The h-index can be applied to any level of aggregation.
- It combines two types of activity (in the original setting this is citation impact and publications).
- It is a mathematically simple index.
- It is a robust indicator, see also (Rousseau, 2007). Increasing the number of publications alone does not have an immediate effect on this index.
- Single peaks (top publications) have hardly any influence on the h-index.
- In principle, any document type can be included.
- Unimportant (hardly ever cited) publications do not influence the h-index.

Disadvantages

- The h-index, in its original setting (Hirsch, 2005), puts newcomers at a disadvantage since both publication output and observed citation rates will be relatively low.
- The index allows scientists to rest on their laurels since the number of citations received may increase even if no new papers are published.
- This indicator is based on long-term observations. It can, moreover, never decrease.
- Like most pure citation measures it is field-dependent.
- There is a problem finding reference standards.
- There exist many more versatile indicators (van Raan, 2005).
- It is rather difficult to collect all data necessary for the determination of the h-index. Often a scientist's complete publication list is necessary in order to discriminate between scientists with the same name and initial.

It has been observed that (in the original context of publications and citations) the h-index is only weakly sensitive to the number of citations received. Indeed, when a scientist's h-index is equal to h then this scientist's first h articles received at least h times h , i.e. h^2 citations. For a given value of the h-index, this lower bound is the only relation that logically exists between publications and citations.

Hirsch-type indices

Because of this weak sensitivity with respect to the actual number of citations received, Leo Egghe proposed another index referred to as the g-index (Egghe, 2006a,b). This g-index is calculated as follows: one draws the same list as for the h-index, but now the g-index is defined as the highest rank such that the cumulative sum of the number of citations received is larger than or equal to the square of this rank. Clearly $h \leq g$. A scientist who writes many articles which are each well-received, but not exceptionally well, will have a high h-index. His g-index will just be marginally larger than his h-index. Stated otherwise the ratio of g/h will be close to 1 (but never smaller than 1!). A scientist who writes a few exceptional articles (maybe some reviews), while her other articles are hardly noticed by the scientific community will have a relatively low h-index and a high g-index. Examples of such cases are shown in (Rousseau, 2006b). Taken together, g and h present a concise picture of a scientist's achievements in terms of publications and citations.

Recently, another index has been proposed by Marek Kosmulski (2006). This index, denoted as $H^{(2)}$ is defined as follows. Consider again the list of publications [co]-authored by scientist S, ranked according to the number of citations each of these publications has received. Then this scientist's $H^{(2)}$ -index is k if k is the highest rank (largest natural number) such that the first k publications received each at least k^2 citations. Obviously: $H^{(2)} \leq h$. According to Kosmulski, the main reason for the introduction of this index is that it reduces the work of checking names, corresponding publications and received citations, and is still highly correlated with the total number of citations received.

It is clear that Hirsch-type indices can not only be used to evaluate lifetime publication-citation achievements, but also in the context of many other source-item relationships (Braun et al., 2005; Egghe & Rousseau, 2006). One of these other applications, illustrated in this article, is the case of library books as sources of loans in a university library. In this context the definition of a Hirsch index for loans is rephrased as follows. Consider the list of a collection of books in a library (it may also be the collection of all books in a particular branch library or book category) and the number of times these books were on loan during a fixed period. This list is ranked according to the number of loans (in decreasing order). Books with the same number of loans are given different rankings (again, the exact order does not matter). Then the Hirsch index for loans of this collection is h if h is the highest rank such that the first h books were on loan each at least h times. We will apply this definition for categories of books, and not for a library as a whole.

Tongji University and its Library Classification System

Tongji University is a comprehensive university situated in Shanghai and offering courses in engineering, science, medicine, management, arts, law and economics. It has special strengths in architecture, civil engineering and oceanography. Indeed, in a recent report Tongji University ranked among the top 5 Chinese universities in architecture, urban planning, civil engineering, environmental sciences, traffic and transportation engineering, equipment engineering and engineering management (Qiu et al., 2006).

The history of Tongji University can be traced back to 1907 when Tongji German Medical School was founded by Erich Paulun, a German doctor living in Shanghai. As the result of a nationwide reorganization Tongji University became in 1952 a university with strengths in engineering, focusing on civil engineering. Since then the university gradually developed to be a comprehensive university, strong in engineering but also offering programs in science, economics, management, arts and law. In 1996 Tongji University merged with Shanghai Institute of Urban Construction and Shanghai Institute of Building Materials. In April 2000, the expanded Tongji University merged with Shanghai Railway University. In this way the university consists of several campuses covering a total area of 141.8 hectares: Main Campus (including South Campus), Hudong Campus, Hubei campus, Huxi Campus, and Jiading Campus, a new campus situated inside Shanghai International Automobile City, a suburban district of Shanghai. There are libraries at each campus, except for the South Campus, which is served by the Main Library (situated at the Main Campus). Since 1978 the university restored its special relationship with Germany, resulting in the establishment of the Chinese-German University College. Nowadays Tongji University registers more than 50,000 students at all levels of education (bachelor, master, doctoral and post-doctoral). There are over 4,200 academic staff members for teaching and/or research.

Tongji University Library applies the Chinese Library Classification System (CLC), a comprehensive library system used in most libraries, information institutes and centres in China. The Chinese Academy of Sciences, however, uses another system. The CLC system has gone through several revisions, its most comprehensive one dating from 1999. According to the CLC system there are 22 main categories, shown in Table 1. These main categories are further subdivided into subcategories.

As Tongji University is a university specializing in science and engineering we will also use the subcategories of category T: Industrial technology. These subcategories are shown in Table 2.

Data collection

Data are obtained from the loan records of the Library of Tongji University, consisting of a Main Library and four branch libraries. The Main Library is situated at the Main Campus (also referred to as Siping Campus) in the centre of Shanghai. Branch libraries are located on four other campuses: Huxi, Hudong, Hubei and Jiading Campus. The Huxi branch library serves mostly freshman students. Senior undergraduate students study on Main campus. Hudong library serves mostly master and doctoral students; Hubei library serves mostly vocational students. While these campuses and their libraries are also situated in Shanghai, Jiading Campus is a new campus located within Shanghai Jiading

International Automobile City, somewhat outside the centre, and built for over 15,000 students. It received its first students in September 2004. Students at Jiading Campus specialize in automobile related subjects and software engineering.

Table 1. Chinese Library Classification System

A	Marxism, Leninism and Chinese communism
B	Philosophy and religion
C	Social sciences
D	Politics and law
E	Military sciences
F	Economics
G	Culture, science (of sciences), education
H	Languages (incl. linguistics)
I	Literature
J	Arts
K	History and geography
N	Natural sciences (general)
O	Mathematics, physics and chemistry
P	Astronomy and geosciences (incl. marine sciences)
Q	Bioscience
R	Medicine and hygiene
S	Agricultural science (including forestry)
T	Industrial technology
U	Transportation
V	Aviation and spaceflight
X	Environmental sciences
Z	Others

Table 2. Subcategories of category T: Industrial technology

TB	Fundamental engineering technology
TD	Mining engineering
TE	Oil and natural gas industry
TG	Metal industry
TH	Mechanics
TJ	Weapon industry
TK	Dynamics and sources of energy
TL	Atomic energy technology
TM	Electrotechnics
TN	Wireless electronics and telecommunication technology
TP	Computer science
TQ	Industrial chemistry
TS	Light industry
TU	Architecture and urban planning
TV	Water conservation and irrigation

Branch libraries apply different policies as to which books are kept in the reading room and which are available for regular check-out. In Huxi library books in categories B, C, D, F, G, H, O, and T are transferred from regular check-out to the reading room, once it is clear that they are in demand. In Hudong library, all books published after 2000 were transferred to the reading room, while all older books are available for regular check-out. Finally, Hubei library transferred all medicine books to the Main Library, but does not do any other transfers. The most recent books, though, are only available in the reading room.

Library loan data for each library (five in total) were collected in August 2006 and cover the period March 21, 2001 till July 31, 2006. The beginning date coincides with the introduction of a computerized library loan system for the whole university. Tongji University uses the Huwen Library System software. This library software contains several statistical functions, but we only use the function which ranks books according to the number of loans. In each library the software provided us with data for the reading room (or reading rooms) and general check-out separately. Jiading Campus Library, however, has only reading rooms for the moment. Moreover, data for this library cover only a two-year period. All these data were copied in an Excel spreadsheet in order to calculate the Hirsch index, the g-index and the $H^{(2)}$ -index. Usually the library owns several copies of a book. In those cases all loans are brought together leading to exactly one number (of loans) for each title.

Structural and statistical data analysis

The Hirsch-index, g-index and $H^{(2)}$ -index are calculated for each class in each library, and this for the reading room(s) and regular check-out separately. In terms of sources and items, each class is considered a source (playing the role of the authors in Hirsch's original setting) while loans are items (playing the same role as citations in Hirsch's original setting). Table 3 shows the results for the Main Library (R stands for Reading room; C for check-out). The first number gives the rank, while the second number gives the value of the index. Results for the other libraries can be obtained from the authors. For example, class TP (*Computer science*) is always number one in the Main Library's reading room. For the reading room it has an $H^{(2)}$ -value of 4, an h-value of 15 and a g-value of 19. Compared to other categories, books on computer science are somewhat less popular for check-out: in this list computer science books occupy the fifth place. Values of the indices, however, are much higher for check-out than for the reading room: $H^{(2)} = 13$, $h = 105$ and $g = 131$. This phenomenon only occurs in the Main Library. At the other libraries the opposite is true.

On logical grounds the inequalities $H^{(2)} \leq h \leq g$ always hold. Interestingly, we observed some non-trivial equalities for h and g. For instance loan data in class V: "Aviation and spaceflight" for the Jiading reading room are shown in Table 4. Clearly the h-index is equal to 4 (the book at rank 4 is loaned out 5 times, while the book at rank 5 is loaned out only 4 times). The g-index is also equal to 4 as $4^2 = 16 < 20$ (the cumulative number of loans for the first 4 books), while $5^2 = 25 > 24$ (the cumulative number of loans of the first 5 books).

Do different Hirsch-type indices lead to the same ranking of library classification classes? Are these ranking the same for reading room data as for check-out data? We tried to answer these questions by calculating Spearman rank correlation coefficients. Table 5 shows the results for the Main Library. Correlations are generally high, and are especially high when correlations between the three Hirsch-type indices for the reading room, and between the check-out data are compared (in bold). Similar tables were obtained for the branch libraries and can be obtained from the authors.

Table 3. Main Library (rank followed by value of the Hirsch-type index in the cells)

		H⁽²⁾ - R	H⁽²⁾-C	h - R	h - C	g - R	g-C
A	<i>Marxism- Leninism and Chinese communism</i>	28-1	26-6	28-2	25-27	30-2	26-36
B	<i>Philosophy and religion</i>	14-2	8-11	14-5	7-80	19-5	8-101
C	<i>Social sciences</i>	14-2	8-11	10-6	11-69	14-6	11-91
D	<i>Politics and law</i>	14-2	16-9	14-5	15-56	14-6	15-71
E	<i>Military sciences</i>	28-1	23-7	28-2	26-25	25-3	25-39
F	<i>Economics</i>	5-3	5-13	10-6	6-92	7-9	6-121
G	<i>Culture- science (of sciences)- education</i>	14-2	19-8	19-4	19-41	22-4	19-56
H	<i>Languages (incl. linguistics)</i>	5-3	3-15	10-6	3-139	12-9	3-177
I	<i>Literature</i>	5-3	1-16	7-7	2-148	10-8	2-186
J	<i>Arts</i>	14-2	7-12	19-4	8-78	19-5	7-105
K	<i>History and geography</i>	14-2	8-11	19-4	8-78	19-5	8-101
N	<i>Natural sciences (general)</i>	14-2	23-7	24-3	23-28	25-3	22-43
O	<i>Mathematics- physics and chemistry</i>	5-3	3-15	3-12	4-110	5-14	4-149
P	<i>Astronomy and geosciences (incl. marine sciences)</i>	14-2	16-9	14-5	17-43	14-6	18-58
Q	<i>Bioscience</i>	5-3	19-8	14-5	20-37	12-7	20-53
R	<i>Medicine and hygiene</i>	1-4	23-7	5-11	22-31	3-16	22-43
S	<i>Agricultural science (including forestry)</i>	28-1	26-6	24-3	29-19	25-3	30-29
TB	<i>Fundamental engineering technology</i>	1-4	12-10	3-12	14-57	4-15	13-81
TD	<i>Mining engineering</i>	32-0	32-0	32-0	32-0	32-0	32-0
TE	<i>Oil and natural gas industry</i>	32-0	32-0	32-0	32-0	32-0	32-0
TG	<i>Metal industry</i>	14-2	26-6	19-4	28-22	22-4	27-35
TH	<i>Mechanics</i>	14-2	12-10	14-5	16-51	14-6	15-71
TJ	<i>Weapon industry</i>	32-0	32-0	32-0	32-0	32-0	32-0
TK	<i>Dynamics and sources of energy</i>	14-2	19-8	24-3	21-34	22-4	21-48
TL	<i>Atomic energy technology</i>	32-0	32-0	32-0	32-0	32-0	32-0
TM	<i>Electrotechnics</i>	5-3	16-9	10-6	18-42	10-8	17-64
TN	<i>Wireless electronics and telecommunication technology</i>	5-3	12-10	6-10	13-64	6-13	13-81
TP	<i>Computer science</i>	1-4	5-13	1-15	5-105	1-19	5-131
TQ	<i>Industrial chemistry</i>	14-2	19-8	19-4	23-28	14-6	24-42
TS	<i>Light industry</i>	14-2	26-6	24-3	27-23	25-3	28-32
TU	<i>Architecture and urban planning</i>	1-4	1-16	2-14	1-149	2-17	1-188
TV	<i>Water conservation and irrigation</i>	14-2	26-6	28-2	30-18	25-3	29-30
U	<i>Transportation</i>	5-3	8-11	7-7	10-73	7-9	10-100
V	<i>Aviation and spaceflight</i>	28-1	26-6	31-1	31-16	30-2	31-22
X	<i>Environmental sciences</i>	5-3	12-10	7-7	12-66	7-9	12-87

Table 4 Loans in Jiading reading room, class V

Rank	(Rank)²		Number of loans	Cumulative number of loans
1	1	Book A	5	5
2	4	Book B	5	10
3	9	Book C	5	15
4	16	Book D	5	20
5	25	Book E	4	24
6	36	Book F	4	28
...

Table 5. Rank correlation according to Hirsch-type indices between library classification classes
(Main Library)

	H⁽²⁾-R	H⁽²⁾-C	h-R	h-C	g-R	g-C
<i>H⁽²⁾-R</i>	1.0000	0.7333	0.9347	0.7467	0.9531	0.7511
<i>H⁽²⁾-C</i>		1.0000	0.8187	0.9900	0.7943	0.9925
<i>h-R</i>			1.0000	0.8306	0.9810	0.8270
<i>h-C</i>				1.0000	0.7965	0.9979
<i>g-R</i>					1.0000	0.7956
<i>g-C</i>						1.0000

Another important issue is whether these Hirsch-type indicators are able to separate sources, i.e. library classification classes in our case. We use the Gini coefficient, a measure of concentration, as an indicator of separation. Table 6 shows the values of the Gini coefficient between library classification classes for each library and each Hirsch-type index.

Table 6. Gini coefficients

	H⁽²⁾	h	g
<i>Main R</i>	0.286	0.399	0.412
<i>Main C</i>	0.270	0.420	0.396
<i>Huxi R</i>	0.479	0.635	0.646
<i>Huxi C</i>	0.407	0.542	0.539
<i>Hudong R</i>	0.332	0.464	0.464
<i>Hudong C</i>	0.326	0.433	0.431
<i>Hubei R</i>	0.335	0.478	0.472
<i>Hubei C</i>	0.322	0.432	0.442
<i>Jiading R</i>	0.280	0.432	0.432
<i>average</i>	0.337	0.471	0.470

Not surprisingly, the H⁽²⁾-index does not discriminate in the same way among classes as the h- and g-indices, leading to smaller values of the Gini coefficient. We see further that the average values of these inequality measures are very similar for the h- and the g-index. In Egghe's example of Price medallists (Egghe, 2006c) the g-index discriminates more than h-index, but, of course, the type of data studied there is totally different from ours.

Qualitative data analysis

In this section we discuss some interesting observations about the actual Tongji data. Hirsch-type indices are always higher for the reading rooms than for check-out in Hudong and Hubei branches, while for the Main Library the opposite is the case. Huxi library shows a mixed pattern. This may be explained by the fact that in some branch libraries (Hudong and Hubei) books available for check-out are usually older books, and hence for some fields less interesting. These older books are mainly the older collections, namely those obtained before the local university merged with Tongji University. This observation suggests that newer books have a higher Hirsch-type index for loans than older ones (under the same loan conditions). *Computer science* usually ranks higher for reading room data than for check-out data. This can be explained by the fact that the most recent books are mainly consulted in the reading rooms. Huxi library shows a mixed pattern because books are transferred between reading room and check-out depending on the classification type. The different pattern for the Main Library is due to the fact that every book owned by Tongji University Library is available for loan at the Main Library.

We observe that values of the Hirsch-type indices are highest for Hudong Campus, followed by Huxi and Hubei, in that order. This illustrates the fact that master and doctoral students use more books than undergraduate students, who, in turn, use more books than vocational students.

Books in the categories H (*languages*), I (*literature*), O (*Mathematics-physics-chemistry*), TP (*Computer science*) and TU (*Architecture and urban planning*) are usually among the top-5 in the rankings according to Hirsch-type index. The high scores for TU (*Architecture and urban planning*) illustrate the fact that this is the topic in which Tongji University excels (Qiu et al., 2006). Even at Jiading campus this category scores relatively high (between 10th and 13th position) although no architecture is taught at this campus.

Some smaller changes in Hirsch-type indices' ranks indicate differences in emphasis in different campus. In the Main Library the categories *Medicine and hygiene* and *Fundamental engineering technology* rank high, but only for the reading rooms. For *Medicine and hygiene* this may be explained by the fact that the valuable medicine books have been transferred from Hubei library (formerly the library of the medical university) to the Main Library. Moreover, all books on medicine are available in a reading room, leading to high Hirsch-type indices in *medicine and hygiene*. Scores for *fundamental engineering technology* can be explained by the fact that this category contains the books which are fundamental for most undergraduates in technologically oriented fields. In Huxi library *philosophy, religion and Marxism* also scores high, as do books on medicine and biosciences (check-out data). Medicine and biosciences score high because all books belonging to these categories are available for check-out. Moreover *philosophy, religion and Marxism* are compulsory courses for undergraduates. Category U (*Transportation*) scores very high at Jiading library, illustrating the fact that the Jiading campus specializes in automobile related subjects and software engineering (*Computer science* ranks first).

Jiading Library's average value for the h-index (reading room) is 19.9 which is of the same order as Hudong's (21.8) and Huxi's (16.5) and higher than Hubei's (10.9). Remembering that we have only data for the latest two year, this shows that this library is very successful and attracts many students to its premises.

Tongji Library has only few books on mining engineering (TD), oil and natural gas industry (TE), weapon industry (TJ), and atomic energy technology (TL). In some libraries such books are not even available. For this reason these categories can usually be found at the end of the ranked lists.

Discussion and conclusions

Applications of bibliometrics to library management have always been an integral part of the information sciences (Burrell, 1980; Egghe & Rousseau, 1990; Leemans et al., 1992; McCain, 1997). In this article we have shown that also the latest developments in the field, i.e. the Hirsch-type indices, can be applied to library management issues. Hirsch-type indices can illuminate the reading interests of readers as shown by their use of a library's collection. Consequently, changes in these indices may reflect changes in users' interest. Yet, they may also reflect changes over time in the quality of different collections.

In this article we considered rather straightforward applications of the original definitions of the Hirsch-type indices. We ask as a research question: could another (related) index be invented which is particularly suited for use in a library management environment?

References

- Ball, P. (2005). Index aims for fair ranking of scientists. *Nature*, 436, p. 900.
- Bar-Ilan, J. (2006). H-index for Price medalists revisited. *ISSI Newsletter* 2(1), p. 3-5.
- Bornmann, L. and Daniel, H.-D. (2005). Does the h-index for ranking of scientists really work? *Scientometrics*, 65, 391-392.
- Braun T., Glänzel, W. and Schubert A. (2005). A Hirsch-type index for journals. *The Scientist*, 19(22), p.8.
- Burrell, Q.L. (1980). A simple model for library loans. *Journal of Documentation*, 36, 115-132.

- Cronin, B. and Mehro, L. I. (2006). Using the h-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, 57, 1275-1278.
- Eghe, L. (2006a). How to improve the h-index. *The Scientist*, 20(3), p. 14.
- Eghe, L. (2006b). An improvement of the H-index: the G-index. *ISSI Newsletter*, 2(1), 8-9.
- Eghe, L. (2006c). Theory and practice of the g-index. *Scientometrics*, 69, 131-152.
- Eghe, L. and Rousseau, R. (1990). *Introduction to Informetrics. Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier.
- Eghe, L. and Rousseau, R. (2006). An informetric model for the Hirsch index. *Scientometrics*, 69, 121-129.
- Glänzel, W. (2006a). On the opportunities and limitations of the H-index. *Science Focus*, 1(1), 10-11 (in Chinese).
- Glänzel, W. (2006b). On the h-index – a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67, 315-321.
- Glänzel, W. and Persson, O. (2005). H-index for Price medalists. *ISSI Newsletter*, 1(4), 5-18.
- Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA*, 102(46), 16569-16572.
- Jin, B. (2006). H-index: an evaluation indicator proposed by scientist. *Science Focus*, 1(1), 8-9. (In Chinese).
- Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter*, 2(3), 4-6.
- Leemans, M.J., Maes, M., Rousseau, R. and Ruts, C. (1992). The negative binomial distribution as a trend distribution for circulation data in Flemish public libraries. *Scientometrics*, 25, 47-57.
- McCain, K. W. (1997). Bibliometric tools for serials collection management in academic libraries. *Advances in Serials Management*, 6, 105-146.
- Qiu, J., Zhao, Y., Yu, Y., Liu, Y., Zhu, S., Yin, Z. and Ma, R. (2006). Names of universities in the top 5 percent of the discipline specialty ranking. *Science Time, 5th May, 2006*. (In Chinese). The same information is also available at: <http://rccse.whu.edu.cn/html/2006/06/20060607172359-1.htm> [last viewed: November 17, 2006]
- Rousseau, R. (2006a). A case study: evolution of JASIS' h-index. *Science Focus*, 1 (1), 16-17 (in Chinese). English version available at E-LIS, code 5430.
- Rousseau, R. (2006b). New developments related to the Hirsch index. *Science Focus*, 1 (4), 23-25 (in Chinese). English version available at E-LIS, code 6376.
- Rousseau, R. (2007). The influence of missing publications on the Hirsch index. *Journal of Informetrics*, 1, 2-7.
- Van Raan, A.F.J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67, 491-502.

A Quantitative Hstoriography of Mexican Integration into the International Standards of Scientific Research¹.

Ma. Elena Luna-Morales^{*}, Francisco Collazo-Reyes^{**} and Jane M. Russell^{***}.

^{*}elena@csb.cinvestav.mx

Centro de Investigación y de Estudios Avanzados del IPN, Unidad de Servicios Bibliográficos, Av. IPN 2508,
Col. San Pedro Zacatenco, 07360, Mexico DF (Mexico).

^{**}fcollazo@fis.cinvestav.mx

Centro de Investigación y de Estudios Avanzados del IPN, Departamento de Física, Av. IPN 2508, Col. San
Pedro Zacatenco, 07360, Mexico DF (Mexico).

^{***}jrussell@servidor.unam.mx

UNAM (Universidad Nacional Autónoma de México), Centro Universitario de Investigaciones
Bibliotecológicas, Ciudad Universitaria, 04510 México DF (Mexico).

Abstract

A study is presented on Mexican science published in mainstream journals during the first half of the 20th century, based on the analysis of the bibliographical elements present in the records from the *Science Citation Index Expanded* from 1900-1950. Organizational structures and patterns of communication, publication and citation were determined which represented the scientific practices of the research community during this period. We found three distinct modes of knowledge production: amateur, institutional and academic which corresponded to different periods in the process of the incorporation of Mexican science into the communication patterns of international science. The production modes were characterized by a variety of indicators: periods and types of research, publication and citation patterns, author production, journals and subject categories, institutional structure, and geographic distribution of production.

Keywords

Mexican science; knowledge production modes; quantitative historiography; scientific communication patterns.

Introduction

The recent availability online of the *Century of Science* initiative within the *Web of Science* presents us with the possibility of developing new perspectives on the growth and evolutionary paths taken by science during the 20th century in countries like Mexico. The wealth of objective research data available can assist historians and sociologists of science and other scholars (Shapin, 1992), to re-examine, reflect on, strengthen or put to the test what has been written about the past in the field of science studies and even in some cases, to correct traditional historical methods (Kragh, 1987). Traditionally, the historical record, as a collection of personal and collective testimonies, has relied heavily on the readings, bibliographical research, and collective memory of the specialist scientific communities, without the added advantage and insight that bibliometric data can provide. Bibliometrics is most often used in the field of library and information science. However, it can be applied to any discipline to learn more about its scholarly content. In the history of science, it is used to elucidate the development of scientific disciplines by tracing the historical movements that are revealed in the results obtained by researchers (Okubo, 1997). While scientometrics and bibliometrics focus on the formal reporting of science, historiography, in contrast, tends to explain the formal on the basis of the informal (Edge, 1979).

¹ This work was partially supported by CONACYT (Mexico) and by the *Programa Transdisciplinario en Investigación y Desarrollo para Facultades y Escuelas, Unidad de Apoyo a la Investigación en Facultades y Escuelas, UNAM: Macroproyecto de las Tecnologías para la Universidad de la Información y la Computación*.

The historiography of science was originally written by philosophers and by practicing and retired scientists (Christie, 2005) as a way to communicate the virtues of science to the public. In the 1930s, effort was directed towards looking at the ways in which scientific practices were allied with the needs and motivations of their context. By the 1960s with the increasing importance of science and technology to modern life, the emphasis was on problematizing the scientific enterprise, thus making it difficult to reach consensus as to the best way to write its history (Suárez, 2005). The professionalization of the historiography of science during the 20th century, increasingly in the hands of the historians of science and now accepted as a legitimate field of academic study (Christie, 2005), brought with it a greater richness and variety of interpretations and viewpoints (Laudau, 2005) and an increasing demand for specialized documental and other information resources. In 1963, Eugene Garfield, influenced by *Shepard's Citation Index* for legal cases, founded the *Science Citation Index* (SCI). Although bibliometric analysis predates it, SCI and its access to ISI's large datasets increased the popularity of bibliometric research, especially outside the field of information science. The SCI and more recently the *Web of Science* have become the most generally accepted basic source for bibliometric analysis. With the advent of *Century of Science* initiative we can now go back and use bibliometrics on the timeline previous to the creation of the SCI.

The prevailing vision of the history of Mexican science has been written mainly by members of the national scientific community, through personal viewpoints, essays and monographs, some as eyewitness accounts of the period (Beltrán, 1952, 1970, 1989; Gortari, 1963), others as eminent saviors of our scientific past by way of general works (Pérez-Tamayo, 2005; Moreno, 1986), biographies of prominent scientists (Coordinación de la Investigación Científica, 2003; Academia Mexicana de Ciencias, 2003), evolution of disciplines (Pérez-Angón, 2006; García, and Pérez, 2006), as well as institutional histories. The emergence of the first studies by specialists in the sociology and historiography of Mexican science from the 1980s and 1990s, (Trabulse, 1983; Casas, 2003; Saldaña, 1982) introduced new ways of writing about Mexican science, the questions asked were more diverse as were the study objectives, and diagnosis went deeper. Mexican science required a greater wealth of historiographic sources and interpretations (Trabulse, 1996, 2003) to complement the mainly descriptive and externalist approaches (Casas, 2003) and to strengthen conceptual study frameworks by incorporating both endogenous and exogenous factors to the analysis of the local maturing process of the sciences (Saldaña, 1994).

The present study, therefore, puts a different perspective on the history of Mexican science by analyzing its production and communication patterns during the first half of the 20th century. We are especially interested in mapping the organizational and disciplinary structures that emerged over this period, as well as publication patterns which characterized knowledge production during this time. We also hope to identify aspects that help characterize the process of transition and rupture between knowledge production based on immediatist intentions, centred on specific regional developments and applications (Pérez-Tamayo, 2005) and the new mode of knowledge production that started to gestate within the universities during this phase of development.

Material and Methods

Information sources:

- *Science Citation Index Expanded (SCIE)*, 1900-1950
- Library holdings and fulltext journals
- Internet
- Literature on the history of Mexican science

Methodology

The new version of SCIE allows information to be recovered by country name from 1940 to 1945 and from 1973 to the present but not between 1945 and 1972 where records did not include author affiliations. The scientific production of Mexico from 1900-1945 was downloaded from the SCIE using the following search strategy "MEXICO NOT NEW MEXICO", and from 1945-1950, using the names of authors who featured prominently in different texts on the history of Mexican science and

were therefore considered likely to have published in the mainstream literature. We consulted available library, fulltext journals and Internet sources to determine the institutional affiliations of these scientists. By way of the analysis of the common elements of bibliographical records (such as author, title, journals, addresses, numbers of references and citations, document type, categories) and the application of quantitative, disaggregation and word frequency techniques, we were able to identify important differences between the institutional structures, forms of organization and scientific communication patterns present during the period analyzed. This allowed us to identify three knowledge production modes: amateur (Saldaña and Azuela, 1994), institutional and academic, each one with its own mechanisms for incorporating the standards of publication of international science which we were able to characterize with the help of 13 different bibliometric indicators.

Results

We recovered a total of 185 mainstream publications spanning the years from 1900 to 1950. More important than the actual numbers of papers is the fact that Mexican science showed a continuous, albeit small, presence in the international databases during this period of greatest political instability in the history of the country caused by the revolution of 1910.

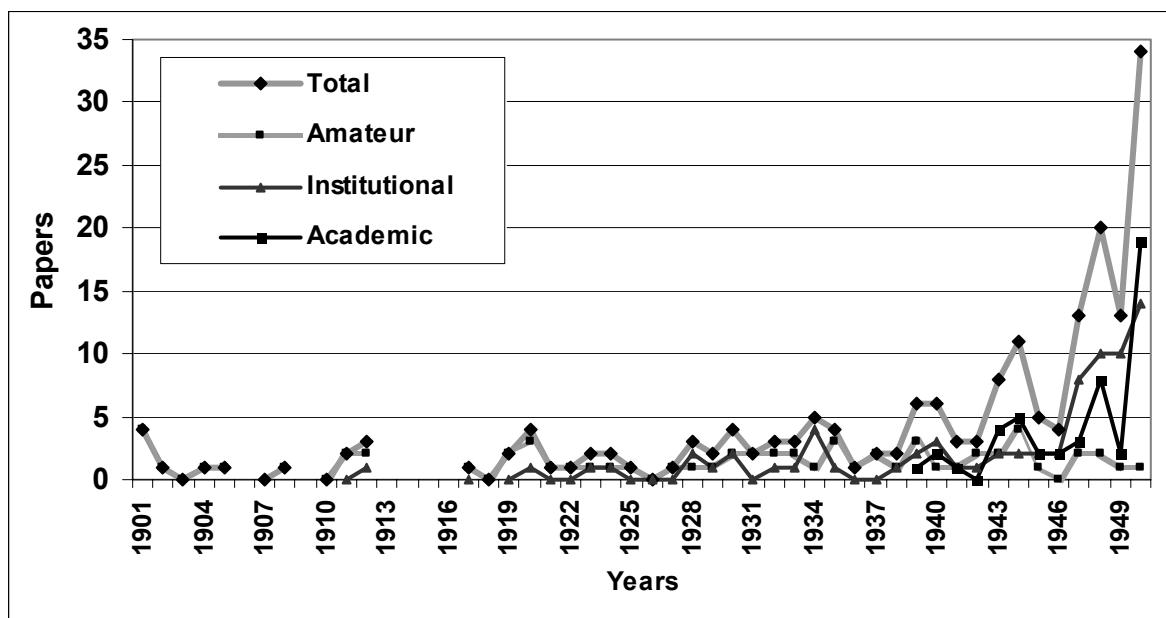


Figure 1. Scientific production modes of Mexican science during the first half of the 20th century.

Figure 1 shows the general distribution of scientific production from 1900 to 1950 according to the three different types of knowledge production and communication. Two distinct periods are apparent, the first covers the long period of reliance (1900-1930) on the customary ways of doing science, based on a single type of scientific practice. This changed during the second period from 1930 to 1950, with the coexistence of three different modes of knowledge production.

One third (62) of papers correspond to the amateur mode which exhibits a static form of production, suggestive of precarious scientific activity related to limited organizational structure and isolated scientific practices. In contrast to the other two modes the amateur mode is present during the most part of the fifty-year period and represents the most traditional way of doing science in Mexico during this time. The institutional mode is responsible for the largest part of the production, 40% (74) appearing in the intermediary decades as an emergent form of production which is consolidated during the last five years of the period under study. The studies undertaken in the academic environment correspond to 26.5% (49) of the total and are concentrated in the final decade of the period. These represent the first international publications produced by the newly-established research institutes and are mainly physics papers authored by the National Autonomous University of Mexico (UNAM).

Each one of the production modes was sustained by distinct circumstances and organizational structures which replicated clearly differentiated patterns of publication, citation and information usage characterized in Table 1.

Table 1. Production modes and publication and citation patterns.

Characteristics	Mode 1	Mode2	Mode3
<i>Denomination</i>	Individual or Amateur	Institutional	Academic
<i>Period</i>	Occurs throughout the period but mainly from 1900-1940	Begins towards the end of the 1910s.	Begins towards the end of the 1930s.
<i>Type of research</i>	Isolated, local, utilitarian	Continuity in the production: topics, authors and institutions.	Scientific practice: institutional, professional and independent
<i>Publication patterns</i>	Sole author: 93 % Co-authorship: 7 %	Sole author: 38 % Co-authorship: -Two authors: 24 % -More than two: 38 %	Sole author: 73 % Co-authorship: -Two authors: 15 % -More than two: 12 %
<i>Reference and citation patterns</i>	Av. no. references: 2 Av. no. citations: 0.16	Av. no. references: 11 Av. no. citations: 13	Av. no. references: 6 Av. no. citations: 6
<i>Document type</i>	-Articles -Technical Reports -Descriptions -Reports	-Articles -Meeting Abstracts -Reviews -Notes -Letters	-Meeting abstracts -Articles -Letters -Notes -Editorial Material
<i>Language</i>	-English -French	-English -French -German	-English -French -Spanish -German
<i>Authors</i>	-Medina, M -Ordóñez, E -Sánchez, PC -Matuda, E -Herrera, AL -Urueña, JG -Gallo, J	-Giral, F -Varela, G -Costero, I -Mazzotti, L -Rosenblueth, A -Mooser, H -Hudson, NP -Gómez, F -Romo, J -Rosenkranz, G -Chavez, I -Castañeda, MR	-Vallarta, MS -Guerra, F -Baños, A -Moshinsky, M -De Oyarzabal, J -Barajas, A -Graef, C
<i>Preferred journals</i>	-Trans Am Geophys Union -Trans Am Inst Mining Metall Eng -J Ind Eng Chem -Am Midland Nat	-Am Heart J -J Infect Dis -P Soc Exp Biol Med -J Cell Comp Physiol -Arch Biochem -Compt Rend Sci Soc Biol -J Am Med Assoc -J Am Chem Soc -Am J Hyg -J Exp Med	-Phys Rev -J Pharmacol Exp Ther -Am J Roentgenol Rad Ther -Arch Int Med
<i>Categories</i>	-Geochemistry & Geophysics -Engineering, Civil -Metallurgy & Metallurgical Engineering -Medicine, General & Internal -Ecology -Biodiversity Conservation -Chemistry, Applied	-Medicine, General & Internal -Medicine, Research & Experimental -Cardiac & Cardiovascular Systems -Infectious Diseases -Biochemistry & Molecular Biology -Immunology -Biology -Pathology	-Physics, Multidisciplinary -Pharmacology & Pharmacy -Multidisciplinary Sciences -Medicine, General & Internal -Psychiatry

		-Chemistry, Multidisciplinary	
Institutional structure	<ul style="list-style-type: none"> -Mining companies -Directorate of Geographic and Climatic Studies -Head office of Geography, Meteorology and Hydrology -Oil companies -Mutada Herbarium Escuincla 	<ul style="list-style-type: none"> -National Institute of Cardiology -Institute of Health and Tropical -Laboratories of Hormon -General Hospital -Institute of Hygiene - Syntex Research Laboratories -Hospital Laboratory America -Pediatrics Hospital 	<ul style="list-style-type: none"> -UNAM -UAL -AUP -IPN
Geographic (distribution by state)	<ul style="list-style-type: none"> -Mexico City: 53 % -Hidalgo: 7 % -Chiapas: 5 % -Coahuila: 5 % -Sonora: 5 % -Otros: 25 % 	<ul style="list-style-type: none"> -Mexico City: 97 % -Yucatán 3 % 	<ul style="list-style-type: none"> -Mexico City: 96% -Puebla 2% -Nuevo León 2%
Word frequency	<ul style="list-style-type: none"> Mexico, Geodetic, Mexican, Mining, Mine, Methods, Pachuca, Cyanide, Ore, Geology, Gravity-Station, Fresnillo, Oil, Plasmogenesis 	<ul style="list-style-type: none"> Typhus, Mexican, Heart, Steroids, Experimental, Guinea pigs, Mexico, Synthesis, Treatment, Rheumatic, Oils, Pigs, Turtle, Patients, Brucellosis 	<ul style="list-style-type: none"> Cosmic, Radiation, Gravitation, Fields, Energy, Magnetic, Spectrum, Rays, Birkhoff's, Particles

The amateur or individual mode is associated with poor, non-professional scientific practice, principally the result of the individual effort of the authors, accomplished locally in isolation from peers and with immediate utilitarian objectives in mind. Within the subject areas related mainly to the mining, oil, railway and health industrial sectors, a variety of different research topics were undertaken which showed little continuity and connection between authors. With the notable exception of A. Medina, the most productive researcher during this period, who published original results on the characteristics and progress of geodetic work in Mexico in the decade of the 30s, the majority of the authors published only one paper and the few who published three papers wrote about different subjects each time. Such is the case of E. Ordóñez who published on the Mexican railway system, the mines from the Pachuca district and oil in the south of the State of Tamaulipas. Another example is P.C. Sánchez who wrote about earthquakes, the history of geodesy in Mexico and about volcanoes.

This production mode replicates the publication patterns of a sole author which have no references and no citations. The papers are typically reports, classification studies, descriptions, technical research reports, implementations of solutions, clinical methods and studies, written in English by researchers such as engineers and physicians, affiliated mainly to foreign companies from the mining, oil, and railway sectors and to a lesser extent, from national institutes in the fields of geography, geology and private hospitals. The specialized journals used are in the subject categories of engineering: geochemistry, geophysics, civil, chemical, mining, and general, internal and clinical medicine, which are no longer covered by the SCI. An interesting feature of this mode of production is that, unlike the institutional and academic modes, research is much less centralized within the federal district of Mexico City.

Mode 2 of production corresponds to scientific practice which shares the characteristics and circumstance of local and external development of modes 1 and 3 throughout the period analyzed. Nonetheless, it also presents important differences with respect to communication, publication and citation patterns as shown in Table 1. This mode is replicated in the institutions of geology, geography, medicine, public health and hygiene, astronomy and biology, created by the state under varying names beginning in the 19th century and is manifest in the international, mainstream literature in a more marked and continuous way in the decade of the 1930s by the attention given to the topic of typhus as a particular local, public health problem by different specialists with M.D. degrees. Doctors, such as G. Varela, M.R. Castañeda, H. Mooser, J. Sozaya, and N.P. Hudson, affiliated to institutions of public

health and hygiene, such as the Institute of Hygiene, the Institute of Health and Tropical Diseases and the General Hospital which created a laboratory specialized in typhus research, as well private laboratories and hospitals. This environment created as a result of typhus research, produced the first instances of national and international scientific collaboration with respect to a local problem, in addition to the first Mexican papers to receive more than 10 or 20 citations.

Mode 2 emerged under different local social and political circumstances and with changes in the authors and in their professional profiles with respect to mode 1. In the decade of the 40s, new specialists joined the researchers studying typhus; among them were exiles from the Spanish Civil War and Mexican researchers who had trained abroad. For instance, I. Chávez, B. Sepúlveda, A. Rosenblueth, I. Costero, F. Giral, J. Romo, G. Haro, and F. Gómez, among others, attached to the Institutes of Cardiology, Pediatrics and institutions from the chemical and pharmaceutical industry, created new openings in the relationship with international science which translated into the consolidation of new patterns of scientific production and communication. These were associated with the adoption of scientific practices involving co-authorship with peers, in both national and international scenarios, diversification of languages and research methods, the presentation of papers in international meetings which were reviewed in mainstream journals, as well as the publication of short communications in the form of notes and letters.

In Mode 2 efforts are directed along different paths towards insertion of Mexican science into the international standards of publication and citation, manifested in accordance with Table 1, by an increasing average number of references cited in the papers accompanied by a rise in the impact of local studies through citations given by peers in the international scientific community. The average of 13 citations achieved per paper, is competitive with the present-day average numbers of citations for Mexican science. These characteristics are associated with changes in the objectives set by scientific practices, above and beyond utilitarian considerations of the results, as shown by basic research of a theoretical and experimental nature which has guaranteed the continuity of the new topics studied and has allowed access to different journals with the highest impact factors in their respective areas. Such is the case of the American Heart Journal, Ecology, Proceedings of the Society for Experimental Biology and Medicine, and the Journal of Cellular and Comparative Physiology where the first studies were published which received more than 50 citations and the Journal of the American Chemical Society with the highest average number of citations, 79 citations per paper and the paper with the highest number of citations, 154, for the period analyzed. Mode 2 was developed in a centralized manner in the federal district of Mexico City.

Mode 3 corresponds to professional research carried out within institutions and linked to academic topics and interests. This implies a greater freedom in the choice of research topics and greater independence from considerations related to the solution of local problems, than with respect to modes 1 and 2. Its origins lie in the incorporation of the Institutes of Geology, and of Biology, of the National School of Medicine, the National Astronomical Observatory, among others, to the National Autonomous University of Mexico (UNAM) in the period from 1930-1950. It initiated its development in a strongly centralized way, 90% concentrated in this institution, particularly in the Institute of Physics and within the federal district of Mexico City.

This academic mode is the least present and appears only at the end of the period analyzed. The production mode and communication patterns identified in Table 1, correspond principally to the physical sciences and, in particular to the area of Cosmic Rays, the theoretical research topic with most influence among physicists at that time. These scientists started disseminating their first studies internationally through their contributions to the Proceedings of the American Physical Society, principally in the meeting celebrated in 1947 in Houston, Texas and in 1950 in Mexico City. This communication channel guaranteed the publication of papers in Physical Reviews, the main international journal in the field, in the form of meeting abstracts. This document type represents 50% of all contributions made under mode 3 and are generally written by a sole author, references are not included and they are rarely cited. These conditions were a fundamental influence on the overall characteristics of the academic mode, identified in Table 1, which refer to publication patterns

exemplified by individual authorship, low average numbers of references and citations per paper and 70% of the production centralized in the journal Physical Reviews.

Discussion

The variables identified in our bibliometric analysis: author, document type, institutional affiliation, geographic location, number of references, number of citations, categories and title words, allowed us, through the application of desegregations techniques quantification and analysis of word frequency, to follow the development of Mexican science during the first half of the 20th century. We established a process of incorporation of the local community into the global scenario by adopting the international standard of publication in mainstream journals. This process evolved accompanying and documenting the construction of new ways for knowledge production that occurred and complemented each other throughout this period. In each of these stages favorable conditions were created for a closer relationship with international scientific culture.

The presence under local conditions of new scientific practices during the period analyzed, were observed using bibliometric methods involving the tracking and quantification of common biographical elements (authors, journals, topics, institutions, references, citations) for the identification of changes occurring in institutional structures, forms of organization and patterns of scientific communication. These biographical elements advise us of evolutionary pathways and other testimonial elements sufficient for the characterization of the development of Mexican science during this period, tracing its journey through three distinct stages for the generation and publication of knowledge in the international environment.

Using the institutional affiliations of authors in the SCI, it was possible to recover the scientific production written by external authors working in foreign companies based in Mexico, corresponding principally to mode 1 where research is carried out locally in mining companies as well as in the transport and health sectors. The bibliographic data associated with this production such as authors, titles, journals and topics, are difficult to document using historiographic methods and for this reason, are not found in the historical works which cover this period.

The general patterns of publication and citation of Mexican science today are defined by the practices of mode 2 scientific research, associated with problems of public health and hygiene, principally as a result of the line of research developed on the study of typhoid disease, by a main group of four scientists: H. Mooser, G. Varela, M.R. Castañeda and N.P. Hudson. Their work generated a body of literature with certain characteristics. Firstly, it represented the first topic in Mexican science whose study continued over a number of years, the 10 years from 1928 to 1938, in which several authors were involved, four principal ones and 15 co-authors. Secondly, 18 papers were published which were the first to cite an average of 10 references and with an average impact of 11 citations each. The papers were published in high impact journals which are still today covered by the SCI: Journal of Experimental Medicine, Journal of Infectious Diseases, Proceedings of the Society for Experimental Biology and Medicine, American Journal of Tropical Medicine. The publication and citation patterns of the scientific literature on typhoid disease are the clearest indication of the division of Mexican science during the 20th century, in terms of changes in communication patterns between the amateur and the professional modes. It is also the first example of the internationalization of a local topic which attains wide impact, carried out by Mexican researchers affiliated to institutions in the area of public health and using resources of these home institutions.

The evolutionary information collected in the present study indicate that the least productive state of Mexican science is associated with scientific practices adopted during times of greatest social and political instability, within a single production mode, carried out principally in amateur form, isolated from the influence of other scientific practices and aligned with utilitarian objectives, and representing the form which experiences the most difficulty in coinciding with the standards of international scientific publication. The phase of greatest growth, as the end of the period analyzed, is the result of changes in local conditions, and the convergence and complement of the institutional and academic modes, primarily.

Disciplines and professions exist which are linked with one mode of production, such is the case of physics, which arose with academic practice. Others such as health employ the three modes in its different areas of clinical research, public health and hygiene, internal, experimental and theoretical medicine. We consider that the quantitative historiographic approach used in the present study made possible by the online availability of mainstream production records from the beginning of the last century, has proved a valuable adjunct in the description of the development of Mexican science from 1900-1950 and particularly for the identification of the mechanisms that were instrumental to the integration of Mexico into international scientific community standards during that period.

References

- Académica Mexicana de Ciencias (2003). *Ciencia y Tecnología en México en el Siglo XX*. Biografías de personajes ilustres. México: Academia Mexicana de Ciencias.
- Beltrán, E. (1952). *Medio siglo de ciencia en México. 1900-1950*. México: SEP.
- Beltrán, E. (1970). Fuentes mexicanas de historia de la ciencia. *Anales de la Sociedad Mexicana de Historia en Ciencia y Tecnología*, 2, 57-112.
- Beltrán, E. (1989). La historia de la ciencia en México en los último cinco lustros (1963-1988). In *Memorias del Primer Congreso Mexicano de Historia de la Ciencia y la Tecnología* (México, DF, 7-30 septiembre, 1988). Tomo I: 79-100.
- Casas, R. (2003). Los estudios sociales de la ciencia y la tecnología: enfoques, problemas y temas para una agenda de investigación. In *Perspectivas y desafíos de la educación, la ciencia y la tecnología* (pp. 139-195). México: UNAM.
- Christie, J.R.R. (2005). El desarrollo de la historiografía de la ciencia. In S.F. Martínez, & G. Guillaumin (comps.), *Historia, filosofía y enseñanza de la ciencia* (pp. 43-65). México: UNAM. Instituto de Investigaciones Filosóficas.
- Coordinación de la Investigación Científica (2003). *Forjadores de la ciencia en la UNAM: conferencias del ciclo mi vida en la ciencia, mayo-agosto de 2003*. México: UNAM. CIC.
- Edge, D. (1979). Quantitative measures of communication in science: a critical review. *History of Science*, 17, 102-134.
- García, A; Pérez, M.A. (2006). High Energy Physics in Mexico: Historical Sketch and Implications. In M.A. Pérez, L.F. Urrutia, L. Villaseñor (editors). *Particles and Fields* (pp. 1-10). Melville: American Institute of Physics. (AIP Conference Proceedings; vol. 857 Part B)
- Gortari, Eli de (1963). *La ciencia en la historia de México*. México: FCE.
- Kragh, H. (1987). *An introduction to the historiography of science*. UK: Cambridge University Press.
- Laudan, R. (2005). La nueva historia de la ciencia: aplicaciones para la filosofía de la ciencia. In S.F. Martínez, & G. Guillaumin (comps.), *Historia, filosofía y enseñanza de la ciencia* (pp. 43-65). México: UNAM. Instituto de Investigaciones Filosóficas.
- Moreno, R. (1986). *Ensayos de historia de las ciencia y la tecnología en México*. México: UNAM. Serie. Historia de la Ciencia y la Tecnología; 2.
- Okubo, Y. (1997). *Bibliometric indicators and analysis of research systems: methods and examples*. OECD Science, Technology and Industry Working Papers. OECD Publishing. Retrieved October 21, 2006, from Doi: 10.1787/208277770603 site: http://econpapers.repec.org/paper/oecstiaaa/1997_2F1-en.htm.
- Pérez-Angón, M.A. (2006). Atlas de la Ciencia Mexicana. Available in: <http://www.amc.edu.mx/atlas.htm>, (October, 2006).
- Pérez-Tamayo, R. (2005). *Historia general de la ciencia en México en el siglo XX*. México: FCE. Colección obras de ciencia y tecnología.
- Saldaña, J.J. (1982). Primer Doctor en Historia de la Ciencia y la Tecnología, creó en 1982 la Sociedad Latinoamericana de Historia de la Ciencia y la Tecnología. *Quipu*, 11(2), 128-130.
- Saldaña, J.J. y Azuela, L.F. (1994). De amateurs a profesionales. Las sociedades científicas en México en el siglo XIX. *Quipu*, 11(2), 135-172.
- Saldaña, J.J. (1994). El sector externo y la ciencia nacional: el conservacionismo en México (1934-1952). *Quipu*, 11(2), 195-217.
- Shapin, S. (1992). Discipline and Boundaries: The History and Sociology of Sciences as seen through the Externalism-Internalism Debate. *History of Science*, 30, 333-369.
- Suárez, E. (2005). La historiografía de la ciencia. In S.F. Martínez, & G. Guillaumin (comps.), *Historia, filosofía y enseñanza de la ciencia* (pp. 19-30). UNAM. Instituto de Investigaciones Filosóficas.
- Trabulse, E. (1983). *Historia de la Ciencia en México*. México: CONACYT: FCE.
- Trabulse, E. (1996). *El círculo roto: estudios históricos sobre la ciencia en México* (3ra ed). México: FCE.
- Trabulse, E. (2003, July). Tradición y ruptura en la ciencia mexicana. In *Proceedings of the International Congress of History of Science XXI*. México: SMHCT: UNAM.