Matching bibliometric data from Scopus with National Databases of Colombian Scientists (ScienTI Col)

Edwin Bernal[1] and Javier Guerrero C.[2]

[1] *ebernal@ocyt.org.co*
Colombian Observatory of Science and Technology, Bogotá D.C (Colombia) Carrera 15 N 37-59

[2] *jeguerreroc@ocyt.org.co*
Colombian Observatory of Science and Technology, Bogotá D.C (Colombia) Carrera 15 N 37-59

Introduction

In this poster we present a first approach to a methodology developed at the Colombian Observatory of Science and Technology in order to validate and accredit articles extracted from Scopus. The Colombian Observatory of Science and Technology (OCyT) publish an annual book of S&T indicators; one of the main problems in the construction of macro/meso/micro indicators of bibliometric production in Colombia is the incorrect attribution to papers, either to the country, city, institution or individual author. In comparing two sets of data, the output of the search from Scopus and the information included in the national databases of scientific information *ScienTI Col*, we developed a methodology to accurately attribute the articles at various levels. There are several methodologies in order to attribute articles at the country or institution level (Egghe, Rousseau, Van Hooydonk, 2000; Nedehof, Moed, 1993), as far as we are not using bibliometric output as an evaluative tool, but a strategy to locate national and institutional productivity in international databases we adopt the total count approach.

Colombian research articles in journals indexed in Scopus

The amount of Colombian research articles (documents signed by an author who in the moment of publication of the article works for a Colombian institution) has growth continually since 2001, as presented in Figure 1. The output was obtained using the advance search tool in the Scopus web service. The search was done using the field AFFILCOUNTRY (Colombia) for the years 2000-2009. The search output was exported to an excel worksheet and a manual approach was carried in order to determine preliminary problems associated with the accuracy of the data. 12.596 research articles associated to Colombian institutions were found, after a manual check 12.264 research articles were left, 15 were exclude due to duplications and 317 due to misassignation.

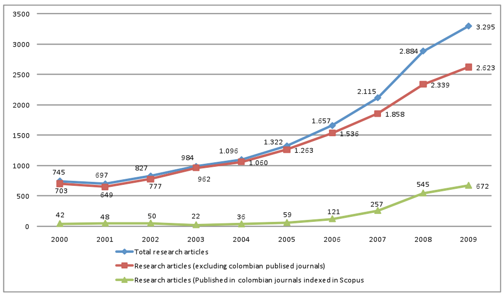

Figure 1. Amount of Colombian research articles in journals indexed in Scopus, 2000-2009

The national database for the management of scientific information: ScienTI Col

The OCyT receives annually from The Administrative Department of Science, Technology and Innovation (Colciencias) information from the Database ScienTI Col. This database contains information about the activity of individual researchers (CvLAC), research groups (GrupLAC) and Colombian S&T Institutions (InstituLAC), Figure 2.

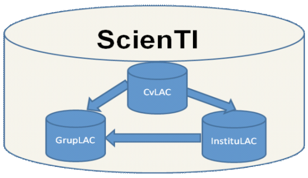

Figure 2. ScienTI Col Model, assignation of institutional affiliations.

For the time lapse 2000-2010 there are 168.289[42] research articles registered in ScienTI, both in national and international journals. Those are our matching universe. Many local articles registered in ScienTI are not covered in international Databases, and cannot be matched; nevertheless, as the information provide in ScienTI is used in Colombia to evaluate the performance of research teams, it is expected that local authors reports publication information as accurate as possible.

Data processing

To process the data we use free software. As a programming language we use PHP. And for data base management we use MySQL. We construct an algorithm in order to separate the data corresponding to authors and institutional affiliation. Then we define four models to match the two data sets and maximising the accuracy of the results.

Algorithms

To separate authors we developed and algorithm, figure 3, which establish the amount of authors per article and extract the data of each one. A plain text is produced separating each author per column:

```
Load data from Scopus
Convert archive to matrix
Determine archive EOF
To x ← 1 to x ← EOF do
          Eliminate spaces
          Convert text to capital
          Replace accents
          Replace strange characters
 End to
To z ← 1 to z ← EOF do
     Yes (authors =1) then
          Create authors_columns vector
          End if
          If (authors > 1) then
                Count authors = author count vector author separated in z
                To t ← 1 to t ← count author do
                     Find position author t+1
                     Create vector author_column in t
                End to
          End if
End to
To u ←1 to u ←EOF do
       Generate line plain archive plain vector authors column_vector
       Generate line jump
End to
```

Figure 3. Author separation pseudocode

An algorithm was developed in order to match articles from the Scopus data set and the articles reported by the researchers in CvLAC. The algorithm check, source title, article title, source ISSN, and publication year, we exclude author names due to normalization problems[43].

```
stablish conection with the database
select all scopus registers
while x ← eof do
     extract issn, year & title
     select scienti registers with same isss to scopus and same year
     while and ← eof2 do
          extract scienti, issn, year, title person id
          compare scopus title vs scienti title
          if (percentage >= 70%) so
               actualize scienti table with coincidence id
          end if
      end while
end while
```

Figure 4. Matching pseudocode algorithm

Results:

We run the algorithms with four different restrictions, the approach models and results are summarized in table 1.

Restriction / % Similarity	Match cases	% of match
Soucer title 80%, article title 70%	305	2,49
ISSN 100%, article title 70%	5.523	45,03
Pub year 100%, article title 70%	6.569	53,56
Article title 70%	6.646	54,19
Total match cases	**7.503**	**61,18**

Table 1. Run models, restrictions and output

We were able to identify 7.503 research articles published in journal indexed in Scopus and registered in ScienTI, and

[42] Every author registers his production. So this data contains duplications due to the times that coauthors register the same articles.

[43] The OCyT is developing a project of normalization of the author and institutions names.

using the identification number of the research articles in the ScienTI Col database we could confirm the institutional affiliation of the researcher.

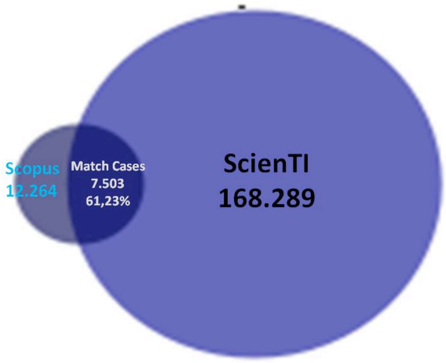

Figure 5. Match cases between the national database and Scopus output

Conclusions

The methodology seems promissory in order to confirm and to correctly attribute papers to Colombian institutions. Indentifying the authors and through the affiliation of the research group in the ScienTI DB we are able to assign institutional affiliations. Nevertheless, the results open up the question concerning the data of the remaining 39% of articles, are those articles included in the national database (ScieTI Col)? How to develop restrictions in order to gain accuracy in the representation of the Colombian research output as represented in the national? Currently we are running different models, matching the bibliometric output of the research groups only (GrupLAC).

References

Egghe, L; Rousseau, R & Van Hooydonk, G. (2000). Methods for Accrediting Publications to Authors or Countries: Consequences for Evaluation Studies. *Journal of the American Society for Information Science*, 51, 145-157.

Nederhof. A.J., & Moed, H.E. (1993). Modeling multinational publication: development of an on-line fractionation approach to measure naiional scientific output. Scientometrics, 27, 39-52.