Extracting and sharing data citations from Google Scholar for collaborative exploitation

Sibele Fausto*, Tiago Rodrigo Marçal Murakami

1sfausto@usp.br
Escola de Comunicações e Artes - University of São Paulo,
Av. Prof. Lucio M. Rodrigues, 443, São Paulo, SP, CEP 05608-020 (Brazil)

2tiago.murakami@fht.xbl.usp.br
Departamento Técnico, Sistema Integrado de Bibliotecas, University of São Paulo,
Rua da Biblioteca, S/N, Complexo Brasilia, Rio Embarcamento, São Paulo, SP, CEP 05508-050 (Brazil)

Background and purpose

There are studies that have drawn attention to the lack of indexing for the titles of scientific journals in the Social Sciences, Applied Social Sciences and Humanities in large commercial databases (Frandsen & Nicolaisen, 2008; Neuhaus & Daniel, 2007). This lack is even more acute when it comes to journals concerned with these areas published in languages other than English and published in developing countries (Archambault & Lari-viére, 2010), which makes it difficult to carry out an investigation of the importance and impact of these journals.

This situation is changing as a result of the new opportunities provided by the emergence of Open Access (OA) and tools as the search engine Google Scholar (GS) and software for data processing such as Publish or Perish - PoP (Harzing, 2007). The increasing shift of Social Sciences and Humanities journals to the Web – including those of Library and Information Science (LIS) is making them more widespread. This is allowing detailed searches to be conducted through GS and the recovery of citations of articles, which can be regarded as an alternative to traditional databases in bibliometrics studies on the impact of scientific production published in these areas. In addition it highlights the fact that GS is a free access source, in contrast with expensive commercial databases. It has a broad coverage of other kinds of material, even in the Social Sciences and Humanities (SSH), such as books, book chapters, conference materials, etc. which are not normally covered by traditional databases and hence it is able to make a comprehensive recovery of open access journals, in languages other than English, some of which come from emerging countries.

However, this apparently favorable context for research into bibliometrics in these areas still faces challenges owing to questions about the reliability of the GS as a data source (Jacsó, 2010). This criticism regarding to GS is a restatement of the need for more research into the tool to finds a rational basis for understanding the full potential of Google Scholar, by various categorizations such as the received citations for each journal, including citations per year and the articles cited, among others. These preliminary exercises were also publicly shared through the following link:

http://public.tableausoftware.com/views/EstudosdeCitacoesrecebidasporperiodicos/ContasCitadasreceivebidosperiodicos?embed=y&display_count=no, e.g. as shown in Figure 1.

Preliminary findings

Citation studies are an important subject research in Bibliometrics and their sources of reliable data were, until recently, a prerogative of restrictive and expensive commercial databases, despite these sources still continue to show inconsistencies as is widely discussed in the literature. Google Scholar provides an alternative source to these studies, particularly in the areas of the SSH, where many journals are not considered by the large databases.

The emergence of tools that facilitate the extraction and data processing from GS, such as PoP and tools like Google Refine, Google Drive and Tableau Public help to simplify the task of validating these data. In our view, the public sharing of pretreated citation data can stimulate more collaborative investigations by the community of Brazilian scientometricians with the aim to demonstrate the capacity of Google Scholar to act as an alternative and reliable data source in the metrical studies of national journals and thus enable better measures of the SSH results in the context of scientific evaluation in Brazil.

Methods

This pilot scheme adopted the following procedures:

a. Conducting a survey of LIS journals titles through compiling lists of those that exist on the web;

b. Carrying out searches using PoP software for Windows, with the journal title as a parameter, and confirming the official titles and abbreviations, in the period from January 28, 2014 to March 02, 2014;

c. Displaying the results in Google Drive spreadsheets, one for each retrieved journal title;

d. Creating a spreadsheet that brings together all the spreadsheets with the articles that had at least one citation;

e. Carrying out statistical tests using Excel and Tableau Public.

Google Drive allows its contents to be shared publicly, and the extracted data to be made available through the following link:

https://docs.google.com/spreadsheets/d/19kxmMnflSOHe60_mey-mycG5Stkpe8y-HHxqfGQ/sheet

Figure 1. Number of Citations per journal and per year

Data extraction from the GS with PoP resulted in a total of 24 Brazilian LIS journals, all in open access. However, the searches recovered some inaccurate data which were then analyzed by article and those with inconsistencies were withdrawn. The data obtained allowed some exploratory exercises to be conducted with Tableau Public, by various categorizations such as the received citations for each journal, including citations per year and the articles cited, among others.

References


