Scientific journal publishers and omitted citations in bibliometric databases: Any relationship?

Fiorenzo Franceschini *, Domenico Maisano, Luca Mastrogiacomo

Politecnico di Torino, DIGEP (Department of Management and Production Engineering), Corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

ABSTRACT

Omitted citations – i.e., missing links between a cited paper and the corresponding citing papers – are the main consequence of several bibliometric database errors.

This paper investigates the possible relationship between omitted citations and publishers of the relevant citing papers. This relationship is potentially meaningful because: (i) publishers generally impose editorial styles, which could affect database errors, and (ii) some publishers may be more efficient than others in detecting and correcting pre-existing errors in the manuscripts to be published, reducing the risk of database errors.

Based on an extensive sample of scientific papers in the Manufacturing Engineering field, this study examines the citations omitted by the Scopus and WoS databases, using a recent automated algorithm. Major results are that: (i) there are significant differences in terms of omitted-citation rate between publishers and (ii) the omitted-citation rates of publishers may vary depending on the database in use.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction and literature review

Bibliometric databases, like any database, are not free from errors. Despite the improved accuracy over the past ten years – probably due to the systematic employ of automatic tools for correcting errors in cited article lists by editors and database administrators (Adam, 2002) – the problem is far from being solved. This is proven by (i) several recent articles documenting the existence of errors of different nature – e.g., (Franceschini & Maisano, 2011a; Jacsó, 2012) – and (ii) the fact that bibliometric database staff constantly encourage users to report any noticed inaccuracy.

A synthetic classification of the major database errors is reported in Table 1, distinguishing between authors' and database mapping errors. The contributions by Buchanan (2006), Jacsó (2006) and Li, Burnham, Lemley, and Britton (2010) and Olensky (2013) show that one of the main consequences of these errors is represented by omitted citations, i.e., citations that should be ascribed to a certain (cited) paper but, for some reason, are lost. In other terms, the link between citing and cited article is not established by the database.

Even though the scientific literature reports numerous notifications of blunders (sometimes grotesque!) by Google Scholar (Labbé, 2010), it often ignores the citations omitted by the two major multidisciplinary databases: Scopus and Web of Science (WoS). According to the study by Buchanan (2006), citations omitted by WoS are likely to be around 5–10%

* Corresponding author. Tel.: +39 0907225.
   E-mail addresses: fiorenzo.franceschini@polito.it (F. Franceschini), domenico.maisano@polito.it (D. Maisano), luca.mastrogiacomo@polito.it
(L. Mastrogiacomo).

Table 1

Classification of bibliometric database errors according to Buchanan (2006).

Error type	Author errors	Database mapping errors
Definition	Errors made by authors when creating the list of cited articles for their publication	Failure to establish an electronic link between a cited article and the corresponding citing articles that can be attributed to a data-entry error
Examples	- Errors in name and initials of the first author - Errors in publication title - Errors in publication year - Errors in volume number - Errors in pagination.	- Transcription errors - Target-source article record errors - Cited article omitted from a cited-article list - Reason unknown

of the "true" number of citations (i.e., the number of citations that would be indexed in the ideal case of absence of omitted citations).

Unfortunately, most of the contributions on database errors rely on the manual analysis of small samples of scientific articles and therefore results are not very robust statistically. To overcome this obstacle, Franceschini, Maisano, and Mastrogiacomo (2013) introduced an algorithm for estimating a database's omitted-citation rate automatically. This algorithm requires the combined use of two or more bibliometric databases and is based upon the hypothesis that the mismatch between the citations occurring in one database and another one is evidence of possible errors/omissions.

The automated algorithm has been recently applied by Franceschini, Maisano, and Mastrogiacomo (2014) to a large set of journals in the Manufacturing field, showing that – with a few exceptions – the differences in terms of omitted-citation rate (p) between these journals is included between 4% and 10% for WoS and between 2% and 8% for Scopus. The same study showed that omitted citations are usually "isolated accidents", which concern a very small portion of the articles published by a certain journal; nevertheless, these errors may affect indicators based on citation statistics significantly.

Going back to the classification in Table 1, it is not unreasonable to guess that some editorial styles imposed by certain publishers could hamper the correct identification or the citing/cited papers by a database. For example, citing papers containing lists of references with (i) abbreviated journal titles (ISO 4:1997, 1997; Thomson Reuters, 2014a), (ii) first authors' names only, or (iii) omitted paper titles (not unusual for journals in the Physics field) could complicate the identification of cited papers.

Apart from the list of references, other features concerning publishers could affect the database propensity to omit citations, such as: the type of data made available (PDF or HTML version of the articles) and the ability to detect and correct pre-existing errors in the cited article list of a manuscript, before publication. We remark that database errors often result from pre-existing author errors, which are unnoticed throughout the production/indexing process of a paper. In the best cases, reviewers, publishers or even database staff (in chronological sequence) are able to detect and correct them. In the worst cases, they may ignore these errors and even generate new ones.

The objective of this paper is to investigate the possible relationship between omitted citations and publishers of the relevant citing papers (e.g., Elsevier, Springer, Taylor & Francis, etc.). The analysis relies on the same dataset used in (Franceschini et al., 2014) – which concern (cited) papers in the field of Manufacturing Engineering – and focuses the attention on the publishers of the citing articles; omitted citations are examined from the perspective of both Scopus and WoS.

The remaining of the paper is organized into four sections. Section 2 recalls the automated algorithm for analysing omitted citations. Section 3 provides a detailed illustration of the methodology of data collection and analysis. Section 4 presents the analysis results. The concluding section summarizes the original contributions of this paper, focusing the attention on the relevant implications, limitations and ideas for future research.

2. Automated algorithm for analysing the omitted citations

Before recalling the algorithm, we present an introductory example to illustrate how it works. Let us consider a fictitious paper of interest indexed by Scopus and WoS. The number of citations received by this paper is 13 in Scopus and 12 in WoS (see Table 2).

The union of the citations recorded by the two databases is a total of nineteen citations. Among the citing articles, only nine belong to sources (i.e., journals or conference proceedings) officially covered by both databases (highlighted in grey in Table 2). Focusing on these nine "theoretically overlapping" (TO) citing articles, two are omitted by Scopus (but not by WoS) and one is omitted by WoS (but not by Scopus). Therefore, from the perspective of the paper of interest, a rough estimate of the omitted-citation rate is $2/9 \approx 22.2\%$ in Scopus and $1/9 \approx 11.1\%$ in WoS. The same reasoning can be extended to multiple papers of interest and more than two bibliometric databases.

Let us now focus attention on the automated algorithm, which is based on the combined use of two bibliometric databases (Scopus and WoS in this case) and can be summarized in three steps:

1. Identify a set of (P) papers of interest, indexed by both the databases.