"The Delphic Oracle" - An Analysis of Potential Error Sources in Bibliographic Databases

Dirk Tunger*1, Stefanie Haustein1,2, Lena Ruppert3, Gaetano Luca2, Simon Unterhalt2

1{d.tunger, s.haustein}@fz-juelich.de
Forschungszentrum Jülich, Central Library
52425 Jülich (Germany)

2Heinrich-Heine-University Düsseldorf, Department of Information Science
Universitätsstraße 1, 40225 Düsseldorf (Germany)

3Cologne University of Applied Sciences, Institute of Information Science
Claudiusstraße 1, 50678 Cologne (Germany)

Introduction
To an increasing extent, decisions with far-reaching consequences, such as the funding of scientific research, are being made on the basis of bibliometric data. Data from one of the two big scientific databases, "Science Citation Index" or "Scopus", usually forms the basis for such bibliometric evaluations. In view of the far-reaching consequences of the decisions the question arises of how reliable the original data sources are.

Method
In order to give a concrete answer to this question, the authors performed a study based on publications in 8 physics journals in the year 2007. Due to the considerable manual effort involved in data cleansing, only a sample of this size could be processed in the study.

Table 1. Share of erroneous articles and lost citations for a sample of eight physics journals

<table>
<thead>
<tr>
<th>Journal of High Energy Physics</th>
<th>Number of publications in 2007</th>
<th>Share of erroneous articles</th>
<th>Share of lost citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature Physics</td>
<td>307</td>
<td>8.0%</td>
<td>5.2%</td>
</tr>
<tr>
<td>Physical Review A</td>
<td>2356</td>
<td>7.9%</td>
<td>13.7%</td>
</tr>
<tr>
<td>Physical Review Letters</td>
<td>3817</td>
<td>6.1%</td>
<td>19.2%</td>
</tr>
<tr>
<td>Physics Letters B</td>
<td>861</td>
<td>8.0%</td>
<td>7.2%</td>
</tr>
<tr>
<td>Physics of Plasma</td>
<td>834</td>
<td>3.7%</td>
<td>16.4%</td>
</tr>
<tr>
<td>Reports on Progress in Physics</td>
<td>40</td>
<td>2.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Reviews of Modern Physics</td>
<td>34</td>
<td>0.0%</td>
<td>13.3%</td>
</tr>
</tbody>
</table>

The figures show the total number of documents from 2007, the percentage of documents with at least one error in the bibliographic data and the percentage of lost citations in the respective journal. The evaluation is subdivided into two phases. In the first phase the bibliographic data of the original publications are compared with the corresponding entry in the Science Citation Index. In the second phase, a search is then made for any lost citations with the aid of the cited reference search.
Results
The results of the two evaluation phases provide material for discussion since the error rate for the bibliographic data is almost 7% and for the citations about 15%! The error rate for the individual journals varies quite considerably in both phases so that possible disadvantages may depend on the journal used. However, the error rate does not depend on the size of the journal. With respect to the bibliographic data, the major error with 384 occurrences in a total of 651 erroneous documents is related to "Chinese given names". Instead of one initial for the given name, which would be logical according to the style in the original articles, in the Science Citation Index in some cases two initials are used for the given name (thus, for example, in the Science Citation Index Zhang, Hongbao becomes Zhang, HB, instead of Zhang, H, as would be expected). This means that in a search using the author's name with just one initial the document would not be found since the author's name is indexed with two initials. Other very frequent errors were that the authors' names and affiliations were missing altogether or spelt incorrectly. The nature of the misspellings (e.g. rn instead of m) lead to the conclusion that these are OCR errors. Errors of this type mean that publications are swallowed up by the database and cannot be taken into consideration for bibliometric analyses if the search uses the names of these persons or their institutions. However, errors do not exclusively occur on the bibliographic side, citations are also lost. A citation is created by matching a publication entry with the footnotes of all database entries. Each footnote that includes a publication is a citation. The problem is to be found in the matching. If the name of the lead author or the journal is misspelt then this is one reason for lost citations as is the switch from two- or three-figure page numbers to six-figure article numbers and the associated notation errors in citing the documents.

Discussion
This paper provides an impetus for a very important discussion. How valid are the big citation databases, where are the problems in the production workflow of the database providers and approximately what level of error must be taken into account in interpretations? Only when these parameters have been clarified will it be possible to ensure that bibliometric data are handled correctly. It is therefore certainly worth considering how the community can work together with the database providers in addressing this problem.

References