How is Bibliographic Data Accuracy Assessed?

Marlies Olensky

marlies.olensky@ibi.hu-berlin.de

Berlin School of Library and Information Science, Humboldt-Universität zu Berlin, Unter den Linden 6, Berlin, 10099 (Germany)

Abstract
This preliminary study presents a state-of-the-art analysis of bibliographic data accuracy assessment. It addresses the problem of inaccurate data in bibliographic data sources. 98 studies investigating the accuracy of references and citation links, or evaluating the data accuracy of bibliometric data sources (or both) are examined in terms of selection of data samples, methods, consulted data sources and error categorization. The analysis of applied methods shows a high tendency towards using the original publication as gold standard for verification. To assess the errors in bibliometric data sources three of the studies apply matchkeys. The examination of the error categorization demonstrates that bibliographic data accuracy is assessed by author name(s), journal title, volume, year and pagination, whereof journal title, volume and year have the highest impact. However, the actual categorization and the level of detail diverge among the studies. Thus, future research needs to be conducted on standardizing error categorization and method(s) to assess bibliographic data accuracy.

Introduction
Reference lists are the raw material for carrying out citation analyses (Garfield, 1972; MacRoberts & MacRoberts, 1989; Dinkel, 2011). These bibliographic data are comprised in bibliographic databases used as bibliometric data sources. In spite of applied matching algorithms to unify references and link source articles with citing articles, errors in the main bibliographic data fields (author name(s), article title, journal title, volume, year, and pagination) still occur (Jacső, 2005). The primary bibliometric data sources are Web of Science (WoS), Scopus and Google Scholar (GS), which means that inaccurate reference lists and in succession the remaining erroneous records in these data sources influence the results of bibliometric calculations. Although these data sources apply data cleaning processes, not all of the errors made by authors can be corrected. Besides, not only authors of inaccurate reference lists are responsible for incorrect data in bibliometric data sources, but the data sources themselves can contain errors (e.g. Garfield, 1974; Hildebrandt & Larsen, 2008).

And although sophisticated algorithms for matching citing and cited reference have been developed in recent years, errors still occur (Neuhaus & Daniel, 2008). In discussions with the author of this paper, it was claimed that data quality problems including bibliographic accuracy in bibliometric data sources have been solved by algorithms in several applied bibliometrics research groups - but that yet, the solutions have not been published to keep strategic advantages. While this may be true it is maintained that in terms of good scientific practice this can only be considered a claim and not as a scientific result unless the methods are published or at least their effectiveness demonstrated in experiments open for scientific access and assessment (the latter doesn't require publication of the actual algorithms and thus would not annihilate strategic advantages).
The implications of erroneous reference lists and bibliometric data sources for bibliometric research are obvious: data collection procedures must be accurate and must properly match cited references to target and source articles (Moed, 2005). In their daily work, researchers, particularly bibliometricians, and the scholarly community would benefit from knowing what the actual data quality and consequently the error rates of the data, they are dealing with, are. Thus, they could fine-tune data cleaning processes in order to achieve better quality and more accurate results.

This preliminary study is part of a dissertation project contributing to research on data quality in bibliometric data sources. It compares studies which investigated data accuracy (as one aspect of data quality) of bibliographic data regarding their selected data samples, methods and in particular their categorization of errors. It aims at summarizing the state-of-the-art of bibliographic data accuracy assessment. In the long term the dissertation project will investigate how data quality in bibliometric data sources can be assessed, what error categorization should be employed and investigate the implications on bibliometric calculations.

**Data quality and bibliometric data sources**

This section will shortly explain the definitions of data quality and bibliometric data sources. The ISO 9000 definition states quality as the “totality of features and characteristics of a product, process or service that bears on its ability to satisfy stated or implicit needs” (ISO 2005). In the 1990s, Redman (1996) developed a system-centered framework that considers the dimensions of data quality according to the three aspects of data: data modeling, data values and data representation. This framework can be applied to a variety of databases. Determining the quality of data values is the first step towards assessing data quality in bibliometric data sources. According to Redman (1996) the dimensions to be assessed are: accuracy, completeness, currency and consistency. To the best of the author’s knowledge only studies assessing data accuracy have previously been carried out and therefore are the subject of this work. Other aspects of data quality assessment in bibliometric data sources comprise, amongst others, coverage, duplicate records, author disambiguation (Franceschini & Maisano, 2011).

Moed (2005, p. 174) defines a bibliometric database as primarily intended for large quantitative statistical analyses and mappings, whereas bibliographic databases merely provide access to a collection of bibliographic records of all kinds of published literature. For this work, the term bibliometric data source will be used instead, as it implies bibliographic databases, bibliometric databases and any other source that could be used for bibliometric calculations.

**Method**

Wagner and Middleton (2008) assessed the effects of technical editing and the reference accuracy in peer-reviewed biomedical journals. Their report provided the basis for the evaluation aspects considered in this study. Taking their review as a starting point, a comprehensive list of studies that either investigated the accuracy of references, or evaluated the data accuracy of bibliometric data sources, or both, was obtained. The main bibliometric data sources WoS, Scopus and GS were searched and cited references of the found articles were explored. 98 studies were selected according to the following criteria: the main goal of the study had to be the investigation of bibliographic data accuracy, it had to be published in English and the full text had to be obtainable online, via library or via interlibrary loan.
At first ten papers of the 98 were randomly selected in order to set up the basic structure of the evaluation form. The studies were then reviewed regarding their selection of data samples, methods, consulted data sources and error categorizations. The structure had to be altered marginally in course of the evaluation in order to represent all studies in a correct way. The main aspects of evaluation were: main goal of study; subject area; data sources; number of journals investigated; number, publication type and year of citing source articles; number and publication type of cited articles; random selection of the data sample; method; error categories. The results were statistically analysed and are presented in the following section.

Findings

General Statistics
The study includes papers published in a 34-year-period from 1977 to 2011. Figure 1 displays the exact distribution of the studies over this period. The peak in 1995 is explained by a group of authors that published a lot of similar studies in that year (examining one journal per published study). It can be observed that 51% of the studies were published in the last third of this period (in or after 2000). The main goal of each study was the investigation of reference accuracy. 13 studies also examined the quotation accuracy of authors. The studies were mainly published in biomedical journals, only 16% in a library and information science related journal (or book chapter). 71% of the publications were articles or proceedings papers. 21% were letters, correspondences, editorial material or notes. The rest consisted of two poster presentation at a conference, two other presentations, one short survey, one master thesis and one book chapter.

Methods and data samples
The studies investigated mainly reference accuracy in biomedical literature. 79% of the studies dealt with medical related journals and 7% with reference accuracy in natural sciences (chemistry, physics, etc.). Four studies investigated the accuracy in library and information science journals and only three studies considered different subject areas in order to compare them.
92% of the studies indicated their main goal is to determine the error rate in reference lists. The other 8% evaluated the data accuracy in a specific database. 80% of the studies checked the references against the original publication. Four of the database studies also used the original as source for verification. 11% checked the original only in case the first data source they used for verification showed a discrepancy. The remaining 9% did not consult the original source at all. Therefore in total 90% of the studies referred to the original publication for verification and used it as a gold standard. Five of the database studies used the search functions of the databases in order to find and evaluate source or cited articles. The other three used matchkeys for finding discrepancies in the records. 57% of the studies specified they consulted only one data source to check the references, whereas 28% used two and 14% more than two. The data sources employed for verification, as well as a comparison of primary and secondary data source are listed in table 1. Due to the high tendency to biomedical literature it is obvious that MEDLINE/PubMed is the second most common data source with 28%. WoS shares rank four with other indexes at 12%.

Table 1. Use of data sources.

<table>
<thead>
<tr>
<th>Data sources (multiple entries)</th>
<th>total</th>
<th>as primary data source</th>
<th>as secondary data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>original publication</td>
<td>90%</td>
<td>73%</td>
<td>7%</td>
</tr>
<tr>
<td>Medline/Pubmed</td>
<td>28%</td>
<td>16%</td>
<td>10%</td>
</tr>
<tr>
<td>other databases (e.g SCIFinder, PsycINFO)</td>
<td>13%</td>
<td>--</td>
<td>13%</td>
</tr>
<tr>
<td>Sciene Citation Index (SCI), WoS</td>
<td>12%</td>
<td>9%</td>
<td>2%</td>
</tr>
<tr>
<td>other indexes</td>
<td>12%</td>
<td>--</td>
<td>10%</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>1%</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Internet</td>
<td>1%</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

35% of the studies investigated the accuracy of references in only one journal followed by 18% that examined three journals. In total 83% of the studies investigated one to five journals. The top number of investigated journals was one study with 34 journals and one study investigated the entire content of the CD-ROM version of the SCI (Science Citation Index) 1980-2004. In contrast, 49% of the studies chose “major journals” in the specific subject area.

Only 35% of the studies mention a range of 15 to 9496 citing source articles. Nine of these studies state they had randomly selected these articles. The publication types of the citing source articles were mainly articles (71%). 67% of the studies used articles from only one publication year; 20% compared two years, whereas 8% investigated a longer period of time. Of the 66 studies considering only one year, 35% just selected one specific issue for their investigation. 56% of the studies indicated the total number of cited articles, of which 62% randomly selected the actual data sample of cited articles. For those studies the author calculated the ratio of the data sample actually used. The calculation reveals 82% used a sample sized between 0.4% and 10%. 53% chose a data sample of 4% or less.

For all studies the size of the actually investigated data samples ranges from 30 to 22 Mio cited articles. 14% of the studies investigated a data sample between 30 and 100 cited articles. 26% of the studies indicated a sample size of 101 to 200 cited articles. In total, 64% of the studies randomly selected their data sample. 44% of those studies reduced the data sample due to non-
verifiable articles and did not replace them with others, whereas 13% replaced them. The remaining 44% investigated all cited articles available from the source articles. 56% of the studies investigated articles only, 22% also included other publication types like reviews, letters and books; the remaining 21% did not specify the publication types of the cited articles.

**Error categorization**

The statistical evaluation of error categorization proved to be quite difficult. The majority of the studies examined all six bibliographic fields for their accuracy: author name(s), article title, journal title, volume, year and pagination. Yet, they divided the error categories differently which will be discussed in this section separately for the studies evaluating the accuracy in databases and those investigating the reference accuracy in specific journals. 81% of all studies recorded multiple errors, 10% only logged the first error and then categorized the record as incorrect and for 9% the author was not able to extract this information from the text. In 7% of the studies, more than one researcher investigated the errors and an interrater-agreement was calculated. 31% compared the different error rates of years or journals and tested them on statistical significance. In most cases, it was unclear if the error categorization was designed prior to the evaluation or errors were grouped together afterwards.

Considering the eight studies evaluating the accuracy of bibliographic data sources, all of them investigated the six bibliographic data fields. Three of these studies (employing the matchkey method) categorized the errors in greater detail than the others (study A: 22, study B: 19, study C: 32 categories). Study A used the bibliographic fields as upper categories, the other ones did not. Only a few of the categories overlapped. Study B only evaluated the author field, pagination and the volume number, yet very granular. Even though study B and C were carried out by the same author, the error categories and the employed matchkeys differed. Study B used in the first round of matching full author name, publication year, starting page number and volume number, whereas study C used the first six characters of the author's family name, his or her first initial, the year of publication, volume number and starting page number. Another study investigated the difference between errors clearly made by authors and those made by the database. Those made by authors were five of the standard bibliographic fields excluding the article title. A different study divided the errors in seven categories, whereof three were related to the author name, one to wrong spellings in different fields, one to missing names and affiliations and two to missing articles or citations. Specific fields to categorize database errors were not really homogeneous throughout the studies. This is a summarized list: transcription error, source article record error, cited article omitted from cited article list // validity of citation counts and authentication of citations, completeness, overlap with other data sources, and accuracy of records (as upper category).

Of the other 90 studies investigating the accuracy of reference lists, three could not be evaluated as they did not specify their error categorization. Of the 87 remaining studies, 97% to 100% evaluated the six standard bibliographic fields. Specific mentioning of errors in authors’ initials is reported in 76%, whereas the number of authors is recorded in 54% and the order of the authors is investigated in 39% of the studies. The issue field was disregarded in 83% of the studies and only 26% also considered errors in punctuation. 62% of the studies based their error categorization on whether the error would impede the retrieval of the original publication or not. Table 2 displays an overview of the distribution of the error categories.
60% of the studies distinguished between major and minor errors, whereof four studies divided the severity of errors into three categories. The definition of these intermediate categories coincided with the other minor error categories and therefore was included into the evaluation of these. The majority defined major errors as those impeding immediate retrieval or identification of a publication. Another 8% used this definition in their error categorization as well, but did not label the categories as major and minor.

Table 2. Data fields as error categories.

<table>
<thead>
<tr>
<th>Bibliographic field</th>
<th>No. of studies</th>
<th>% of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>author names</td>
<td>87</td>
<td>100%</td>
</tr>
<tr>
<td>author initials</td>
<td>66</td>
<td>76%</td>
</tr>
<tr>
<td>author number</td>
<td>47</td>
<td>54%</td>
</tr>
<tr>
<td>author order</td>
<td>34</td>
<td>39%</td>
</tr>
<tr>
<td>article titles</td>
<td>84</td>
<td>97%</td>
</tr>
<tr>
<td>journal title</td>
<td>87</td>
<td>100%</td>
</tr>
<tr>
<td>volume</td>
<td>87</td>
<td>100%</td>
</tr>
<tr>
<td>issue</td>
<td>15</td>
<td>17%</td>
</tr>
<tr>
<td>year</td>
<td>85</td>
<td>98%</td>
</tr>
<tr>
<td>pagination</td>
<td>87</td>
<td>100%</td>
</tr>
</tbody>
</table>

Of the 52 studies distinguishing between major and minor errors, 96% counted an erroneous volume number field as major error. 92% considered an incorrect year as major error, whereas two studies did not record errors in the year field at all. 59% indicated that an erroneous journal title was a major error, whereas 14% stated it could be major or minor depending on the impact on retrieval. 63% of the studies reported incorrect page numbers as major and minor, whereupon major was defined as incorrect or omitted first page number or the page numbers did not overlap with the correct ones. Minor errors were defined as incorrect or omitted last page number or page numbers not overlapping with the correct ones. 27% regarded any error in pagination as major. 22% indicated author errors could be major or minor errors (depending again on the impact on retrieval). 18% considered author errors as major and 53% as minor. Yet, seven studies specifically mentioned that an error in the first author’s name was considered as major. If recorded or specified at all, authors’ initials, number and order of authors were mainly considered as minor errors. 45% of the studies categorized article title errors as minor, 18% categorized them as major, whereas 31% classified them as major and minor (depending again on the severity). 16% of the studies regarded an incorrect issue as a major error (eight of nine studies that recorded issue errors at all).

Discussion

The evaluation of the applied methods in the studies shows a very high tendency towards using the original publication as gold standard for verification. Even the database studies consulted the original in most cases. Merely the studies employing the matchkey method did not refer to the original. In order to identify whether the mistakes were made on the author or on the database
side, it is necessary to verify the references against the originals. Therefore it depends on the intended goal of the assessment whether the original should be consulted.

The majority of studies investigated one to five journals from one year. Yet, the sizes of the actual data sample diverge among the studies. The databases studies selected larger data samples than other studies. For those studies where a data sample ratio could be calculated, the majority used a sample between 0.4 and 10%. Studies tended to randomly select the cited articles. Just 31% of the studies calculated whether the results are statistically significant, i.e. the majority of studies are rather case studies than statistically representative assessments of bibliographic data accuracy.

The studies show that bibliographic data accuracy is assessed by author name(s), journal title, volume, year and pagination, whereof journal title, volume and year have the highest impact. Still, the granularity of error categories in these fields varies. For example, only half of the studies specifically mentioned the number and order of authors as separate error categories. This does not necessarily imply that the other studies disregarded them. Yet, getting the number or order of authors wrong in a bibliometric data source could mean mismatching or no matching at all for a cited article. Especially with regard to all the initiatives that work on facilitating author name disambiguation (e.g. ORCID – Open Researcher and Contributer ID, ResearcherID (WoS), AuthorID (Scopus), Author Claim, etc.), it seems necessary to thoroughly investigate all aspects related to an author’s name. Furthermore, the three database studies employing the matchkey method did this by using quite different bibliographic fields. Therefore, no standardized matchkey could be identified.

**Conclusion**

The results of this study contribute to research on data quality assessment in bibliometric data sources. It investigated studies assessing bibliographic data accuracy with regard to methods, data samples and in particular error categorization. The results show that bibliographic data accuracy is characterized by author name(s), journal title, volume, year and pagination and assessed by consulting the original publication. However, when assessing the accuracy of bibliographic data sources, matchkeys are an important method that needs to be considered. Due to the variation of granularity in the error categorization, no standardized error categorization for assessing bibliographic data accuracy could be determined.

Furthermore, Wyles (2004) states that the categorization of major and minor errors as used in most studies is not up to date anymore and depends on the applied method of retrieval. The distinction between major and minor error, yet is still valid as some errors impede retrieval or immediate identification and others are simply annoying. Thus, the author’s future research will concentrate on finding a standardized error categorization and method to assess bibliographic data accuracy. In particular, it will be interesting to investigate the granularity of the author name field (including initials, order and number) and its impact on retrieval in bibliometric data sources. Subsequently, the standardization will allow for more comparable error rates and contribute to research on data quality assessment in bibliometric data sources.
References


Appendix


