Journal of Library and Information Studies 12:2 (December 2014) p.19-38
doi:10.6182/jlis.2014.12(2).019

Testing an Automated Accuracy Assessment Method on Bibliographic Data

Marlies Olensky[1]

Abstract

This study investigates automated data accuracy assessment as described in data quality literature for its suitability to assess bibliographic data. The data samples comprise the publications of two Nobel Prize winners in the field of Chemistry for a 10-year-publication period retrieved from the two bibliometric data sources, Web of Science and Scopus. The bibliographic records are assessed against the original publication (*gold standard*) and an automatic assessment method is compared to a manual one. The results show that the manual assessment method reflects truer accuracy scores. The automated assessment method would need to be extended by additional rules that reflect specific characteristics of bibliographic data. Both data sources had higher accuracy scores per field than accumulated per record. This study contributes to the research on finding a standardized assessment method of bibliographic data accuracy as well as defining the impact of data accuracy on the citation matching process.

Keywords: Data Accuracy Assessment; Bibliographic Data; Bibliometric Data Sources; Web of Science; Scopus

1. Introduction

The discussion about which bibliometric data source, be it Web of Science (WoS), Scopus, Google Scholar or others, should be used for citation analyses seems to be a never-ending story. While a variety of studies (e.g. Archambault, Campbell, Gingras, & Larivière, 2009; Meho & Yang, 2007) compared the two main bibliometric data sources WoS and Scopus in terms of coverage, overlap, citation counts, etc., only a few studies (also) investigated the underlying problem: data quality of the data values (Hildebrandt & Larsen, 2008; Larsen, Hytteballe Ibanez, & Bolling, 2007; Moed, 2005). Even though, most of the comparative studies mention they had to carry out some kind of data cleaning and normalization (Wallin, 2005), they did not further investigate how accurate the data from these data sources actually is. However, the accuracy of the citation data has direct influence on the accuracy of the citation matching process

[1] Humboldt-Universität zu Berlin, Berlin School of Library and Information Science, Berlin, Germany
E-mail: marlies.olensky@ibi.hu-berlin.de

(Moed, 2005) because inaccurate data can cause a non-match of citing references to their cited articles. Even though sophisticated algorithms for matching citation data have been developed by several applied bibliometric research groups (e.g. CWTS (Note 1), iFQ (Note 2), Science-Metrix (Note 3)) that should rectify the majority of inaccuracies in citations, discrepancies in the two main bibliometric data sources, WoS and Scopus, still occur (Neuhaus & Daniel, 2008). Due to competitive advantage these algorithms are not publicly available and have not been evaluated so far. Only iFQ revealed parts of the research process of developing such a matching algorithm (Schmidt, 2012).

Non-matched or incorrectly matched references can influence the results of bibliometric calculations. Since those results become increasingly important in the context of research evaluation, it is important to ask whether the citation matching processes work sufficiently accurate for this purpose (Olensky, 2014). In this context, the need was identified to find a suitable method to assess bibliographic data accuracy, since the accuracy of the bibliographic references plays an important role in the citation matching process. Hence, this study investigates if an automated assessment, as described in the data quality literature, could be used to determine the accuracy of bibliographic data. The accuracy of two bibliographic datasets from both, WoS and Scopus, is assessed in an automated way and compared to the results of a manual assessment process. The results contribute to the research on determining the impact of data accuracy on citation analysis.

The paper is organized as follows: the sections *Inaccuracies in bibliometric data sources* and *Data quality / accuracy assessment* present the background of the research problem. The following section 4 discusses the research questions of this study, the applied methodology and the selected data sample. Section 5 presents the results of the assessments and section 6 concludes the paper.

2. Inaccuracies in Bibliometric Data Sources

The first commercial data source used for bibliometric analysis, WoS, was actually built as literature retrieval database for journal articles (Hood & Wilson, 2003) and the use as source for citation analyses was a succeeding development. WoS is the web portal provided by Thomson Reuters for searching different citation indexes (e.g. Arts & Humanities Citation Index, Science Citation Index, etc.). Elsevier's counterpart is Scopus, launched in 2004 as a reaction to the monopoly held by Thomson Reuters. Both databases offer functionalities for searching, browsing, sorting,

saving and exporting to citation management software, as well as citation counts and basic citation analyses. They are both subject to subscriptions. A cost-free alternative is Google Scholar, also launched in 2004. Contrary to WoS and Scopus, Google does not provide information about the number of records, indexed titles, covered subject areas or the time span in their database, which makes comparability and quality control even harder than with the two commercial ones. Hence, in this study we focus on the assessment of WoS and Scopus.

Researchers have studied data quality problems in bibliometric data sources. Moed and Vriens (1989) were the first ones to conduct a study on the accuracy of citation counts, pitfalls during data collection and the influence of random and systematic errors on citation analysis. They outlined errors and variations occurring in the fields: author name, journal title, publication year, volume and starting page number. Table 1 lists the areas of concern that have been identified in the literature inter alia by Moed (2005) and Jacsó (2008). Inaccuracies in references can be caused either by the author (e.g. provides inconsistent versions of his name or institutional affiliation), the citing author (e.g. jumbles the digits of the volume number) or by the database (e.g. interprets the issue number as the volume number) (Buchanan, 2006; Hildebrandt & Larsen, 2008; Moed & Vriens, 1989). All of these inaccuracies can be responsible for a non-link in the citation matching process.

Very few studies (e.g. Hildebrandt & Larsen, 2008; Larsen et al., 2007; Moed, 2005) have studied inaccuracies in WoS in more detail. Larsen et al. (2007) investigated WoS's automatic matching and linking algorithm, identified patterns of errors and suggested improvements to the algorithm. The overall results showed that of 33,024 citations 6.2% were erroneous with at least one error. In total, they found 2,626 errors in those 33,024 citations. They compared two time periods (1995-1997, 2000-2002), which did not show a strong improvement (6.4% vs. 5.8%). The domains performed quite differently with Law showing the highest error rate of 31.1% (Political Science 5.2%, Information Science 4.3%, Medicine 2.4%, Computer Science 2.3%, Biology 1.7%). In general, they concluded that the WoS algorithm must be quite conservative and some improvements could be made but would require more analysis. Hildebrandt and Larsen (2008) took a closer look at the high error rate in the field of Law. They found the most common errors in WoS were in the fields cited page, author names and year. Some errors originating from the original references were corrected in the citation index, others were

unfortunately added. Moed (2005) carried out the largest study on data accuracy in WoS by investigating 22 million citing references. He employed different match keys, as used in citation matching processes, in order to match the references to their 18 million target articles. 7.7% of references were discrepant and resulted in a non-match in WoS, i.e. a lost citation. Other studies reported rates between 6 and 12% (6.2%: Larsen et al., 2007; 7%: Tunger, Haustein, Ruppert, Luca, & Unterhalt, 2010; 9.4%: Moed & Vriens, 1989; 12%: Hildebrandt & Larsen, 2008). Overall, none of these studies looked further into finding a standardized method of how these errors could be categorized and how the knowledge of these inaccuracies could be leveraged to improve citation matching algorithms.

3. Data Quality / Accuracy Assessment

The Oxford English Dictionary (2013) defines data as "an item of information". The ISO 9000 standard defines 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). Hence, data quality can be defined as the "fitness for the purpose of use" (Maydanchik, 2007, p. 245; Wang & Strong, 1996) of an item of information. In the context of databases, data quality is usually defined along four data quality dimensions: accuracy, completeness, consistency, and timeliness (or currency) (Batini & Scannapieco, 2006; Bovee, Srivastava, & Mak, 2003; Jarke, Lenzerini, Vassiliou, & Vassiliadis, 2003; Naumann, 2002; Wand & Wang, 1996; Wang & Strong, 1996). Most studies have identified data accuracy as the key dimension of data quality (Batini & Scannapieco, 2006; Wand & Wang, 1996) and, therefore, we focus on the investigation of the accuracy of bibliographic data values.

The literature provides a variety of techniques to assess data quality in databases and summarizes those in different data quality assessment frameworks (Batini, Cabitza, Cappiello, & Francalanci, 2008; Even & Shankaranarayanan, 2007; Lee, Strong, Kahn, & Wang, 2002; Scannapieco, Virgillito, Marchetti, Mecella, & Baldoni, 2004; Su & Jin, 2004; Wang, 1998). These mostly describe how enterprises can maintain quality in their databases by employing record linkage, business process rules and similarity measures (Batini, Cappiello, Francalanci, & Maurino, 2009). The measurement of data accuracy is basically defined as the ratio of correct and incorrect values and can be expressed in different ways (cf. Table 2). The definition of what constitutes a data unit is up to the individual assessment. It could be a data field, data record or even

Table 1. Reported Inaccuracies in Bibliometric Data Sources

Area of concern	Problem description	Example
Inconsistent and erroneous spellings of author names	Author names with accented characters	Stalnioniené or Ludányi
	Names with prefixes	van Hooland
	Double middle initials with or without punctuation	Weng, C.-H. vs. Weng, C-H vs. Weng, C.H.
	Misspelling of author names with many adjacent consonants	Mühlberger or Sprecher
	Problematic names in non-Latin alphabets when transliterated	Chang vs. Chung
Lack of journal title standardization	Various abbreviations and punctuations in journal titles	Heteroatom Chemistry vs. Heteroat. Chem. vs. Heteroatom Chem
Numeric bibliographic fields (publication year, volume number, pagination)	Transposed digits	p. 564 vs. p. 654
	Plus or minus one digit	1997 vs. 1998

Note. Adapted from "Google Scholar - A new data source for citation analysis," by A.-W. Harzing, 2008, Retrieved from http://www.harzing.com/pop_gs.htm; "Informetric studies using databases: Opportunities and challenges, " by W. W. Hood and C. S. Wilson, 2003, *Scientometrics*, *58*(3), pp. 587-608; "The plausibility of computing the *h*-index of scholarly productivity and impact using reference-enhanced databases, " by P. Jacsó, 2008, *Online Information Review, 32*(2), pp. 266-283; "Error rates and error types for the Web of Science algorithm for automatic identification of citations, " by B. Larsen, K. Hytteballe Ibanez, and P. Bolling, 2007, September, *Presented at the 12th Nordic Workshop on Bibliometrics and Research Policy*, Copenhagen, Denmark; "Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar, " by L. I. Meho, and K. Yang, 2007, *Journal of the American Society for Information Science and Technology, 58*(13), pp. 2105-2125; "Citation analysis in research evaluation, " by H. F. Moed, 2005, *Information Science and Knowledge Management*, *9*, Dordrecht: Springer; ""The Delphic Oracle": An analysis of potential error sources in bibliographic databases, " by D. Tunger, S. Haustein, L. Ruppert, G. Luca, and S. Unterhalt, 2010, In CWTS (Ed.), *Proceedings of the 11th International Conference on Science and Technology Indicators*, pp. 282-283, Leiden, Netherlands.

Table 2. Data Accuracy Metrics

Data accuracy	Metric	Source
Accuracy	$\frac{incorrect\ values}{correct\ values}$	Loshin (2001)
Free-of-error dimension	$1-\frac{count\ of\ data\ units\ in\ error}{total\ number\ of\ data\ units}$	Pipino, Lee, & Wang (2002)
Free-of-error-rating	$1-\frac{number\ of\ data\ units\ in\ error}{total\ number\ of\ data\ units}$	Lee, Pipino, Funk, & Wang (2006)
Accuracy score	$\frac{count\ of\ rel.rec.-count\ of\ err.\ rec.}{count\ of\ relevant\ records}$	Maydanchik (2007)
Syntactic Accuracy	$\frac{number\ of\ correct\ values}{number\ of\ total\ values}$	Batini et al. (2009)

an entire dataset. Data accuracy can also be assessed in a more complex way by measuring the distance between the values stored in the database and the correct one (Batini et al., 2009; Batini & Scannapieco, 2006). Redman (1996) and Scannapieco et al. (2004) suggest using a distance function to calculate the data accuracy score. For example, the Levenshtein distance function is a widely used method to measure the distance between two strings, i.e. how many edits it takes to convert value *v*, the assessed value, into value *v'*, the correct value (Levenshtein, 1966). In contrast, the Jaro-Winkler distance function (Winkler, 1995) measures the similarity of two strings, i.e. how many characters two strings have in common. However, before the accuracy of a value can be assessed, one needs to define what the correct value is and what qualifies as an incorrect data value.

The accuracy of references in research articles and bibliometric data sources has been studied before, but not by employing any of the above-mentioned frameworks from the data quality literature. In a literature study, we, therefore, investigated all of these studies (98 studies in total) to determine whether inaccuracies in references are assessed and categorized in a standardized and/or automated way (Olensky, 2012). The main aspects of the evaluation were: main goal of study; data sources employed; number of journals investigated; number, publication type and year of citing articles; number and publication type of cited articles; selection of the data sample; assessment method; type of error categorization. The majority of studies was carried out by researchers in their own field to point out negligent references that would

impede fellow researchers from retrieving their sources. Moreover, a few studies (e.g. García-Pérez, 2010; Moed, 2005; Moed & Vriens, 1989) assessed data accuracy of bibliometric data sources.

Table 3 illustrates the results and shows that bibliographic data is measured by the accuracy of the following fields: *author name(s)*, *journal title*, *volume*, *year* and *pagination*. Most studies verified the accuracy of the references by consulting the original article, i.e. defined as the correct values *v'*. Even more than half of the database studies used the original publication as verification standard, except for two, which employed match keys to identify inaccurate data. The study also revealed bibliographic data errors are not categorized in a standardized way and the granularity of categories varies. Half of the studies divided the errors into the groups *major* and *minor*; some added an *intermediate* category, others just listed and described the nature of the inaccuracies (e.g. page number missing, small variation in author name (Moed, 2005); wrong cited year, swapping of digits (Larsen et al., 2007)).

Table 3. Aspects of Bibliographic Data Accuracy

Bibliographic field	% of studies
Author name(s)	100
Author initials	76
Author number	54
Author order	39
Article titles	97
Journal title	100
Volume	100
Issue	17
Year	98
Pagination	100

Note. Adapted from "How is bibliographic data accuracy assessed? " by M. Olensky, 2012, In É. Archambault, Y. Gingras, and V. Larivière (Eds.), *Proceedings of the 17th International Conference on Science and Technology Indicators*, pp. 628-639, Montreal, Canada. Retrieved from http://2012.sticonference.org/index.php?page=proc

4. Research Questions and Methodology

Since match keys, as used in citation matching processes, can identify inaccurate data records, but do not provide information about how inaccurate data is, we identified the need to test a different method to assess bibliographic data accuracy in an automated way. The Levenshtein distance function was chosen, because it is a widely used distance metric in the data quality literature and it is one parameter in the matching algorithm of iFQ (Schmidt, 2012).

We address the following research questions:

- Is data accuracy assessment as described in the data quality literature suitable for assessing data accuracy of bibliographic records?
  - (1) Do the data sources differ with regard to their data accuracy scores when assessed automatically and manually?
  - (2) Does the Levenshtein distance score really reflect which data source provides the most accurate data?

To answer the main research question and the two sub-research questions, the assessment process was divided into an automated and a manual assessment of the data. The manual assessment was used to verify the results of the automated assessment method. Comparing the accuracy scores of the different data sources for the two assessment methods revealed whether an automated accuracy assessment as described in the data quality literature is an adequate means to assess bibliographic data. The applied methodology is graphically represented in Figure 1.

Bibliographic data accuracy is characterized by the data fields of *author name*(s) (including *first* and *second initial* of the given names), *article title*, *journal title*, *volume number*, *publication year* and *pagination* (Olensky, 2012), which are therefore used as assessment parameters. The investigation of other fields, such as Times Cited (WoS) and Cited By (Scopus), is not in the scope of this study. The original publication is used as a *gold standard* to verify the accuracy of those bibliographic data fields (Olensky, 2012). To explore the reasons behind the inaccuracies is not in the scope of this study. The study follows the wording of Maydanchik (2007) and uses

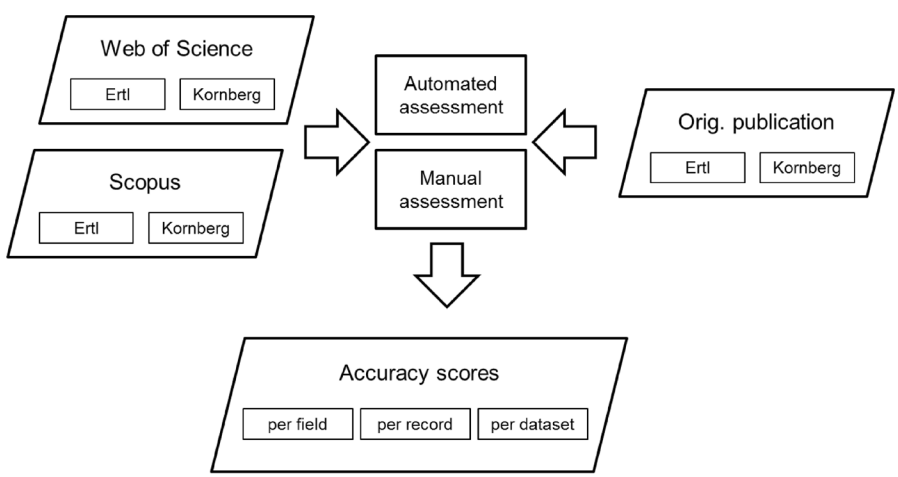

Figure 1. Graphical Representation of the Applied Methodology

the term *accuracy score* to represent the data accuracy metric, but expresses it in percentages.

4.1 Automated assessment

The automated evaluation calculated the Levenshtein distance score for the data value of each bibliographic field compared to the value of the original publication. The scores were then accumulated for each record per dataset and the total accuracy scores were calculated. Additionally, the data source providing the lowest Levenshtein distance score per record was determined and accumulated per dataset.

Some journal titles in original articles only give the abbreviated journal title; therefore, the Levenshtein distance function would provide a fairly high score for these fields. Hence, for the WoS data, we carried out an additional evaluation replacing the field *SO* (Publication Name) with *JI* (ISO Source Abbreviation) to check if these data reflect a more accurate picture. This was not necessary for Scopus data, as the values in the fields *Source Title* and *Abbreviated Source Title* do not differ at all for both data samples.

4.2 Manual assessment

During the manual assessment we went through the discrepancies the automatic evaluation had found and assigned *inaccuracy points* (IAPs) to the respective data fields: 0 = accurate, 1 = minor, 2 = medium, 3 = major inaccuracy. Afterwards, the scores were accumulated for each record per dataset and the total accuracy scores were calculated. Additionally, the data source providing the lowest score of IAPs per record was determined and accumulated per dataset. The findings of a previous literature study (Olensky, 2012), where we investigated the definition of what an inaccuracy is as well as the weighting of these in the different bibliographic fields, were considered. Discrepancies in the fields *author name* (including given names' *initials*), *journal title*, *volume number* and *publication year* were rated higher than inaccuracies in the other fields because these fields disambiguate a publication and are often used in the citation matching process. The following paragraphs describe the rating procedure in more detail and Table 4 summarizes the IAPs.

Punctuation discrepancies were not counted as inaccuracy. Special characters like the Scandinavian å, or the German umlauts (ä, ö, ü) are represented accordingly in Scopus, whereas WoS converts these into normal letters, yet not consistently. Therefore, we made a difference for the two databases and rated the lack of special characters in WoS as minor inaccuracy (1 IAP) per field.

Inaccuracies in the *authors' last name*, such as spelling mistakes or incorrect names,

Table 4. Overview of IAPs – Manual Assessment

Bibliographic data field	Inaccuracy description	IAPs
Any	Punctuation	0
Any	Special characters	1
Author's last name	Spelling mistakes	3
	Incorrect	3
First initial	Incorrect	3
	Missing	3
	Hyphenation	1
Second initial	Incorrect	1
	Missing	1
	Added	0
Article title	Different abbreviations	1
	Additional information for the reader	0
	English translations of foreign article titles	2
Publication name	Abbreviated	0
	Subtitle	1
Publication year	Incorrect	3
Volume number	Incorrect	3
	Incomplete	2
Starting page	Incorrect	2
	Missing	2
Ending page	Incorrect	1
	Missing	1

were rated as major inaccuracy (3 IAPs). An inaccurate, wrong or missing *first initial* (of the given name) was reflected by 3, an inaccurate *second initial* by 1 IAP. If an author's first name was hyphenated, like Wei-Hau, this is represented correctly in Scopus as W.-H. In WoS, those are listed as separate initials. During the manual assessment, only 1 IAP was assigned to the *first initial* for this discrepancy instead of the automatically assigned 2 IAPs to the *first initial* and another IAP to the *second initial* (in total 3 IAPs). If the data source contained a *second initial* that was not in the original article, the automatic assessment would assign 1 IAP

to this field. The manual assessment corrected this to 0 IAP, as this additional information supports author name disambiguation. Yet, if the original gave the *second initial* and the data source lacked this information, this was counted as minor inaccuracy (1 IAP).

In the field *article title*, chemical elements were used as abbreviation (e.g. Ag) or with their full name (e.g. Silver). This discrepancy was rated as minor inaccuracy (1 IAP). In some cases, the *article titles* in the data source contained additional information for the user (e.g. reprint information) which resulted in a higher Levenshtein score. During the manual assessment, this was disregarded. WoS and Scopus treat *article titles* in other languages than English differently. WoS gives the translated English title, whereas Scopus gives the original title and the English translation in brackets. In this case, we assigned 0 IAP to Scopus and 2 IAPs to WoS.

As mentioned before, many original articles give only the *abbreviated publication name*. The manual assessment considered this fact and manually checked the *abbreviated* with the *full publication name*. In case of a correct match, 0 IAP were assigned. *Journal titles* can have subtitles, which are registered in the data sources. Yet, they might not be indicated in the original publication; this discrepancy was reflected by a minor inaccuracy (1 IAP).

A wrong *publication year* was reflected with 3 instead of 1 IAP from the automatic assessment. If the *volume number* was not given in the original article but the database held a *volume number*, this was not counted as inaccuracy in the manual assessment. An incorrect *volume number* was rated as a major inaccuracy (3 IAPs). An incomplete *volume number* (e.g.: 11 instead of 11-12) was rated as medium inaccuracy (2 IAPs).

An inaccuracy in the *starting page* field was rated more severe than in the *ending page* field. Any inaccurate or missing *starting page* was rated as medium inaccuracy (2 IAPs). Any inaccurate or missing *ending page* was rated as minor inaccuracy (1 IAP). For the field *ending page*, the inaccuracy scores according to Levenshtein are quite high if the article is only one page long, because in Scopus one-page-articles have no ending page. The manual assessment corrected this by assigning a minor inaccuracy (1 IAP) to these fields.

4.3 Data sample

We chose Nobel Prize winners as data sample, as these have been the topic of previous bibliometric studies (Gingras & Wallace, 2010) and one can assume they publish and get cited. WoS and Scopus provide good coverage of Chemistry literature that is why two Nobel Prize winners from that domain were chosen.

We selected one English-speaking and one from a non-English speaking country: Roger D. Kornberg, an American biochemist and Professor at Stanford University School of Medicine, won the Nobel Prize for Chemistry in 2006. Gerhard Ertl, a German physicist and Professor emeritus at the Department of Physical Chemistry at Fritz-Haber-Institut der Max-Planck-Gesellschaft in Berlin, Germany, won it in 2007. Publications of both Nobel Prize winners were retrieved for a publication period of 10 years counted back from the last winning year (1998-2007), regardless whether they were the first or co-author. Articles and proceedings / conference papers were analysed, all other publication types were excluded.

Both author names (including spelling variations) were searched in the web versions of WoS (Note 4), Scopus (Note 5). In contrast to the two related Danish studies (Hildebrandt & Larsen, 2008; Larsen et al., 2007) not the cited references but the full bibliographic records of the actual publications were investigated. Those were downloaded from both data sources in September 2012. To verify the records of the "right" Kornberg and Ertl, we checked the "typical" co-authors, journal titles as well as cross-checked the data with the institutional website. The websites were also consulted to verify the completeness of records from both data sources. The requirements for the publications to be included into the study besides the publication type were defined as: they had to be written in English or German, be obtainable online, in the library or via inter-library loan and the publications had to be indexed in both databases. The original publications were downloaded and their bibliographic data manually gathered. This process was double-checked by two co-researchers to ensure the accuracy of these data. In the Ertl data sample, there were 5 records unique to WoS, 3 unique to Scopus. In total 134 publications were investigated. In the Kornberg data sample, there was only 1 publication unique to Scopus (a conference paper) and none to WoS. In total 63 publications were investigated. The records were automatically pre-processed and stored in a MySQL database for further analysis.

5. Results

This section summarizes and compares the results of the two assessment processes. The data sources are abbreviated as follows: Scopus is not abbreviated, Web of Science is abbreviated as WoS and the variant evaluation of WoS, where the abbreviated publication name was compared to the original publication, is abbreviated as WoSAbb. The data samples are abbreviated as follows: Ertl = E; Kornberg = K.

The results of the automated evaluation show that Scopus provides data with the least discrepancies. Following the metric described in section 3 on *Data quality / accuracy assessment*, the accuracy scores were calculated. E: 37% and K: 38% of Scopus' records did not contain any discrepancy to the original publication, whereas WoS only provided E: 30% and K: 27% of absolutely accurate records (cf. Table 5). In order to validate the Levenshtein distance function as means to evaluate bibliographic data accuracy, a manual assessment of both data samples following the methodology described above was carried out. First of all, the results reveal that both data samples are more accurate than the automated evaluation showed. Yet, Table 5 affirms that Scopus still provides the most accurate data. The Δ between the automated and manual assessment lies between 29% and 57%, which indicates that the Levenshtein distance score does not reflect true accuracy scores of the bibliographic data. For both data sources the Δ is lower in the Ertl data sample than in the Kornberg sample and higher for both samples in Scopus than in WoS and the highest for the WoSAbb variant.

Taking into account the absolute Levenshtein distance score for each record, we compared the data sources to determine the one that provides the records with the lowest scores (cf. Table 6). The automated assessment process reveals a less distinct ranking of data sources, whereat Scopus provides slightly more accurate data. However, in the manual assessment the majority of records (E: 67% and K: 75%) from both data sources were equally accurate. In other words, in 67% of the records in the Ertl data sample WoS and Scopus inaccuracy scores were equally accurate and the same is true for 75% of the records in the Kornberg data sample. Yet, looking at the remaining records, Scopus, again, provides records with fewer discrepancies. The Δ between the automated and manual assessment lies between 1% and 50%. The Δ of records that are equally accurate in both data sources (row 3) is the highest in

Table 5. Automated vs. Manual Evaluation, Compared to Original Publication, Record Level. How Many Records Are 100% Accurate?

	Ertl			Kornberg		
	Automated(%)	Manual(%)	Δ(%)	Automated(%)	Manual(%)	Δ(%)
Scopus	37	76	39	38	90	52
WoS	30	59	29	27	73	46
WoSAbb	17	59	42	16	73	57

Table 6. Automated vs. Manual Evaluation, Compared to Original Publication, Record Level. Which Data Source Provides the Most Accurate Records?

	Ertl			Kornberg		
	Automated(%)	Manual(%)	Δ(%)	Automated(%)	Manual(%)	Δ(%)
Scopus	26	27	1	32	19	13
WoSAbb	26	-	-	29	-	-
Scopus & WoS	25	67	42	25	75	50
WoS	15	6	9	2	6	4
All the same	5	-	-	13	-	-
WoS & WoSAbb	2	-	-	-	-	-

both data samples. The results of this analysis corroborate that the discrepancy between the automatically and the manually calculated accuracy scores are quite high and that the Levenshtein distance score does not reflect the severity of inaccuracies correctly.

Calculating the accuracy scores on the data field level for each data set, the Δ between the automated and the manual assessment process diminishes. The Δ is equalized for all data sources and data sets and lies between 3% and 5% (cf. Table 7). However, the results still show that the manual assessment process draws a truer picture of accuracy scores than the automated one. Additionally, the results show that the accuracy assessment of bibliographic data should be carried out on a bibliographic data field level and not be accumulated per record.

6. Conclusion

This study investigated data accuracy in the two main bibliometric data sources, WoS and Scopus, per data field, per data record and accumulated the results per dataset. It tested whether an automated assessment method using the Levenshtein distance function, as described in the data quality literature, can be applied to bibliographic data. The main result is that the Levenshtein distance function is a good means to determine whether a data record contains discrepancies, but the score does not provide a true picture of how inaccurate a field is without the application of additional rules. Therefore, a modified assessment method is needed. The rules spelled out in the manual assessment method reflect most of the required adjustments that could be made to an automatic assessment method. They mirror specific characteristics of bibliographic data:

Table 7. Automated vs. Manual Evaluation, Compared to Original Publication, Data Field Level. How Many Fields Are 100% Accurate?

	Ertl			Kornberg		
	Automated(%)	Manual(%)	Δ(%)	Automated(%)	Manual(%)	Δ(%)
Scopus	94	97	3	96	99.5	3.5
WoS	93	96	3	94	98.0	4.0
WoSAbb	91	96	5	93	98.0	5.0

- different presentation of data (e.g. one-page-articles in Scopus have no end page)
- abbreviated publication names (check with ISO abbreviation or ISSN)
- translated article titles
- punctuation
- special characters (e.g. German Umlaut)
- non alphanumeric characters (e.g. α)
- domain-specific abbreviations (e.g. *Ag* // Silver)
- different weighting of bibliographic fields
- different weighting of inaccuracies (omitted // inaccurate // incomplete)

Additionally, during the manual data accuracy assessment of author names, it was noticed that author data, just because they are accurate if compared to the original, does not necessarily mean, they are also adequate for the use in citation analysis. The example of an author named Zei M.-S. (variants found in the data: Zei M., Zei M. S., Zei Mau-Scheng) shows that the consistency of author names throughout a data source might be, therefore, more important than the accuracy compared to the original source. In some cases, the original source might serve as source for author disambiguation but not all of the journals fully print the author's first name.

Overall, the results lead to the conclusion that the Levenshtein distance score does not reflect a true inaccuracy rate for the two data samples investigated. The results show that the accuracy scores per record draw a very different picture than the ones per data field. We, therefore, recommend using the accuracy scores per data field to describe bibliographic data accuracy. In future work, we will apply the assessment method enriched by the additional rules for bibliographic data to a larger, more representative data sample that includes cited as well as citing articles in order to determine the impact of data accuracy on citation matching.

Notes

Note 1 Centre for Science and Technology Studies in Leiden, http://www.cwts.nl/

Note 2 Institut für Forschungsqualität in Berlin, http://www.forschungsinfo.de/

Note 3 Science-Metrix in Montreal, http://www.science-metrix.com/

Note 4 Searched was only the Science Citation Index Expanded (SCI-EXPANDED).

Note 5 The subject areas Social Sciences & Humanities were excluded to narrow down the search results.

References

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology, 60*(7), 1320-1326. doi: 10.1002/asi.21062

Batini, C., Cabitza, F., Cappiello, C., & Francalanci, C. (2008). A comprehensive data quality methodology for web and structured data. *International Journal of Innovative Computing and Applications, 1*(3), 205-218. doi: 10.1504/IJICA.2008.019688

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), 16:1-16:52. doi: 10.1145/1541880.1541883

Batini, C., & Scannapieco, M. (2006). *Data quality: Concepts, methodologies and techniques*. Berlin, Germany: Springer.

Bovee, M., Srivastava, R. P., & Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems, 18*(1), 51-74. doi: 10.1002/int.10074

Buchanan, R. A. (2006). Accuracy of cited references: The role of citation databases. *College & Research Libraries, 67*(4), 292-303. doi: 10.5860/crl.67.4.292

Data. (2013). In *Oxford English Dictionary: Online Version* (3rd edition).

Even, A., & Shankaranarayanan, G. (2007). Utility-driven assessment of data quality. *SIGMIS Database, 38*(2), 75-93. doi: 10.1145/1240616.1240623

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of *h*-indices in psychology. *Journal of the American Society for Information Science and Technology, 61*(10), 2070-2085. doi: 10.1002/asi.21372

Gingras, Y., & Wallace, M. (2010). Why it has become more difficult to predict

Nobel Prize winners: A bibliometric analysis of nominees and winners of the chemistry and physics prizes (1901-2007). *Scientometrics, 82*(2), 401-412. doi: 10.1007/s11192-009-0035-9

Harzing, A.-W. (2008). *Google Scholar - A new data source for citation analysis*. Retrieved from http://www.harzing.com/pop_gs.htm

Hildebrandt, A. L., & Larsen, B. (2008, September). *Reference and citation errors: A study of three law journals*. Paper presented at the 13th Nordic Workshop on Bibliometrics and Research Policy, Tampere, Finland.

Hood, W. W., & Wilson, C. S. (2003). Informetric studies using databases: Opportunities and challenges. *Scientometrics, 58*(3), 587-608. doi: 10.1023/B:SCIE.0000006882.47115.c6

International Organization for Standardization (ISO). (2005). *ISO 9000:2005: Quality management systems – Fundamentals and vocabulary*. Geneva, Switzerland: International Organization for Standardization.

Jacsó, P. (2008). The plausibility of computing the *h*-index of scholarly productivity and impact using reference-enhanced databases. *Online Information Review, 32*(2), 266-283. doi: 10.1108/14684520810879872

Jarke, M., Lenzerini, M., Vassiliou, Y., & Vassiliadis, P. (2003). *Fundamentals of data warehouses* (2nd ed.). Berlin, Germany: Springer. doi: 10.1007/978-3-662-05153-5

Larsen, B., Hytteballe Ibanez, K., & Bolling, P. (2007, September). *Error rates and error types for the Web of Science algorithm for automatic identification of citations*. Paper presented at the 12th Nordic Workshop on Bibliometrics and Research Policy, Copenhagen, Denmark.

Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2006). *Journey to data quality*. Cambridge, MA: MIT Press.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management, 40*(2), 133-146. doi: 10.1016/S0378-7206(02)00043-5

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10*, 707-710.

Loshin, D. (2001). *Enterprise knowledge management: The data quality approach*. San Diego, CA: Morgan Kaufmann.

Maydanchik, A. (2007). *Data quality assessment*. Bradley Beach, NJ: Technics Publications.

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58*(13), 2105-2125. doi: 10.1002/asi.20677

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht, Netherlands: Springer. doi: 10.1007/1-4020-3714-7

Moed, H. F., & Vriens, M. (1989). Possible inaccuracies occurring in citation analysis. *Journal of Information Science, 15*(2), 95-107. doi: 10.1177/016555158901500205

Naumann, F. (2002). Quality-driven query answering for integrated information systems. In *Lecture notes in computer science: Vol. 2261*. Berlin, Germany: Springer. doi: 10.1007/3-540-45921-9_6

Neuhaus, C., & Daniel, H.-D. (2008). Data sources for performing citation analysis: An overview. *Journal of Documentation, 64*(2), 193-210. doi: 10.1108/00220410810858010

Olensky, M. (2012). How is bibliographic data accuracy assessed? In É. Archambault, Y. Gingras, & V. Larivière (Eds.), *Proceedings of the 17th International Conference on Science and Technology Indicators* (pp. 628-639). Montreal, Canada. Retrieved from http://2012.sticonference.org/index.php?page=proc

Olensky, M. (2014). *Data accuracy in bibliometric data sources and its impact on bibliometric data sources* (Unpublished doctoral dissertation). Berlin School of Library and Information Science, Humboldt-Universität zu Berlin, Germany.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211-218. doi: 10.1145/505248.506010

Redman, T. C. (1996). *Data quality for the information age (Artech House computer science library)*. Boston, MA: Artech House.

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., & Baldoni, R. (2004). The DaQuinCIS architecture: A platform for exchanging and improving data quality in cooperative information systems. *Information Systems, 29*(7), 551-582. doi: 10.1016/j.is.2003.12.004

Schmidt, M. (2012). Development and evaluation of a match key for linking references to cited articles. In É. Archambault, Y. Gingras, & V. Larivière (Eds.), *Proceedings of the 17th International Conference on Science and Technology Indicators* (pp. 707-718). Montreal, Canada. Retrieved from http://2012.sticonference.org/index.php?page=proc

Su, Y., & Jin, Z. (2004). A methodology for information quality assessment in the designing and manufacturing processes of mechanical products. In I. N. Chengalur-

Smith, L. Raschid, J. Long, & C. Seko (Eds.), *Proceedings of the Ninth International Conference on Information Quality (ICIQ-04)* (pp. 447-465). Cambridge, MA.

Tunger, D., Haustein, S., Ruppert, L., Luca, G., & Unterhalt, S. (2010). "The Delphic Oracle": An analysis of potential error sources in bibliographic databases. In CWTS (Ed.), *Proceedings of the 11th International Conference on Science and Technology Indicators* (pp. 282-283). Leiden, Netherlands.

Wallin, J. A. (2005). Bibliometric methods: Pitfalls and possibilities. *Basic & Clinical Pharmacology & Toxicology, 97*(5), 261-275. doi: 10.1111/j.1742-7843.2005.pto_139.x

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86-95. doi: 10.1145/240455.240479

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM, 41*(2), 58-65. doi: 10.1145/269012.269022

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-33. Retrieved from http://www.jstor.org/stable/40398176

Winkler, W. E. (1995). Matching and record linkage. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, & P. S. Kott (Eds.), *Business Survey Methods* (pp. 355-384). New York, NY: Wiley. doi: 10.1002/9781118150504.ch20

(Received: 2014/7/28; Accepted: 2014/11/7)

圖書資訊學刊　第12卷第2期　（103.12）　頁19-38
doi:10.6182/jlis.2014.12(2).019

書目資料準確性評估自動化之測試研究

Testing an Automated Accuracy Assessment Method on Bibliographic Data

Marlies Olensky[1]

摘　要

本研究探討資料品質文獻所提及的自動化資料準確性評估法，以瞭解其用於評估書目資料時的適切性。本研究用來測試的書目資料為兩位諾貝爾化學獎得主10年內之出版品，書目資料檢索自Web of Science與Scopus；在準確性評估上，分別以自動化與人工兩種評估法進行書目資料準確性測試，之後再跟原始出版品比對，以瞭解人工與自動化評估的高下。研究結果顯示，人工評估法的準確性得分較高，自動化評估法還需要納入更多能反映書目資料特質的評估規則，始能提高準確性。在兩組書目資料的測試中，單一分欄資料準確性的評估，都比整體書目記錄評估的表現要好。本研究之貢獻在於增進對書目資料準確度標準評估法的探討，並說明了資料準確性在引文比對過程中的重大影響。

關鍵字：資料準確性評估、書目資料、書目計量資料來源、Web of Science、Scopus

[1] 德國柏林洪堡大學圖書館與資訊科學學院
Humboldt-Universität zu Berlin, Berlin School of Library and Information Science, Berlin, Germany
E-mail: marlies.olensky@ibi.hu-berlin.de

註：本中文摘要由圖書資訊學刊編輯提供。
以APA格式引用本文：Olensky, M. (2014). Testing an automated accuracy assessment method on bibliographic data. *Journal of Library and Information Studies, 12*(2), 19-38. doi: 10.6182/jlis.2014.12(2).019
以Chicago格式引用本文：Marlies Olensky. “Testing an automated accuracy assessment method on bibliographic data.” *Journal of Library and Information Studies* 12 no. 2 (2014): 19-38. doi: 10.6182/jlis.2014.12(2).019