Development and Evaluation of a Match Key for Linking References to Cited Articles[1]

Marion Schmidt

schmidt@forschungsinfo.de
iFQ Institute for Research Information and Quality Assurance, Schützenstr. 6A, Berlin, 10117 (Germany)

Research in Progress Paper

Abstract

The very foundation of citation analyses is the processing and linking of references to articles in databases. As it is not possible to evaluate the relevant database providers' algorithms directly, published literature on this issue has been restricted mostly to small case studies. As part of a bigger on-going project focussing on reliability measures for bibliometric indicators, the development of an own algorithm for linking references to cited articles in Web of Science was initiated. In the first part of this research in progress paper the design of this developed algorithm is described. In the second part, the algorithm is evaluated based on an intellectual analysis of matched and non-matched references of a random sample. By way of this case study, we were able to determine to which extent the algorithm yields results that are true and to test and implement improvements.

Introduction

The underlying data quality and the mechanisms of linking citations to articles are at the very foundation of citation analysis. Due to the fact that the procedures and algorithms database providers use for processing and linking references are not documented, those algorithms cannot be evaluated directly. In a number of case studies the quality of reference lists were analysed.

Sweetland (1989) evaluated four studies published in the 1970s and estimated that 7% of author errors would make it difficult to locate a publication because of errors in publication title, year, volume and pagination. 15% of references contained errors in author names. Buchanan (2006) reviewed a range of case studies performed since then. He reported error rates between 3 and 15% in studies defining author errors as errors in the key elements author name, source title, volume, pagination and year. In a case study of altogether 11,108 articles in SCIE and SciFinderScolar he analysed the interplay of errors in reference lists and database mapping errors. He determined 10% missing links between articles. While this is highly informative, studies like these can only give a glimpse at the quality of matching at certain times as changes in referencing habits and, more importantly, changes in the database's matching procedure can render these studies obsolete.

Developing an own match keys facilitates this task. Moed (2002) referred to an analysis performed by the Centre for Science and Technology Studies (CWTS) that was based on 20 million cited references. He determined an overall number of 7% discrepant cited references

[1]Work on this project has been funded by the German Federal Ministry of Education and Research. The author thanks William Peter Dinkel as former contributor to this project as well as Daniel Sirtes, Jörg Neufeld and Jasmin Schmitz for helpful discussions

which could not be obtained by a simple matching procedure. He described significant effects on country statistics. In his previous work on this project Dinkel (2011) suggested that especially publications with authors from Russia, China, South Korea, India, Taiwan and Japan tend to be underestimated because of erroneous author names.

By constructing and applying an own match keys for Web of Science data and comparing its results to the Thompson Reuters match key, several goals are pursued: It will be explored

- to which amount it is possible to assign erroneous reference data unambiguously
- how this would change citation counts in relation to an exact matching approach
- to which amount reference data can only be assigned ambiguously or only on the level of the cited journal

Finally we aim for the analysis of potential biases and monitoring changes in the Thompson Reuters match key which is easier with this comparative algorithmic approach than through single case studies.

In the process of matching references to cited documents, there are three main aspects and sources of error: Firstly, citing authors make careless mistakes or forget metadata like volume and pagination. Secondly, citations to articles in press which have become more and more established due to repositories or informal scientific communication cause problems as they may lack complete or contain wrong metadata. To a lesser degree, the same problem applies to references to advance online articles, as Thomson Reuters indexes all articles with respect to their print publication.

Finally, the providers of citation databases have to process and standardize these references, which are initially published in a variety of possible citation styles. In the course of this step, errors occur.

In the main step of matching the standardized references to cited documents, the database providers have to deal with the described first and second types of mistakes:

By way of a very exact matching procedure, this would cause false negatives and citation counts would tend to underestimate the real numbers. By way of a weak match key, the number of false positives rises.

A much referred match key has been suggested quite early by Braun, Glänzel & Schubert in 1985. It consists of four letters of the author name, the last two digits of the publication year, the first character of journal title, volume and first page. Since then, a number of match keys highlighting different elements has been suggested in the literature (see Synnestvedt 2007 for an overview).

Description of the data and the heuristically defined algorithm

Our algorithm matches each reference pointing to a specific year is matched against every article[2] of that year. As first step an exact matching procedure is applied, where all bibliographic data categories of references and articles have to be complete and identical. It happens in single cases nevertheless that by way of the exact matching procedure several target articles for the same

[2] The algorithm matches against all items with no restriction to specific document types

reference are identified. This may apply, for example, to several meeting abstracts on the same page of a journal by an author of the same name.[3] For the set of references where it was not possible to identify the target article by way of the exact matching procedure, a sequence of iterating fault-tolerant matching procedures is performed for every reference-article pair: The basic match key that ensures the highest level of precision iteratively becomes weaker by way of allowing deviations or even skipping single or combinations of bibliographic metadata. The iterative procedure is aborted for a specific reference-article pair if a match is achieved, but proceeds with regard to the whole data corpus in order to check if more exact matchings may occur. In the end, only the most exact matching will be saved for each reference. If the same match key assigns the reference to several target articles, all possible target articles are saved and the matching is coded as non-unique.

The development of the fault-tolerant part of the algorithm has been done in two steps: At first all design principles and thresholds are defined heuristically. Secondly, the algorithm is evaluated and further developed on the basis of an intellectual analysis of a random sample.

Data

The reference dataset consists of entries of references lists, represented by a string consisting of five different types of bibliographic metadata:

1. name and initial of the first author; by Thomson Reuters standardized to the format 'Smith, J'
2. name of the source for publication, by Thomson Reuters processed into a standardized abbreviation format
3. first page
4. number of the volume
5. year

These attributes correspond to the structure of the reference lists as provided by Thomson Reuters. Additional bibliographic data, notably the article title, are not available in the database.

The first step of our algorithm comprises data cleaning and correction of both the references and the articles data. All author names and source titles are transformed to upper case. Blanks and special characters in author names as well as special characters in volume and first page data are removed. Double blanks in source titles are removed. As input for the fault-tolerant part of the algorithm, special characters and blanks in source titles are generally removed in order to compare them as single strings.

The bibliographic data of the Web of Science database comprise four different types of journal abbreviations. In the reference data, journal titles seem to be generally standardized by Thomson Reuters to one of the four types. As first tests showed that sometimes other abbreviations occur nevertheless in references, our algorithm uses all four abbreviations for matching. They are treated by the algorithm as structurally equivalent—the occurrence of another than the designated standard abbreviation is not treated as deviation, but the information is nevertheless captured in the later analysis.

[3] In the conclusion several options to deal with ambiguous matches are discussed.

First part of algorithm development: heuristic definitions

The algorithm's general hierarchical-iterative structure has been defined heuristically as follows: Publication year—source title—first author—first page—volume. The publication year is not varied at all, meaning all references pointing to a specific year are only matched against the articles volume of this year. This decision was mainly made due to performance restrictions, but also following the assumption that otherwise different articles of the same person in different years, but in the same journal with incomplete or erroneous volume and pagination data would be harder or even impossible to distinguish.[4]

The source title abbreviation is defined as second most important attribute as we know that source titles are basically standardized by Thomson Reuters when processing the references.[5] Errors should more easily be identified by the citing authors than in comparison to author names. Author names are defined as third most distinctive attribute, followed by first pages and volumes, as numeric data seem to be most prone for mistakes. First pages are preferred over volumes as they seem more distinctive—they are covering a wider numeric range.

Deviations are allowed by way of different rules. We implemented the relative Levenshtein distance in the algorithm. By this measure similar author names and source titles in the articles are identified if the initial source title cannot be matched at all with the articles set or, in case of the author names, cannot be matched in combination with the source title. The Levenshtein distance calculates the number of insert, delete and replace operations necessary to transform a first string into a second one. The relative Levenshtein distance is calculated as relative indicator with respect to the absolute word lengths of both strings that are compared.

An alternative approach would be to simply match only the first few characters of both source titles and author names, respectively. But this would yield to much more imprecision in these categories. The application of the Levenstein distance will give us the opportunity to allow deviations in a very controlled manner, so that matching to the strings with the most similarity may produce unambiguous results even in combination with other deviating or missing metadata. First tests showed that, if deviating source titles are assigned according to the Levenshtein distance, they tend to be assigned to the correct ones quite precisely—in comparison with author names the absolute number of different source titles in the database is very restricted. Because of this, that source title that is identified by the Levenshtein algorithm as the most similar one is taken, while in case of author names, all variants above a specific threshold are tried in order to find the best match in the whole procedure. This suggests that one should allow source title deviations very early in the iteration process. False source titles are allowed only in the last step of the algorithm—that is, with least priority, as a result of testing with the manually evaluated random sample.

For numeric data, deviations are allowed in terms of thresholds for the numeric value of the deviation as well as in terms of missing single digits.

[4] The results of this study have shown that it is probably worth it to abandon this restriction.
[5] http://thomsonreuters.com/products_services/science/free/essays/cited_title_unification/

Evaluation and further development of the algorithm

Description of the evaluation

The main goal of the evaluation process is to analyse to which extent the fault-tolerant part of the algorithm produces correct matches (true positives) and incorrect matches (false positives) as well as misses correct target articles (false negatives). For this purpose the correct target article to each reference has to be identified. The standardized Web of Science references do not contain the article titles of the cited articles anymore. It is therefore not possible to decide beyond doubt if a matching is correct or not based only on the information contained in the database. Therefore the citing articles that contain the references have been searched on the internet and their full texts have been retrieved in order to capture the respective references in their original formatting and completeness.

A random sample of 500 normalized references is collected following a complete run of the matching procedure on references pointing to 2009 articles with the 2009 articles volume.

The calculation is performed on a database with Web of Science data in a consolidated state dating from spring 2011. This database has been constructed by the Competence Centre for Bibliometrics[6] for the purpose of bibliometric research.

The year 2009 was chosen in order to reflect those recent trends in publication and referencing habits like references to articles in press or advance online articles mentioned above.

As the outcome of the exact part of the matching algorithm is neither relevant for the error analysis nor for the optimization of the algorithm, this partition of the calculated data was skipped before applying the random selection process. Those references that could not be assigned to Web of Science source journals neither as exact phrases nor by the Levenhstein algorithm have been skipped as well: Although it is of course possible that a certain part of source items are missed by the algorithm because of massively wrong source titles in references, the extension of the random procedure to these data would result in a lot of references to real non-source articles which are irrelevant for optimizing the algorithm because they cannot be assigned anyway.

The random sample therefore consists of references that are matched with one of the match keys of the second, fault-tolerant part of the algorithm, and of references, which could be identified as pointing to source journals by this part of the algorithm, but could not be matched to individual articles. By way of random sampling a data set is collected that simulates the underlying complete data volume in terms of, e. g., disciplinary structure. If a reference is cited by more than one article, only one full text is accessed, but if we did not have full text to the first we checked the other ones.

Two restrictions inevitably apply: Firstly, some journals' citation styles do not incorporate the article title. We have excluded these references because, as described, no unambiguous judgement on the correctness of the matchings can be made in these cases. This phenomenon seems to occur predominantly in physics journals.

Secondly, with the exception of open access articles, we depend on library licenses in order to get access to the references in their original completeness and formatting. As most journals are licensed by way of big publisher packages, no bias in terms of high versus low impact journals is

[6] The Competence Center consists of the partner institutions Institute for Research Information and Quality Assurance, Fraunhofer Institute for Systems and Innovation Research ISI, Leibniz Institute for Information Infrastructure and Institute of Science and Technology Studies at the University Bielefeld and is funded by the German Federal Ministry of Education and Research. http://www.bibliometrie.info/en/home.html

to be expected; a systematic disciplinary bias is not to be expected either. Finally, in some cases it was not easy to identify the correct articles even if article titles in the references are accessible, as the metadata, including the article titles, were massively erroneous.

In an effort to enhance the initial data base for the evaluation, further extensive search was carried out and those cases could be dissolved into further identified reference-article pairs as well identified non-source articles.[7] More full text articles could be identified as well by way of extensive search as well.

Thus, on the basis of intellectually identified correct reference-article pairs, an error analysis is performed by collecting and coding all errors in the sample's references in relation to the correct target articles.

The bibliographic metadata of the references in the database and the bibliographic metadata of the respective correct target articles in the database are then compared in detail and a description of the error is captured in a codified manner: In case of deviations in strings, the positions of the deviating letters are determined, in case of numerical deviations the respective numerical value.. If metadata is missing or completely false, they are codified as wrong.

Second part of algorithm development: data-driven threshold definitions and improvements

The coded values of deviations are used as data base for final specifications of thresholds. After that, the algorithm is being optimized by way of varying specific aspects of the algorithm, iteratively re-calculating the algorithm on the random sample's data and analyzing the respective results in relation to the intellectually defined correct reference-article pairs.

A string deviation for numerical data is defined as matching of otherwise identical numerical data if a single digit is missing.

Initially thresholds for numeric deviations are defined comparatively broad (volumes may differ from the ones in the target articles by a numeric value of 8, first pages by 30, respectively). On the basis of the error analysis in the random sample, all errors in the respective metadata categories are being taken into consideration and the final threshold is now defined according to them. We tried to identify a peak in order to include the vast majority of errors and to exclude a tail of increasingly haphazard and large deviations from the original. Otherwise very broad thresholds enhance the possibility of false positive matchings. Both—volume and first page threshold—are thus set at the definite value of 5.

The calculation of the Levenshtein distance is very time consuming—comparing every source title with every other in case no exact match has been reached is impossible when calculating whole volumes. Therefore, as starting basis for the evaluation of the algorithm and in order to be able to initially distinguish probable source items from non-source items, a predefinition is set, which is as broad as possible while basically enhancing the performance: Source titles are compared with all other source titles having the first letter in common. Author names are compared with all author names having the first letter and the respective source title in common.

As it is still necessary to improve the performance, further thresholds for the application of the Levenshtein distance are specified, again on the basis of the error analysis. Possible thresholds are more than one common character in the beginning, or a maximum value of different word lengths.

[7] In some cases target articles that have been identified by the algorithm as probable source items are actually not source items because they are meeting abstracts in source journals that are themselves not indexed.

Based on the coding of positions of deviating letters in source title strings, we identify a threshold of 3 characters on the beginning of the string that the two strings have to have in common in order to be compared by the Levenshtein algorithm. Additionally, we define a maximum value of different string lengths of 7, also on basis of the coding in the error analysis.

In case of deviations of author names, a threshold of 2 common characters is analogously defined as condition to enter the Levenshtein procedure, apart from having the source title in common.

A similarity threshold of 0.8 is defined for the Levenshtein distance based on testing different values and comparing the results in terms of precision.

Tests showed that, even given the fact that the Levenshtein distance is calculated as an indicator relative to the respective absolute word lengths, especially Chinese names with less than 6 characters are matched to a great number of equally short strings with identical family names and varying first name initials. Thus, another precondition is additionally set: names need to have a minimum length of 6 characters (including first name initials) in order to enter the Levenshtein procedure.

After the thresholds are fitted, the algorithm is re-calculated for the data of the random sample.

It is then evaluated, how many unique correct matches and unique false-positive matches as well as non-unique matches are resulting from the algorithm. The major criterion for the evaluation is the maximization of unique correct matches together with the minimization of unique false-positive matches.

The error analysis is used as a starting point for further evaluation of the algorithm by varying and testing specific aspects. The analysis also gave rise to possibilities to further differentiate and partly stretch the algorithm's fault-tolerance without losing precision.

The results obtained from this analysis included the following:

First pages have the highest number of wrong cases, followed by volumes, publication years, first authors and sources: The initial order of first pages and volumes is now reversed in the algorithm, which enhances the number and precision of unique matches. Before skipping the author name, a further match key is implemented that combines both numeric data and the third letter of the respective author names.

While initially wrong source titles have not been allowed in the algorithm, now an extension is programmed and tested including several further iterative match keys: It is first checked if a match can be achieved if the fourth letter of the source string and volume and first pages are identical (as counterbalance for the Levenshtein thresholds), then if only both numeric data are identical and then if the fourth letter and either volume or first page are identical. Before arriving at this constellation, different aspects like the order of numeric data and the usage of a source string letter have been tested against each other.

Source titles have then the highest number of deviating cases, followed by first authors, first pages and volumes. The order of deviations in numeric data and author names is thus reversed and the sample recalculated while allowing deviations in author names before deviating and wrong first pages and then volumes. But as wrong volume and first pages do not co-occur very often with deviating author names, there is no significant change of the precision: The results only change with respect of one less correct unique match. The order is therefore not changed for the time being.

Findings

Of the 500 initial references, eventually 351 references could be accessed in full text in their original formatting. Of those, 65 were excluded because of missing article titles while in one case a correct reference in the full text could not be identified. The data basis of the evaluation finally consists of 285 references. In case of false positives and apparently false negatives we tried to identify the correct target article and to perform the error analysis with respect to this one. While target articles that could not be found in the Web of Science database were excluded at first, the searching was then extended to the internet and the articles are coded as non-source articles if that could be assured. No errors are being coded in these cases. In case of non-unique matchings it is analysed if any of the multiple target articles that the match key identifies is correct.

Table 1. Description of the fault-tolerant algorithm and number of matches.

Deviation Source: In the beginning of the fault-tolerant part of the algorithm all reference source titles that did not match source journals enter the Levenshtein procedure. The second match key identifies references with deviant source titles exclusively, whereas in all following keys deviating source titles are allowed, but correct source titles may also occur
Deviation Author: A deviation specified by way of the Levenshtein distance occurs
Numeric Deviation Volume and First Page: A deviation in the range of +- 5 occurs
String Deviation: The reference has a missing digit
Exchange First Page with Volume [Volume with First Page]: The citing first page is matched with the cited volume; the respective citing volume is not matched
Exclusion: The respective category is not being matched

Description	Number of non-unique Matches	Number of unique Matches
2. Deviation Source		17
3. Deviation Source, Numeric Data without Characters		18
4. Deviation Source, Numeric Deviation First Page		5
5. Deviation Source, String Deviation First Page		5
6. Deviation Source, Numeric Deviation Volume		4
7. Deviation Source, String Deviation Volume		2
8. Deviation Source, Numeric Deviation Volume and Numeric Deviation First Page		1
9. Deviation Source, Numeric Deviation Volume and String Deviation First Page		
10. Deviation Source, Numeric Deviation First Page and String Deviation Volume		
11. Deviation Source, String Deviation Volume and String Deviation First Page		
12. Deviation Source, Exclusion First Page	9	26
13. Deviation Source, Exclusion Volume		13
14. Deviation Source, Exclusion First Page, Numeric Deviation Volume	1	2
15. Deviation Source, Exclusion First Page, String Deviation Volume		
16. Deviation Source, Exclusion Volume, Numeric Deviation First Page		1

17. Deviation Source, Exclusion Volume, String Deviation First Page		
18. Deviation Source, Exchange Volume for First Page, Exclusion First Page		
19. Deviation Source, Exchange First Page for Volume, Exclusion Volume		
20. Deviation Source, Deviation Author		10
21. Deviation Source, Deviation Author, Numeric Deviation First Page		1
22. Deviation Source, Deviation Author, String Deviation First Page		1
23. Deviation Source, Deviation Author, Numeric Deviation Volume		
24. Deviation Source, Deviation Author, String Deviation First Page		
25. Deviation Source, Deviation Author, Numeric Deviation First Page, Numeric Deviation Volume		
26. Deviation Source, Deviation Author, Numeric Deviation Volume, String Deviation First Page		
27. Deviation Source, Deviation Author, Numeric Deviation First Page, String Deviation Volume		
28. Deviation Source, Deviation Author, String Deviation Volume, String Deviation First Page		
29. Deviation Source, Deviation Author, Exclusion First Page	1	3
30. Deviation Source, Deviation Author, Exclusion Volume		1
31. Deviation Source, Deviation Author, Numeric Deviation Volume, Exclusion First Page		1
32. Deviation Source, Deviation Author, String Deviation Volume, Exclusion First Page		
33. Deviation Source, Deviation Author, Numeric Deviation First Page, Exclusion Volume		
34. Deviation Source, Deviation Author, String Deviation First Page, Exclusion Volume		
35. Deviation Source, Deviation Author, Exchange Volume with First Page, Exclusion First Page		
36. Deviation Source, Deviation Author, Exchange First Page with Volume, Exclusion Volume		
37. Deviation Source, Exclusion Volume, Exclusion First Page	9	37
38. Deviation Source, Deviation Author, Exclusion Volume, Exclusion First Page		2
39. Deviation Source, Third Letter Author		3
40. Deviation Source, Exclusion Author	7	24
41. Fourth Letter Source		4
42. Exclusion Source		10
43. Fourth Letter Source, Exclusion First Page		3
44. Fourth Letter Source, Exclusion Volume		

The most relevant result is the fact that unique matches lead to correct results to a great amount. Of 194 unique matches, the fitted algorithm produces 183 correct reference-article pairs, the precision ratio is 94%.

Table 2. Structure of Matches.

Description			**Total Number**
	Number unique Matches	**Number non-unique Matches**	
Match	194	27	221
Match is correct; resp. one of the target articles is correct	183	24	
Match is wrong (false positive), correct target article is missed	2		
Match is wrong (false positive), correct target article is not in database	5		
Match is wrong (false positive), correct target article has another publication year	4	3	
No Match			
No match (false negative), compared with correct target article	9		64
No match, correct target article is not in database	12		
No match (false negative), correct target article has another publication year	43		
All evaluated references			**285**
Reference entails no article title and is not evaluated			65
Correct reference could not be identified unambigiuously in the full text			1
No access to full text			149
All references in the random sample			500

In the case of the eleven false positive unique matches nine correct target articles could be identified by the algorithm having false publication years or not being in the database. In the two remaining cases the first page and the author are wrong. The respective incorrect target articles have been identified by match keys with higher priority (deviating first page in the first case and the combination of deviating author and numeric values in the latter case).

In the case of false positive non-unique matches the publication year is wrong in three cases.

The false negatives are highly characterized by wrong publication years. In thirteen cases it could be ascertained that the original target article was not in the database. There are two different reasons for this phenomenon: On the one hand, in our consolidated version of Web of Science dating from spring 2011 there are some articles of 2009 missing which have been added to the

database only very lately. In some other cases, the references point to meeting abstracts that are not regularly indexed for Web of Science, although their journals are correctly identified as source journals by the algorithm. In the remaining seven cases the references entail error combinations of author names and deviating or wrong volumes and first pages.

The error analysis gives evidence that false publication years cause an essential share of wrong matches as well as non-matches. However, they co-occur very significantly with other errors: In 25 cases, at least two other deviating or wrong metadata occur in combination with wrong publication years. These cases might quite probably result from citations to articles in press. If the algorithm will be extended with respect to wrong publication years, it would be feasible to not allow any other deviations together with wrong years in order to keep the precision high.

Conclusion and outlook

One aspect of the results of the evaluation is especially striking: The possibility to achieve a high recall and precision rate by a very fault tolerant approach which allows combinations of up to four wrong and deviating metadata.

As mentioned above, it could be a possibility to improve both recall and precision by including an almost exact matching procedure with data of other publication years—this could be tested with regard to performance. Probably it would suffice to expand the exact matching procedure only to the respective following publication year in order to at least match the citations to advance online articles with deviant publication year.

Apart from that, a fault-tolerant approach to matching references would very probably always lead to a small share of incorrect matches when, as happened here, a combination of errors in wrong and correct target articles occur that is converse to the specified hierarchy—either in terms of the hierarchy of metadata or in terms of the priority of deviation before error.

The algorithm as specified so far is feasible for calculating whole volumes of Web of Science albeit the iteration of altogether more than 40 match keys and notably the Levenshtein algorithm is time consuming. The calculation of all references pointing to articles of 2009 takes about two days.

A drawback of the performed evaluation is certainly the fact that the random sample and finally the number analyzed references is quite small. The differences in the results of various tests that have been performed in order to extend and fit the algorithm are partly lying only in the range of very few matches. This is because the majority of the analyzed references are not highly multiply erroneous and therefore can be assigned solidly. With regard to the enormous time effort to analyze data intellectually, it is unfortunately not feasible to repeat this study with much larger samples.

Because of the fact that references in Scopus entail article titles, it could be an option for the medium term to adopt the algorithm to Scopus and make use of the article titles. Comparing single article title words of uniquely matched reference-article pairs of the fault-tolerant part of the algorithm could serve as approximated indicator for the truth of the match, respectively.

Further research will focus on the comparison of the results of our exact and fault-tolerant match key with those of the Thomson Reuters match key, with the intention to analyze the quantitative differences in large scale.

The occurrence of ambiguous or non-unique matches raises the question in which respect they should be attributed or counted. It would be of course misleading to attribute them in the same way as unique matches. It would be possible to count them fractionally following a probability approach, but we would consider it more informative to use them as a complementary or error indicator of matches which can be assigned only on the level of source journals of the database.

In this regard it would possibly make sense to reduce them by those which are resulting of those match codes at the end of the algorithm that do not apply the Levenshtein procedure anymore—as in these latter cases theoretically it cannot be guaranteed anymore that all of the non-unique target articles have the same source title.

We will also be able to calculate any of our other bibliometric analyses, such as evaluations of single institutions or countries, with our exact and fault-tolerant match key in comparison with the Thomson Reuters citation counts. Thus it can be analyzed how different match keys might actually change bibliometric results.

References

Buchanan, R.A. (2005). Accuracy of Cited References: The Role of Citation Databases. *College & Research Libraries*, 67 (4), 292-303.

Braun, T, Glänzel, W. & Schubert, A. (1985). *Scientometric Indicators*. Philadelphia: World Scientific.

Dinkel, W.P. (2011). How Do Matchkeys Affect Citation Counts? First Steps Towards an Error Calculus for Bibliometric Indicators. *Proceedings of the ISSI 2011 Conference* (pp. 175-180).

Fellegi, I & Sunter, A (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64 (328), 1183-1210.

Gomatam, S., Carter, R., Ariet, M. & Mitchell, G. (2002). An Empirical Comparison of Record Linkage Procedures. *Statistics in Medicine*. 21 (10), 1485-1496.

Kondrak, G. (2005). N-Gram Similarity and Distance. *Proceedings of the 12th International Conference on String Processing and Information Retrieval* (pp. 115-126).

Lawerence, S., Giles, L.C. & Bollacker, K.D. (1999). Autonomous Citation Matching. *Proceedings of the Third International Conference on Autonomous Agents*, Seattle, Washington, May 1-5, ACM Press, New York.

Lok, C., Chan, M. & Martinson, I.M. (2001). Risk Factors for Citation Errors in Peer-Reviewed Nursing Journals. *Journal of Advanced Nursing*, 34 (2), 223-229.

McLellan, M.F., Case, L.D. & Barnett, M.C. (1992). Trust, But Verify. The Accuracy of References in Four Anesthesia Journals. *Anesthesiology* 77 (1), 185-188.

Moed, H.F. & Vriens, M. (1989). Possible Inaccuracies Occurring in Citation Analysis. *Journal of Information Science*, 15 (2), 95-107.

Moed, H.F. (2002). The Impact-Factors Debate: The ISI's Uses and Limits. *Nature*, 415 (6873), 731-732

Moed, H.F. (2005). *Citation Analysis in Research Evaluation.* Springer: Dordrecht.

Newcombe, H.B., Kennedy, J.M., Axford, S.J. & James, Y.P. (1959). Automatic Linkage of Vital Records. *Science*, 130 (3381), 954-959.

Sweetland, J.H. (1989). Errors in Bibliographic Citations: A Continuing Problem. *Library Quarterly*, 59 (4), 291-304.

Synnestvest, M.B. (2007). *Data Preparation for Biomedical Knowledge Domain Visualization: A Probabilistic Record Linkage and Information Fusion Approach to Citation Data*. Thesis,. http://handle.net/1860/2532