Journal of Informetrics 4 (2010) 194–200

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Combining commercial citation indexes and open-access bibliographic databases to delimit highly interdisciplinary research fields for citation analysis

Andreas Strotmann[a,*], Dangzhi Zhao[b]

[a] School of Public Health, University of Alberta, 13-103 Clinical Sciences Building, 11350 - 83 Avenue, Edmonton, AB, Canada T6G 2V2
[b] School of Library and Information Studies, University of Alberta, Edmonton, AB, Canada T6G 2J4

ARTICLE INFO

Article history:
Received 2 September 2009
Received in revised form
28 November 2009
Accepted 7 December 2009

Keywords:
Field delimitation
Citation analysis
Bibliometrics
Information science
Multiple databases

ABSTRACT

Field delimitation for citation analysis, the process of collecting a set of bibliographic records with cited-reference information of research articles that represent a research field, is the first step in any citation analysis study of a research field. Due to a number of limitations, the commercial citation indexes have long made it difficult to obtain a comprehensive dataset in this step. This paper discusses some of the limitations imposed by these databases, and reports on a method to overcome some of these limitations that was used with great success to delimit an emerging and highly interdisciplinary biomedical research field, stem cell research. The resulting field delimitation and the citation network it induces are both excellent. This multi-database method relies on using PubMed for the actual field delimitation, and on mapping between Scopus and PubMed records for obtaining comprehensive information about cited-references contained in the resulting literature. This method provides high-quality field delimitations for citation studies that can be used as benchmarks for studies of the impact of data collection biases on citation metrics, and may help improve confidence in results of scientometric studies for an increased impact of scientometrics on research policy.

1. Introduction

Field delimitation is normally the first step in citation analysis studies of a research field – the process of (1) identifying a set of research articles that represent this research field for a given time period, and (2) composing bibliographic records for this set of papers that contain all the information needed for citation analysis. The collective view of the authors of the articles that represent the field, as indicated by their citing behavior, is then analyzed with regard to which documents or authors they find most useful for their research and how these documents or authors are related to each other. The bibliographic records that are collected in this step therefore need to include explicitly indexed reference lists of these articles with sufficient information for the intended citation analysis, and currently, Thomson Scientific's Web of Science and Elsevier's Scopus are the only well-structured databases available for this purpose, neither of which is open-access, though.

Since they are the only well-structured citation indexes available, most if not all citation analysis studies have so far relied on these commercial citation indexes exclusively, mostly Web of Science and sometimes Scopus. These databases, however, have a number of limitations (Harzing & van der Wal, 2008) – limited or biased coverage, limited subject indexing

* Corresponding author.
E-mail addresses: andreas.strotmann@ualberta.ca (A. Strotmann), dzhao@ualberta.ca (D. Zhao).

1751-1577/$ – see front matter © 2010 Published by Elsevier Ltd.
doi:10.1016/j.joi.2009.12.001

support, incomplete and noisy cited-reference data – that limit their usefulness for citation analysis, especially when it comes to studying emerging and highly interdisciplinary research fields such as nanotechnology or stem cell research that have recently attracted considerable interest among science and technology policy researchers.

Large-scale open-access citation indexes that could provide reliable and well-structured data for citation analysis do not currently exist, unfortunately, although initiatives and projects that may eventually lead to such a resource have been ongoing and improving, such as Google Scholar, arXiv.org, RePEc.org and some other digital libraries.

To overcome difficulties in using the commercial citation indexes, researchers like Zitt and Bassecoulard (2006) developed sophisticated algorithms to combine citation analysis, keyword search, and document clustering techniques to delineate the nanosciences literature using the Web of Science databases. Here, by contrast, we describe an approach that combines the strengths of a commercial citation index with a well-established high-quality open-access bibliographic database instead of working within the confines of a single database. This approach does not require sophisticated algorithms such as those above in order to overcome limitations inherent in Web of Science or Scopus.

Ingwersen and Christensen (1997) is an early study that combines several commercial databases within the framework of a single commercial database online search provider, Dialog, to overcome coverage limitations they observed when delimiting a field of physics using the Science Citation Index (SCI) alone, and describe methodology they developed to deal with the problem of removing duplicate records that are contributed to the result set by more than one database using Dialog queries. Our field delimitation method addresses a different field delimitation issue, namely, how to complete the records of a field delimitation obtained from a bibliographic database with the full cited-reference information required by citation analysis studies. We thus use multiple databases for different purposes – the bibliographic database to identify and collect citing papers, and the citation index to identify cited references – while Ingwersen and Christensen (1997) used multiple databases for the single purpose of collecting citing paper records.

Zhao and Strotmann (2008) used a relatively small-scale hand-completed dataset extracted from Scopus for an all-author co-citation analysis study of the information science field, while Schneider, Larsen, and Ingwersen (2009) used an exceptional dataset of bibliographic records with fairly complete cited-reference information for a field of engineering to study the advantages of all-author co-citation analyses. In this paper, we show how to create citing-cited datasets of similar quality for a wide range of fields, in particular for biomedical research fields.

In this paper, we discuss some limitations of the two main commercial citation indexes and describe our approach for overcoming them in the context of a case study that combines Scopus, a commercial citation index, and PubMed, a public and open-access bibliographic database, in order to retrieve the data required for an all-author co-citation analysis in the stem-cell research field, a highly interdisciplinary biomedical research area. While the particular method we discuss here will work only for studies on biomedical research areas, as it relies on the PubMed database for bibliographic information, it generalizes in a straightforward manner to other fields, e.g., engineering, for which large-scale bibliographic databases exist, albeit with different success rates from the ones we can report here.

2. Limitations of commercial citation indexes for field delimitation for citation analysis in emerging and highly interdisciplinary research fields

The two main citation indexes have a number of limitations in their respective levels of support for citation analysis studies. Web of Science in particular is deemed to have a limited coverage of emerging interdisciplinary fields because of its emphasis on “most important” (and therefore established) journals or through other collection biases and database limitations – see Harzing and van der Wal (2008) for a comprehensive and current review of issues. Articles retrieved from Web of Science using sophisticated algorithms like in Zitt and Bassecoulard (2006) and Bassecoulard, Lelu, and Zitt (2007) may therefore yield a sufficient sample for some citation analysis studies, but may not be complete enough for large-scale citation mining or for studies that aim to shed light on particularly subtle research policy issues. In addition, this database only indexes the first of authors of each cited reference, which results in incomplete data for citation mining, especially for author citation analysis studies in highly collaborative fields.

For a traditional well-defined research field such as mathematics or sociology, articles in a set of core journals in this field published during a period of time are often used to delimit a field, and this method is well-supported by the two citation indexes. This method, however, does not work well for emerging, interdisciplinary or multidisciplinary research fields because research articles in these fields are published in a wide range of journals that do not have a clear core. For example, stem-cell research is published in journals from a wide range of areas of the biomedical sciences because stem cells are fundamental to the workings of each and every organ of the human body and consequently implicated, on the one hand, in a wide range of diseases as well as, on the other hand, showing promise as a research tool and as a therapeutic vehicle in just as wide a range of areas. In addition, stem-cell research logistics require significant support from bioinformatics and biomedical engineering.

As Zitt and Bassecoulard (2006) show for the nanosciences, and Suomela and Andrade (2005) for stem cell research, keyword searching does not generally work well to delimit these emerging and interdisciplinary fields, either. In addition, subject classifications are generally not detailed enough in these commercial citation indexes to help with delimiting research fields of this type.

Compared to Web of Science, Scopus is less incomplete in terms of data on cited authors as it indexes up to 8 cited authors of each reference. This is important for conducting author-based citation analyses. Scopus also has more extensive

coverage in many research fields. For example, Scopus claims to cover the entire PubMed database plus conference papers for the last ten years or so, likely making it the most complete database for recent research in the biomedical research fields. However, unlike Web of Science, Scopus does not support – in fact, explicitly forbids and actively discourages – the systematic download of large search result sets. There is no provision, for example, to download any search results of more than 2000 records, a limit currently put on the maximum number of records for output. Work-arounds do exist, such as searching one year and a few journals at a time, but they only work for relatively small research areas and turned out to be hopeless for such a large, active and diverse research field as stem-cell research, in which in the order of 10 000 refereed papers are published annually, spread across hundreds of journals.

3. Combining PubMed and Scopus as a basis for citation analysis of the stem-cell research field

In contrast to the situation with citation indexes, large-scale, high-quality open-access bibliographic databases for research articles do exist, and the PubMed/MEDLINE database of the U.S. National Library of Medicine (NLM) is a prime example. While it does not provide cited-reference information, its detailed and high-quality medical subject headings provide an excellent basis for field delimitations (in fact, bioinformatics research such as (Suomela & Andrade, 2005) uses the subject classification as a benchmark to aspire to), and with a mandate to build a comprehensive database of biomedical research and the financing to back it, its coverage of its target literature is exemplary (see below for details). Furthermore, it supports large-scale downloads and processing. Consequently, techniques for mining this knowledge base are the topic of a sizable number of bioinformatics publications.

In order to delimit the stem cell field for citation analysis, i.e., to collect a set of records of articles in the stem-cell research field that is complete as well as clean, i.e., which includes as many articles as possible in this field along with their complete references, but as few articles as possible on topics outside of this field, we carried out a three-step mapping process between PubMed and Scopus: (a) using PubMed's MeSH term search facility to obtain citing papers, (b) mapping these papers to Scopus in order to obtain their cited references; and (c) mapping the Scopus cited references to PubMed both to obtain complete data on cited references and to disambiguate (i.e., clean) these references.[1]

3.1. Obtaining a complete and clean set of citing articles on stem-cell research

PubMed/MEDLINE is a service of the U.S. National Library of Medicine (NLM) that includes over 18 million bibliographic records from medical and other life science journals for biomedical articles back to 1948. PubMed provides rich bibliographic data created by information professionals or provided by journal publishers, and maintained by the NLM. At its core, MEDLINE is the NLM's premier bibliographic database that contains records on journal articles in the life sciences with a concentration on biomedicine. A distinctive feature of MEDLINE is that the records are extensively and manually indexed with NLM's Medical Subject Headings (MeSH).

Suomela and Andrade (2005), the latter of whom was a primary investigator of the Canadian Stem Cell Network at the time, studied the recall and specificity of hundreds of potential search terms for the field of stem cell research. In the context of the present study their results imply that the PubMed MeSH term search provided an excellent field delimitation of the field, while the interdisciplinary nature of the field appeared to make it impossible to construct a keyword search strategy that would match its quality.

Relying on their results, we therefore conducted a search for "stem-cell" in MeSH terms in PubMed, confident that the search results would be considerably cleaner and more complete compared with results from keyword or subject searching supported by Web of Science or Scopus. We retrieved more than 10 000 records for 2007 alone (i.e., 10 915), and were able to download the full search result set in XML format, all at once – a feature that is not supported by the commercial citation indexes such as Web of Science or Scopus.

3.2. Mapping PubMed search results to Scopus

We then developed Java programs to create search strings from these PubMed search results which identify records for the same articles in Scopus. To avoid time-out errors in the Scopus database, each search string aims to retrieve 500 articles from Scopus, although the maximum number of downloadable results in Scopus is 2000 in principle.

Simply put, the Scopus query we construct for a single article looks for its title or starting and ending page numbers, for last names of its authors, for its source journal's ISSN or title, and for volume and issue numbers:

(ISSN or SRCTITLE) and VOLUME and ISSUE and AUTHORLASTNAMES and (TITLE or (PAGEFIRST and PAGELAST)).

TITLE and SRCTITLE fields are normalized. Here is a random example:

(ISSN(0007-1048) OR SRCTITLE("british journal of haematology")) AND VOLUME(111) AND ISSUE(2) AND AUTHLASTNAME(moscardó) AND (TITLE("graft-versus-tumour effect in non-small-cell lung cancer after allogeneic peripheral blood stem cell transplantation.") OR (PAGEFIRST(708) AND PAGELAST(710)).

[1] Step (c) as described here significantly improves on a simpler method reported in (Strotmann et al., 2009).

As just mentioned, each of the search queries we created aimed at 500 articles, i.e., an actual search query submitted to Scopus consisted of 500 single article search strings as above connected with the Boolean OR operator.[2]

To comply with Scopus licensing limitations, which forbid automated downloads, we manually issued these search queries in Scopus using its "Advanced Search" facility, one at a time, and downloaded the complete records of the search results, which included data on cited references. For the 2007 stem cell research data set, 10 693 of 10 915 (or 98%) of the search results from PubMed were found in Scopus. This confirms Scopus' claim that it provides excellent coverage of medical journal articles, lends credence to its claim to include all of MEDLINE, and at the same time indicates that our search algorithms worked quite well, with a miss rate of 2% if we assume that Scopus does include a full 100% of PubMed/MEDLINE as it claims.

We also developed a Java program that removed false positives from the Scopus search results – records retrieved from Scopus that were not included in the original PubMed download. This type of noise was very rare (less than 1%).

Very recently, Scopus added the PubMed ID of a journal paper to its bibliographic records, and made this a searchable field. The above search strategy can thus be greatly simplified in the vast majority of cases; however, initial experiments indicate, surprisingly, that the hit rate using a search strategy purely based on the PMID field is significantly lower than the search strategy above. This new Scopus feature allows us to improve the accuracy of both the search and the subsequent mapping between records retrieved from Scopus and records retrieved from PubMed for those cases where the PMID field is available in the Scopus record and it matches a PubMed record with the same ID. The above search strategy as well as the above algorithms for matching search results to PubMed and filtering out false positives are still necessary for the remaining records.

3.3. Mapping Scopus cited references to PubMed records

At this point, some citation analysis studies will already have completed their field delimitation, as the combination of PubMed citing records and Scopus cited references, which include up to eight author names (surname and initials), publication year and title, and journal name, volume, and number, is quite extensive already. However, there are still limitations of the cited-reference data that other studies may need to overcome.

Scopus data on cited references include up to 8 of the authors of each cited reference, but are quite inconsistent in terms of author names, article titles, journal names, etc. – five to ten different versions of the same reference are not unusual. Retrieving full bibliographic records of these cited references from Scopus was not an option as it was effectively prohibited by Scopus licensing terms.[3] We therefore made use of the PubMed Batch Citation Matcher[4] for this purpose.

We developed Java programs to create search strings from Scopus' data on cited references that look for records of the same articles in PubMed to be issued to the PubMed Batch Citation Matcher. We chose to use this tool rather than directly searching PubMed for a number of reasons. For one, the Batch Citation Matcher handles really large queries quite nicely, speeding up the process considerably. Second, with the exception of the journal name and publication year fields, this tool ignores a field's value if all other fields match – a feature that is useful for correcting erroneous spellings or page numbers found in the Scopus data. The citation matcher thus importantly acts as a disambiguation mechanism for Scopus cited references, and makes it easy to detect duplicates in the original Scopus data as well: the 414 976 cited references that were successfully matched against PubMed corresponded to 215 372 distinct articles in PubMed, i.e., each cited paper was cited 1.9 times on average.

Mapping Scopus cited references to PubMed records using Batch Citation Matcher can therefore not only complete data on cited references but also help with cleaning (i.e., disambiguating) Scopus cited-reference data because different formats of the same cited reference will be mapped to the same PubMed record. Especially for author citation analysis studies, this completes the cited references very well, as these records include a complete list of all authors, mostly with their full names.

The structure of a query for the Batch Citation Matcher is simple:

Journal name | publication year | volume | starting page |first author name |

As an example, "Arterioscler Thromb|1994|14|25|Ahlswede KM|" is the query for this cited reference found in Scopus: Ahlswede, K.M., Williams, S.K., Microvascular endothelial cell sodding of 1-mm expanded polytetrafluoroethylene vascular grafts (1994) Arterioscler Thromb, 14, p. 25.

The PubMed citation matcher returns the PubMed identifier (PMID) corresponding to this article if it is found, or the string "NOT_FOUND" otherwise. These PMIDs were then used to retrieve the corresponding full records.

We manually issued the above search queries to the PubMed Batch Citation Matcher by email in blocks of several thousand each, and obtained in return the PMIDs of these articles, also by email. At first, about 90% of the cited references were successfully matched this way.

We then attempted to improve the hit rate by modifying and reissuing the queries for those 10% articles that had not been found. We first resubmitted the queries that returned NOT_FOUND, but with the journal name left blank, which identified

[2] Note that Web of Science limits the number of search terms per query to 500. A corresponding search strategy would therefore be able to retrieve only a fraction of this number of articles per query from that database, were it possible in that database at all.

[3] Scopus licensing terms are available at http://www.scopus.com/scopus/standard/termsandconditions.url.

[4] The PubMed Batch Citation Matcher is located at http://www.ncbi.nlm.nih.gov/entrez/getids.cgi.

55% of the previously unmatched articles. Secondly, we resubmitted the remaining unmatched queries with journal name in place but (a) leaving out the publication year or (b) increasing or (c) decreasing the year of publication by one. These three queries together identified a further 8% of those cited references that had not been found in the previous two steps. In all, the 433 133 cited references contained in the Scopus records for the 2007 stem cell research field literature yielded 414 976 matches in PubMed using this method, with a total hit rate after these three steps of 96%.

This is an excellent result: Persson (2001) found that only 10% of the cited references from Information Science publications found in the Social Sciences Citation Index were to papers indexed in that database – i.e., using only the traditional data source, SSCI, his corresponding "hit rate" was 10%, compared to 96% in our case. Of course, SCI's coverage of the biomedical literature is presumably considerably greater than its coverage of information science, so that a study like Persson's in the stem cell field would likely have netted a significantly higher "hit rate", but we are not aware of any studies that attempt to verify this in the biomedical sciences. Our own informal experiments indicate that matching from cited reference to full record is a difficult and error-prone task within SCI, yielding large numbers of false positives or negatives due to limitations in its search facilities, which makes it hard to estimate the attainable performance of a method equivalent to ours that uses only the facilities of the SCI database. As for Scopus, its licensing terms forbid its use in this way due to the size of the dataset we deal with.

In a final step of this mapping, we wrote a small Java program to automatically download the full PubMed records of these cited references, as long as they were not to documents in the citing set. Using the NCBI EFetch[5] facility – one of its Entrez Programming Utilities available for use with PubMed – this program automatically retrieved the XML formatted full records for all matched cited references, in blocks of 500 records.[6]

3.4. Creating input files for citation analysis

In a final step, the data we collected this way were assembled by a Java program we developed into a form that supports citation analysis. The full PubMed records that corresponded to cited references were merged with the original PubMed download of citing papers as the two overlap substantially, especially when a long time period is studied. The citing-cited relationships were recorded in a separate file during this merging process.[7] For those references that were not found in PubMed, such as books, we kept data extracted from Scopus cited-reference strings but converted them into an XML format similar to that of the PubMed records.

The three-stage mapping process thus produced three XML documents: one that consists of the original download of citing papers from PubMed and of the PubMed records for cited references, one document for those references from Scopus whose full records were not found in PubMed, and one document for the citing-cited relationship between these papers. These documents can serve as data for all kinds of citation mining studies, and have the following advantages comparing with data obtained from any of the data sources alone.

- They are cleaner (e.g., the citing-cited relationship is clarified) and more complete (e.g., including full author information for nearly all citing papers and cited references, a requirement for all-author co-citation analysis studies).
- Papers are normalized and uniquely identified to a more extensive degree.
- They are in XML format for easier parsing and processing.

In the case of our 2007 data set, a search for "stem-cell" in MeSH terms in PubMed retrieved 10 915 records published in 2007, 10 693 of which were found in Scopus using our methodology. These Scopus records included 433 133 cited references, i.e., roughly 40 cited references per citing paper on average. 414 976 of these cited references were identified with a record in PubMed, and found to refer to a total of 215 372 distinct papers.

4. Discussion

These are excellent results, showing a significant improvement over earlier related attempts (Strotmann, Zhao, & Bubela, 2009). They indicate that it is possible, at least in biomedical research fields, to come quite close to the ideal of a field delimitation for citation analysis studies: to find all refereed publications in that field, each judged by a professional to belong to that field, missing very little of the field's literature, and to do so in such a way that the literature records include complete cited reference and other information relevant to a range of citation analysis studies.

PubMed does an excellent job at putting together a comprehensive collection in this sense of all biomedical research publications – the fact that 96% of all cited references in our dataset were to documents found in PubMed attests to the comprehensiveness of the database. In addition, its MeSH term assignments appear to be recognized as a gold standard of

[5] Efetch is described at http://eutils.ncbi.nlm.nih.gov/entrez/query/static/efetch_help.html.

[6] Again, we could have used larger blocks in principle, but found that errors tended to creep into the download results if we used significantly larger block sizes.

[7] This was possible because the PubMed ID of the citing paper that contained each reference was recorded in the PubMed Batch Citation Matcher queries, so that it was not necessary to include the Scopus results in this phase.

subject categorization within those fields, certainly in the bioinformatics literature (e.g., Suomela & Andrade, 2005) with corroboration from library and information scientists (e.g., Funk, Reid, & McGoogan, 1983; Qin, 2000), which means that we can use them for high-quality field delimitation within the biomedical literature. PubMed does not provide cited references, though.

Based, therefore, on a high-quality field delimitation using PubMed, we find that the Scopus citation index provided full records for 98% of the documents that comprise this field, including cited references. Of the cited references identified by Scopus in this set, 96% were identified as referring to publications found in PubMed, from where we were able to disambiguate, clean, and complete the cited reference information provided originally by Scopus. Together, these success rates may be interpreted to mean that we managed to construct a document citation network that likely contains in excess of 95% of all citation links in the "real" literature of this field – i.e., we are quite close to the "real" citation network that defines this field's 2007 (refereed citing) literature and the references it contains.

While it can be expected that this level of quality is not always necessary for performing valid scientometric studies, there are a number of scenarios where it may prove quite valuable.

Completeness of the delimited literature can play a significant role in the design and evaluation of science metrics themselves, for example. In particular, by comparing the outcomes of identical metrics using different data sources with different completeness characteristics for the same target field, we can analyze the robustness of these measures with respect to incompleteness and/or biases in the data collection methodology. In this scenario, a complete dataset would serve as a benchmark against which to measure the accuracy and reliability of metrics computed from less complete datasets.

If the ultimate goal of a research project is to find hard statistical evidence of subtle scientometric effects, e.g., of research policies on research outcomes, it will likely be necessary to employ sufficiently sensitive rather than robust metrics. The level of completeness of the dataset used for such a study then becomes a determinant for the accuracy of measurements performed on it: the more complete the dataset, the less likely the sensitive measurements of the target effect "signal" will be buried in the statistical "noise" due to missing data. Especially in scientometric studies, where Zipf-type exponential distributions are common, even relatively small errors in the data may be amplified to a significant error in an analysis.

Finally, even though incomplete and even biased datasets may lead to valid analyses, especially when using statistically robust methods such as author co-citation analysis, higher levels of completeness and accuracy of the underlying data can always be counted upon to increase confidence in the results of an analysis.

5. Conclusions

This paper has discussed some difficulties involved in using existing commercial citation indexes for field delimitation for citation analysis studies, and described an approach for overcoming them in the context of a case study that combines Scopus and PubMed in order to retrieve data for citation analyses in the stem-cell research field, a highly interdisciplinary biomedical research field. This three-step approach used high-quality and detailed subject headings assigned by the NLM to PubMed records to delimit the field of stem cell research, mapped the literature thus identified in PubMed to Scopus records to retrieve cited-reference information on these documents, and then used PubMed services to obtain complete bibliographic records for the referenced documents thus obtained.

For the biomedical fields covered in PubMed, we argue that this is an excellent field delimitation strategy for citation analysis studies, at least for the recent literature. MeSH heading assignments in PubMed are very detailed, and are given by trained information professionals to serve as a reliable resource for the U.S. (and indeed world) medical profession's information needs. As an added boon for the information scientist, PubMed supports large downloads from its database, in an XML format that eliminates many ambiguities.

Scopus claims to cover all of PubMed's content of the last decade or so, and our sample appears to corroborate this claim by mapping 98% of a large sample literature dataset from the latter to the former in order to complete the cited-reference information for the records in this set. By mapping the resulting cited references back to PubMed records (with a success rate of 96%), we were able to both disambiguate and complete the records on cited papers in the citing literature we study here, thus completing the field delimitation that began by identifying the field's (citing) literature in PubMed.

The result is an excellent field delimitation, with a resulting citation network that is likely 95% complete (or better), providing exceptional opportunities for citation analysis and mining.

This confirms earlier results (Strotmann et al., 2009) that indicated a number of significant advantages to a multi-database approach when compared to employing only a single citation index that is used by most citation mining studies. In particular, we again find that, in general, a multi-database approach allows the researcher to circumvent limitations imposed by individual databases that make it difficult to retrieve a complete and clean set of articles and their references to represent a research field being studied. We can amend this to say that, at least for recent biomedical research areas, it appears to be possible to actually overcome and not just circumvent these limitations, and to come quite close to the ideal of a field delimitation with complete cited-reference information.

This multi-database approach has the potential to enable large-scale policy studies using citation analysis methods that are otherwise difficult to accomplish. Most if not all citation analysis studies to date have relied on Web of Science data which have been widely recognized as incomplete in terms of coverage, search facilities and indexing of cited references. Scopus data is less incomplete, but does not allow large-scale data collection. As a result, studies have been limited in terms of the scale and/or completeness of their data, and consequently in terms of the impact of their results, although numerous

sophisticated methods have been proposed to circumvent limitations of the citation indexes. It is through the combination of multiple databases that these underlying difficulties may be overcome, and the impact of citation analysis methods may be enhanced for science policy and management.

The practical approach introduced in the present paper has the potential of serving in a wide range of future studies. To this end, we plan to make the Java programs that we developed in the course of this study available as open source software along with code being developed for additional data cleaning and handling in the course of other ongoing studies, as soon as the code has reached sufficient maturity.

Acknowledgements

The research reported here has been funded in part by the Social Sciences and Humanities Research Council of Canada, by the Stem Cell Network, a Canadian Network of Centres of Excellence, and by Genome Canada and Genome Prairies. The authors wish to thank the anonymous reviewers for insightful comments. The authors would also like to thank Gencheng Guo for programming the software and algorithms they designed and developed, and for collecting the data for this study.

References

Bassecoulard, E., Lelu, A., & Zitt, M. (2007). A modular sequence of retrieval procedures to delineate a scientific field: From vocabulary to citations and back. In *Proceedings of the 11th international conference of the International Society for Scientometrics and Informetrics* ISSI: Madrid,

Funk, M., Reid, C. A., & McGoogan, L. S. (1983). Indexing consistency in MEDLINE. *Bulletin of the Medical Libraries Association*, *71*(2), 176–183.

Harzing, A. W. K., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*, 61–73, doi:10.3354/esep00076.

Ingwersen, P., & Christensen, F. H. (1997). Data set isolation for bibliometric online analyses of research publications: Fundamental methodological issues. *Journal of the American Society for Information Science*, *48*(3), 205–217.

Persson, O. (2001). All author citations versus first author citations. *Scientometrics*, *50*(2), 339–344.

Qin, J. (2000). Semantic similarities between a keyword database and a controlled vocabulary database: An investigation in the antibiotic resistance literature. *Journal of the American Society for Information Science*, *51*(3), 166–180.

Schneider, J. W., Larsen, B., & Ingwersen, P. (2009). A comparative study of first and all-author co-citation counting and two different matrix generation approaches applied for co-citation analysis. *Scientometrics*, *80*(1), 105–132.

Strotmann, A., Zhao, D., & Bubela, T. (2009). A multi-database approach to field delineation. In *Proceedings of the 12th International Conference of the International Society for Scientometrics and Informetrics, July 14–17* Rio de Janeiro, Brazil, (pp. 631–635).

Suomela, B. P., & Andrade, M. A. (2005). Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics*, *6*, 75, doi: 10.1186/1471-2105-6-75.

Zhao, D., & Strotmann, A. (2008). Comparing all-author and first-author co-citation analyses of information science. *Journal of Informetrics*, *2*(3), 229–239.

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing and Management*, *42*, 1513–1531.