A Multi-Database Approach to Field Delineation

Andreas Strotmann,[1] Dangzhi Zhao[2] and Tania Bubela[3]

[1]*Andreas.Strotmann@ualberta.ca,* [3]*tbubela@ualberta.ca*
School of Public Health, University of Alberta, Edmonton, AB, T6G 2G3 (Canada)

[2]*dzhao@ualberta.ca*
School of Library and Information Studies, University of Alberta, Edmonton, AB, T6G 2J4 (Canada)

Abstract

Bibliometric studies have long relied on single citation databases as data sources for their field delineation phase, although a few have gone so far as to compare results obtained through the use of several databases separately. In this paper we report on an effort to systematically combine a number of databases to collect metadata records for the literature that defines a research field. In three case studies, we have combined information retrieved from PubMed with metadata from Scopus, and employed the NCBI Entrez Genes database to delineate research areas. At the core of the multi-database field delineation approach that we describe are: first, a method for matching full records back and forth between the Scopus and PubMed databases (95% or higher match rate) augmented by a method for matching Scopus cited references to PubMed records (90% success rate); and second, the wealth of information and services available from NCBI that is connected to individual PubMed records through its various Entrez database connections. Our approach allows us to combine strengths and overcome weaknesses of multiple databases, for an excellent coverage of the target area – interdisciplinary biomedical research.

Background

While field delineation has always been an integral part of bibliometric and scientometric studies, and indeed has long been recognized as a fundamental and largely unsolved problem of scientometrics (e.g., van Raan, 1996; Zitt, 2006), it has only recently become a research area of its own (e.g., Bassecoulard & Zitt, 2007). With a burgeoning interest of science and technology policy researchers in highly interdisciplinary fields such as nanotechnology or stem cell research, traditional field delineation methods – complex keyword search strategies or "definitive" journals or author oeuvres – appear to have reached a limit. Researchers like Zitt and Bassecoulard (2006) overcome these limitations through sophisticated algorithms that combine citation analysis and keyword searches to delimit the nano-sciences literature using ISI, while Suomela & Andrade (2005) developed software to rank all papers in PubMed by the degree to which they are relevant to the core stem cell research field.

However, in both the traditional and the more recent examples found in the bibliometric literature, the literature that defines a field for a given study is still taken from a single data source. This is true even for those studies that compare different data sources as to, say, their relative coverage of a field (e.g., Aksnes et al., 2000; Lundberg, 2007; Zhao, 2005).

In this paper, by contrast, we show how, by combining the strengths of several databases instead of working within the confines of a single one, we can improve traditional field delineation methods without requiring the extra algorithmic complexity of other approaches.

Research Network Sibling Literature Using Scopus and PubMed

In this section we will discuss in detail our experience with the multi-database method of field delineation in a network analysis of Canadian stem cell research.

The ultimate goal of the study which begins with the field delineation described here is to analyze the impact of funding structures and research policy on the evolution of a national, but geographically dispersed collaborative research network; the degree to which that national network is embedded internationally; and finally the impact of other policies such as encouraging and facilitating commercialization on the collaborative research environment.

Proceedings of ISSI 2009, pages 631–635.
Edited by B. Larsen and J. Leta.

Methodologically, the full study (Bubela & Strotmann, 2008) began by defining a literature that is both similar to, and inclusive of, the research outputs produced by the interdisciplinary Canadian Stem Cell Network (SCN). No traditional field delineation methods worked in this context, which necessitated developing a new approach similar to, but simpler than, that of Bassecoulard, Lelu & Zitt's (2007) field delimitation method.

Outline

Starting from a *seed* set of all publications funded by the SCN and provided by the SCN, we retrieved full metadata on those seed documents that were available in Scopus. We also retrieved full metadata for all papers indexed in Scopus as citing each seed document. The literature for the international research network that both embedded and performed meaningfully similar research to the Network was then defined as the union of (a) the seed documents, (b) those papers that were cited by the seed documents, (c) those papers that cited the seed documents, and (d) those papers that were cited by those papers that cited the seed documents (siblings). The set of citing documents in this literature was thus comprised of subsets (a) and (c), while the set of cited documents in this literature was comprised of subsets (b) and (d), the latter of which contains almost all of (a). For the cited references we retrieved full metadata from PubMed, both in order to complete and in order to disambiguate the cited reference information found in Scopus records.

The remainder of this section describes the methodology and the results in detail.

Seed documents

The SCN provided us with a list of all publications funded through the Network since its inception in 2001 until 2007. The list included scientific journal publications, conference presentations, theses and dissertations, as well as articles in the popular press. The list contained 1,409 documents, including duplicates.

We retrieved the full bibliographic records that corresponded to the publications listed in this database from two bibliographic databases – Elsevier's Scopus and the U.S. National Library of Medicine's PubMed. For this purpose, we created complex search queries for these databases using simple Python scripts. The Scopus search yielded 507 distinct document records of scientific journal publications funded by the SCN.

Sibling literature

In addition to records for the seed documents, we retrieved full bibliographic records including cited references for all publications that were indexed in Scopus as citing a seed document. These searches were run in the week of November 27, 2007 and yielded 8,678 (non-unique) full Scopus records. After adding the seed set of SCN publications as retrieved from Scopus and removing duplicates, a total set of 6,996 distinct records of documents indexed in Scopus was used as the basis for subsequent steps.

The sibling literature of the seed set of SCN funded papers was defined as those papers that were cited in those documents that included references to seed papers, as well as all seed documents. We also included all publications referenced by SCN seed publications.

The roughly 7,000 distinct citing documents in our data set contained 418,023 cited references in total, or about 60 references per paper on average. For these cited references, we needed full bibliographic records (except their cited references). Scopus licensing terms[1] prohibited us from retrieving this information from that database, so we made use of the National Library of Medicine's Batch PubMed Citation Matcher[2] for this purpose.

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

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

First, we created software to parse the cited references included in records downloaded from Scopus, and to extract from these cited references the fields required by the PubMed citation matcher. The resulting search, split into blocks of a few thousand queries each, was submitted to the Batch Citation Matcher. Results showed a hit ratio of about 80% of all cited references in our data set. For the remaining 20% or so unmatched references, we resubmitted the query without journal name, increasing the total hit rate to 90% of all cited references (377,505 of 418,023). This is an excellent result: Persson (2001) found that only 10% of the cited references from Information Science publications found in SSCI were to papers indexed in that database – i.e., using only the traditional data source, SSCI, his corresponding "hit rate" was 10%, compared to 90% in our case. A "hit" in this case was defined as either an exact match or a set of up to four "ambiguous" matches as returned by the PubMed citation matcher ("ambiguous" matches were returned very rarely). Except for the journal name and the publication year fields, this matcher ignores a field's value if all other fields match – this feature was useful for correcting erroneous spellings or page numbers found in the source dataset. The citation matcher thus importantly[3] acted as a disambiguation mechanism for cited references, and made it easy to detect duplicates in the original dataset as well: the 377,505 references that were successfully matched against PubMed corresponded to 162,555 distinct articles in PubMed, for an average citation rate of 2.3 references per cited paper.
The citation matcher only provided us with a list of PubMed identifiers (PMIDs) of the documents that matched the cited references in our original dataset. In a final step of our data collection we therefore retrieved the corresponding full records from PubMed. To this end, we wrote another small program to create a series of PubMed queries for the records with these PMIDs. Since it was possible to run these queries in blocks of a few thousand, we were able to download the more than 150,000 resulting XML records within less than a day.

Field delineation results

The resulting field literature data set we collected thus consisted of roughly 7,000 full records (including cited references) retrieved from Scopus, and more than 150,000 full records retrieved from PubMed. All of the full Scopus records had an exact match with a PubMed record in this set, and 90% of the cited references listed in the Scopus records were successfully matched against PubMed.

Bioinformatics-Based Field Delineation

In a series of ongoing large-scale science and technology policy studies that expand upon the Stem Cell Network study outlined above, we are taking the multi-database approach to field delineation a step further by incorporating bioinformatics databases. In these studies, we aim to determine the recent scientific literature on mouse genomics and specific mouse genes. This research again addresses the policy question of funding and commercialization policies on patterns of publication and research collaboration. In associated research, we have created a detailed patent landscape of the mouse genome. We are therefore ultimately interested in the impact of gene patents on publication network measures.
Gene names are notoriously hard to search in literature databases (Yoneya, 2005), and restricting our search to mouse genes would have been practically impossible given the fact that almost all mouse genes have orthologues in humans and other mammals. However, the NCBI's Entrez Genes database catalogues extensive information on genes of a number of species, including references to the literature that corroborates or is the source of each piece of information on a gene. We are thus able to solve this particular type of field delineation problem by (a) identifying all (relevant) mouse genes in Entrez Genes; (b) following the link

[3] Cited reference strings in Scopus are not highly standardized – five to ten different versions are not unusual.

provided by Entrez to the literature referenced in these gene entries in PubMed; and (c) applying any additional restrictions on the literature we are interested in (e.g., time of publication or country of author affiliation) in PubMed.
As in the case of the SCN study described above, we expand the literature for a particular mouse gene to its sibling literature, requiring a mapping from PubMed to Scopus followed by the same series of steps undertaken in the SCN study – in all, truly a multi-database approach to field delineation.

Preliminary results

At the core of the practical applications of our multi-database field delineation methodology, we have been developing tools and methods for mapping back and forth between search result sets retrieved from PubMed and from Scopus; in the latter case, this includes mapping both full records and cited references to their corresponding full PubMed records.
We have successfully applied an early version of this methodology to the literature related to research conducted with funding from the Stem Cell Network (SCN), a Network of Centres of Excellence funded jointly by the three Canadian federal research councils (natural sciences and engineering (NSERC), medicine (CIHR), and humanities and social sciences (SSHRC)) as well as by Industry Canada, the federal Canadian agency charged with advancing Canadian industry in the widest sense. The research network is interdisciplinary – besides its core biologists, medical researchers and clinicians, it includes biomedical and tissue engineers, computer scientists, and a sizable contingent of researchers interested in the legal, ethical, and other social aspects of stem cell research.
Having retrieved the SCN-funded literature and its citing literature with full records including cited references from Scopus (see above for details), we successfully mapped a full 100% of those roughly 7,000 records to equivalent records in PubMed. Of the more than 400,000 cited references contained in these Scopus records, we were able to match 90% to equivalent PubMed records using our own parser for Scopus cited references and a two-step application of the PubMed Batch Citation Matcher. We are currently undertaking other studies in which we map PubMed records to Scopus records, i.e., going in the opposite direction. In this case, preliminary results indicate success rates greater than 95%.
At least for the recent biomedical literature, our methods for combining Scopus and PubMed search results into a joint field delineation data source have been remarkably successful. The fact that 90% of cited references in our sample dataset were successfully matched to a PubMed record indicates that these two combined data sources are quite complete.

Discussion

This paper has explored the potential value of combining several databases during the field delineation phase of a scientometric study. Preliminary results from a series of large-scale applied scientometric studies, in which we combine the strengths of several databases, indicate a number of significant advantages to this multi-database approach when compared to traditional scientometric field delineations that employ only a single data source:
(a) In general, a multi-database approach allows the researcher to circumvent limitations imposed by individual databases that make it difficult to conduct the intended field delineation. In some cases, a multi-database approach is required to delineate the intended research field because no single database contains all the information that defines the intended literature.
(b) The multi-database approach may also help simplify an otherwise highly complex search strategy while retaining the ability to retrieve all the metadata required for a study.
We have been employing the multi-database field delineation approach with some success in an ongoing series of large-scale science policy studies in health biotechnology, where we

have combined the National Library of Medicine's PubMed citation database and its accompanying on-line services with the National Center for Bio-Informatics' gene databases and with the Scopus citation index.

Potential Significance of Findings

We expect these results to hold some significance in two directions.

First, the multi-database approach to field delineation has the potential to enable large-scale policy studies using bibliometric and scientometric methods that are otherwise difficult to accomplish – it is not an accident that we developed these methods in this context. Especially the practical methods introduced here for delineating a field by expanding a well-defined seed literature to its citation-sibling literature and fully expanded cited-reference metadata has the potential of serving in a wide range of future studies.

Second, we hope that this study will encourage those who influence policy making within Scientometrics to start lobbying for improved access to data sources. In the biomedical and physical sciences, efforts are under way to develop research policies that emphasize open access to publications and research outputs, particularly those produced using public research funding. In the field of genomics, this will hopefully result in a wide range of scientific databases available and interoperable with the NCBI databases as discussed above.

A similar discussion within the Scientometrics and research policy research areas is due, with the goal of making more data available to policy researchers and analysts than is available today. As we have seen here, it is important to enhance the direct interoperability of databases, both in terms of cross-cutting user interfaces and in terms of common data structures and authority files. The NCBI Entrez suite of databases again serves as an examplar for this approach. The limitations of today's citation indexes, trapped as they are in a commercial database vendor environment, need to be overcome for scientometrics to become a field with real-world research policy applications.

References

Aksnes, D.W., Olsen, T.B., & Seglen, P.O. (2000). Validation of bibliometric indicators in the field of microbiology : a Norwegian case study. *Scientometrics*, 49(1), 7-22.

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 ISSI 2007,* ISSI: Madrid.

Bubela, T., & Strotmann, A. (2008). Designing Metrics to Assess the Impacts and Social Benefits of Publicly Funded Research in Health and Agricultural Biotechnology. In: The International Expert Group on Innovation and Intellectual Property (2008). *Toward a new era of intellectual property: from confrontation to negotiation*. Available online at www.theinnovationpartnership.org/data/ieg/documents/cases/TIP_Innovation_Metrics_Case_Study.pdf

Lundberg, J. (2006). *Bibliometrics as a research assessment tool : impact beyond the impact factor.* Dissertation, Karolinska Institutet: Stockholm.

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

van Raan, A.F.J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics,* 36 (3), 397-420

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

Yoneya, T. (2005). PSE: a tool for browsing a large amount of MEDLINE/PubMed abstracts with gene names and common words as the keywords. *BMC Bioinformatics,* 2005 (6), 295.

Zhao, D. (2005). Challenges of scholarly publications on the Web to the evaluation of science - A comparison of author visibility on the Web and in print journals. *Information Processing and Management*, 41(6), 1403-1418

Zitt, M. (2006). *Scientometric indicators : a few challenges*. Unpublished. http://eprints.rclis.org/6306/

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