

Abstract proposal for **European Network of Indicator Designers (ENID) Conference in Rome, 7th-9th September 2011**

Session: "Coverage and completeness of research output in data sources", organized by Tim Engels, Centre for R&D Monitoring (ECOOM), University of Antwerp, and Gunnar Sivertsen, NIFU, Oslo.

Comparison of international publication data bases with the publication registers at Finnish universities

Yrjö Leino, CSC - IT Center for Science Ltd., P.O. Box 405 FI-02101 Espoo, Finland

Hanna-Mari Puuska, CSC - IT Center for Science Ltd., P.O. Box 405 FI-02101 Espoo, Finland

CSC — IT Center for Science Ltd is administered by the Finnish Ministry of Education, Science and Culture. CSC provides IT support and resources for academia, research institutes and companies.

Background

The Finnish Ministry of Education and Culture is actively exploring the use of publication and citation data for the purposes of gathering information on the amount and on the quality of research pursued at the Finnish universities. In co-operation with universities and other public research organizations the Ministry is currently developing a national publication register.

Presently, the 16 Finnish universities collect the data of their publications in their own publication databases, which are based on different solutions. The publications are reported annually to the Ministry.

In order to achieve a good and reliable coverage of the academic research activities, the lists of publications delivered by the universities will in the future be compared to international publication data bases (Thomson Reuters ISI Web of Science and Elsevier Scopus) and completed whenever missing publications or data are found.

In this study, we shall report the results obtained by a ministerial working group in a pilot work with Finnish universities, where the data in two Finnish universities' publication registers in 2009-2010 were compared with ISI Web of Science data (WoS). We will also outline plans for further work.

Research subjects and Hypotheses

In the first phase of our study, the data is gathered from the University of Tampere (UTA) and Aalto University's School of Science and Technology (AaltoTech, former Helsinki University of Technology). Both universities' publication registers are used for collecting information on all scientific, artistic and societal activities by the university's research staff. UTA's database SoleCRIS is provided by a commercial company, Solenovo Oy. AaltoTech's database is developed and maintained by the university's library. Data on both universities' research activities are publicly available in the universities' web pages.

The other data set is the list of publications received from Thomson Reuters ISI Web of Science as an answer to a query with the author's address specified as University of Tampere or Aalto University's School of Technology.

The primary object of our study is to find out to what extent the entries in Web of Science are present in the university's data. As the data in the universities' own registers are based on the researchers' own reporting activity, we do not expect them to have complete lists of publications. We will also investigate possible reasons for missing publications in the universities' registers.

The secondary aim is to explore the share of publications in the universities' registers not covered by the WoS data. As known, the publications in medical and natural sciences, which are traditionally oriented towards publishing in international journals, are well presented in the WoS, whereas in other disciplines, the coverage is weaker (see, *e.g.* Moed, 2010). In social sciences and humanities, books and national publications play an important part (*e.g.* Hicks 2005), and in engineering the research results are often published in conference proceedings (*e.g.* Glänzel et al. 2006). UTA is a multidisciplinary university with strong focus on social sciences and humanities (48 % of the university's total expenditure in 2009) and medical sciences (30 %) (source: KOTA database). Meanwhile, AaltoTech is strong mainly on fields of engineering and natural sciences. Thus, peer reviewed scientific articles in international journals form only a minor part of all the entries in these universities publication registers. Accordingly, it would be unreasonable to expect that Web of Science could cover major parts the universities publications.

Thirdly, because the ultimate goal is to handle the publication data for all Finnish universities in a reliable, yet economic way, we also wished to create and test a process for identifying publications that appear in either one (or both) of the Finnish data sets and WoS data. The process is highly automated, thus eliminating the need for error prone manual work.

During the first half of 2011 we will continue the work by studying other universities as well and taking into comparison also the Thomson Reuters raw data on Finnish publications. Thus, we

will have the opportunity to deduce the affiliations directly from the full addresses instead of trusting on the affiliation information given by Web of Science.

Methods

The challenge was to identify a publication from a data set once it has been picked from the other one. For the identification of publications we have applied three separate methods:

- 1) Digital Object Identifier (DOI)
- 2) A set of exactly matching fields of data records (ISSN of a journal/ISBN of a book, volume, issue and page number) combined with an approximate matching of the title
- 3) A partly matching set of fields of data (the same fields as in method 2, and additionally the surname of the first author) combined with an approximate matching of the title, but now with more stringent conditions than in method 2.

A computer program was written to implement these methods. After identification of matching publications, the counts of overlapping and missing publications of the two data sets were calculated in order to analyze the coverage of WoS and missing publications in universities' data.

Results

When writing this abstract, complete data for both universities for the period 2009-2010 were not yet available. Thus, we are here presenting the results concerning only UTA's publications in 2009.

The size of UTA's own publication data comprised 2076 items, whereas the WoS data had 670 entries. Of the 670 publications assigned to the University of Tampere in 2009 by Web of Science, 508 (76 %) were present also in the university's own register. Of the missing 162 publications, 62 were classified as "Meeting Abstracts" which are not covered by UTA's register. With 21 publications classified in other smaller categories, there were all in all 79 "Articles" in Web of Science of which the university had no record.

There are a couple of reasons for the omissions. First, the search for publications in Web of Science may have included articles produced by authors who are not directly affiliated with the university, but who, nevertheless, have included the University of Tampere in the article's address list. The registration of publications in UTA's database is limited only to researchers whose salary is paid by the university. Secondly, the researchers are themselves responsible for keeping the register up-to-date, and this can lead to missing articles when, for instance, a researcher leaves the university for another job.

As for the identification methods, we found DOI very reliable and useful. Of the 508 publications present in both data sets, 441 were correctly identified through DOI, with no false positive identifications. On the other hand, there were only three publications with differing DOI given in the data sets, and in all of these cases the DOIs differed by one character only, indicating that these were probably just typing errors.

Not all publications had a DOI in both data sets, and thus method 2 was also useful and identified 59 additional publications. Method 3, which was designed to complement method 2 to cover for the cases when there are errors or even omissions of fields in data, brought us the final matching 8 publications.

A more detailed analysis on the types of publications and on the disciplines they represent revealed clear variations in the coverage offered by Web of Science. At the general level, of all publications written in English and classified as peer reviewed scientific articles in UTA's own register, 65 % were also present in WoS.

References

1. Glänzel, W., Schlemmer, B., Schubert, A., Thijs, B. (2006): Proceedings literature as additional data source for bibliometric analysis. *Scientometrics* 68(3): 457-73.
2. Hicks, D. (2005): The four literatures of social science. In Moed, H.F.; Glänzel, W.; Schmoch, U. (eds.) (2005), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems.* Kluwer Academic Publishers, Dordrecht.
3. KOTA database. Online database of statistical data on Finnish universities.
<https://kotaplus.csc.fi/>
4. Moed, H.F., 2010. *Citation Analysis in Research Evaluation*, Springer. ISBN 978-90-481-6938-2.