A Fully Automated Method for the Unification of Funding Organizations in the Web of Knowledge[1]

Daniel Sirtes* and Mathias Riechert*

*sirtes@forschungsinfo.de; riechert@forschungsinfo.de
iFQ – Institute for Research Information and Quality Assurance, Schützenstraße 6a, D-10177 Berlin (Germany)

Introduction

As of August 2008 Thomson Reuters includes funding acknowledgements in their Web of Knowledge (WoK) database. One of the major obstacles in using this resource for assessing the output of funding organizations (FO) is the vast amount of aliases included in the funding organizations list (over 10000 entries for the German Research Foundation (DFG)). Numerous further problems with FO data have been discussed previously (see Costas & Yegros-Yegros, 2013; Rigby, 2011a, 2011b; Sirtes, 2013; Yegros-Yegros & Costas, 2013). This proof-of-concept paper presents a highly efficient, precise and fully automated method with minimal manual configuration to unify many of these aliases and almost all of the publications associated with a funding organization.

One of the things that we have learned from our semi-automated method developed for cleaning the DFG data was that many aliases include only the sub-programme of the DFG instead of the German Research Foundation itself (see (Sirtes, 2013)). However, we have also realized that in many cases these aliases appear together with the DFG acronym (e.g. DFG cluster of excellence). This circumstance led to the idea of fishing for the different names and sub-programmes of a funding organization with the help of its acronym(s).

Method

The approach incorporates findings from previous studies on funding acknowledgement data in the WoK with the help of the in-house database developed by the Competence Center for Bibliometrics. In order to further stimulate the debate on which steps should be included in automated FA data cleaning, they are described explicitly:

As previously proposed (Sirtes, 2013; Wang & Shapira, 2011), we first create a thesaurus of funding organization aliases.

(1) Get all WoK funding organizations aliases that were used in more than 80 publication items.

(2) Extract abbreviations out of the funding organization text by using a regular expression that selects strings with at least two capital letters. The resulting list was manually reviewed, and 23 combinations were blacklisted as they are not funding organizations (e.g. USA).

(3) Extract a list of unified short funding organization texts (USFO) from the original funding organization field by removing commas, brackets, hyphens and the abbreviations themselves. Furthermore, we only consider terms that include at least one blank space, as we are searching for the long form of the abbreviations or full

[1] This work was supported by the German Federal Ministry of Education and Research (BMBF, project number: 01PQ08004A).

names of sub-programmes. Additionally, for each of the USFO, the dominant country of all items stating that FO is computed, to ensure that the term is not used by multiple organizations in different countries[2]. Out of all 549 unified funding organization texts, two had differing dominant countries: “Chinese Academy of Sciences (LAMOST)” (with China and USA) and “Cancer Institute” (with USA and Australia).

(4) Assign the 549 USFO to the 668 abbreviations by identifying terms containing both the funding organization text and the abbreviation (for example: “Deutsche Forschungsgemeinschaft (DFG)”. This results in a thesaurus with abbreviation-USFO mappings ranging from 16 different variants (for example NIH) to only abbreviations (for example IDRIS), where no corresponding text was found. As in the previous step, dominant countries are controlled for to address synonyms. Out of the 668 abbreviation-USFO mappings, 12 differ concerning the dominant country of the mapped USFO (most prominently the America NSF and NSF of China).

Building on the thesaurus the search for funding organizations and publication items stating them is implemented as a VB.net program combining two search strategies. The first strategy searches for the abbreviation with regular expressions. In order to include misspellings in funding organizations (which cumulate to about 40% of all funding organization texts) we additionally compute Levenshtein distances of possible variants of the WoK funding organization text and the USFO texts in the second search strategy.

(5) Search in the WoK funding organization full text (FFO) for the extracted abbreviations using regular expression search patterns. The abbreviation can be surrounded by a non-character letter or build the start or the end of the term. This is computed for all 668 abbreviations. The found FFOs are then inserted into a mapping table, if the abbreviation has a common dominant country (see last step). For those abbreviations with multiple dominant countries, only the second search strategy can be applied in order to prevent semantic mismatching.

(6) Search in the WoK FFO for the extracted USFOs: Each USFO text is split and each part-term’s first two letters are used as the basis for the regular expression search. Searching for “Deutsche Forschunsgemeinschaft” for example, uses the following regular expression string: “(^|\s|[[:punct:]])(De\w*.Fo\w*)”, meaning that the funding organization can be at the start of the term, after a blank space or punctuation character. This results in a wide range of different terms, which are possibly misspelled variants of the USFO. Then we compute the Levenshtein-wordlength-ratio (LWLR) to assess the closeness to the USFO. Out of the found funding organizations texts, only funding organizations with a LWLR<0.4 or 0.5 (different variants were tested) are added to a matching table. Again, multiple dominant countries are controlled for.

Finally, the results of both strategies are combined by matching those USFO according to the three possible dominant country combinations:

(7) If the mapping of the USFO to the abbreviations has only one dominant country (276 of 668 cases): In this case, all FOs of a unified short form have the same

[2] A dominant country is the country, which is most often associated with publications of a funding organization. We have also flagged the FOs with low share of the dominant country and whitelisted FOs that have a multinational nature, like the ERC.

dominant country[3]. Both search strategies can therefore be combined. Consequently, the mappings from the funding organization to abbreviations and the mappings from the funding organizations to the USFO texts are unified for each abbreviation.

(8) If the mapping of the USFO to the abbreviations has multiple dominant countries (12 of 668 cases): In this case, combining both search strategies would confound different organizations from different countries. Therefore, the USFO mappings from the same dominant countries are combined. Consequently, each FO/dominant country combination gets unified into an abbreviation with a country index (e.g. NSF_CHN for the National Science Foundation of China)[4].

(9) If there is only an abbreviation and no USFO text (381 of 668 cases): Only the abbreviation matches are used to assign funding organizations from the WoK.

Results and Outlook

DFG

We are in the fortunate position of having a complete manually cleaned list of all publications in the WoK associated with the DFG, aided by the semi-automated method described in Sirtes (2013). We have restricted our comparison with the data generated in our current fully automated method to the 21,963 publications from the year 2010. Our new fully automated method has associated 21,072 items with the DFG. Of these 21,072 we found 21,002 again in our manual set, which amounts to a recall of 95.6% and an astounding precision of 99.7%, with the University of Georgia Research Foundation as the top culprit with 14 false positive publications. However, if one compares the success of the method on the basis of FO aliases instead of publications, then the picture looks considerably grimmer, as most items are concentrated in a few aliases. Out of the 3,061 aliases for 2010, only 1,834 have been found, which amounts to a recall of 59.9%. The precision however, is again very high at 98.3% with 31 false positives. The high recall in publications compared to aliases is explained by two factors: First, the 1,227 missing aliases amount only to 2,201 publications, and second, 1,240 of these have a second funding organization alias associated with it, that is included in our set.

NIH

As the German Research Foundation is probably the funding organization with the most diverse list of aliases and is therefore extremely hard to unify, we compared our method to the largest funding organization in the database, the NIH. We used one external bit of knowledge to enhance our search, which is the list of national institutes with other acronyms than NIH itself. We used 26 of the 28 acronyms listed on the front page of www.nih.gov (we left out the two letter acronyms CC and OD). Thus, we have used the 27 acronyms (including NIH) and their associated texts in the most common occurrences to search for aliases in the WoK. We compared this dataset with 200 publications from 2010 that we have randomly picked from the NIH's own database of their publications: the NIH RePORTER (http://projectreporter.nih.gov/reporter.cfm). We have found 188 of these publications in the WoK (WoK recall: 94%). Out of these publications, 163 had a funding organization associated with it (Funding acknowledgement to NIH RePORTER recall of 81.5%, share ofWoK items with FA 86.7%). Our method was successful in finding 155 out of these 163, which amounts to a recall of 95.1%. However, 7 out of the false negatives did not credit the NIH at all (including one "funding organization" called 'Public Service Grant', which might

[3] Or the USFO is whitelisted as multinational.

[4] The FO aliases with less than 80 articles (or in future versions less than 30) cannot be used for this kind of homonymous FO names as of now there is no way to determine their dominant country.

or might not allude to the public health service grants of the NIH). Thus, our method caught 155 out of 156 publications that can be associated manually to the NIH i.e. a recall of 99.3%. A single publication has evaded our method due to the fact that some researchers are rather sloppy with the prepositions in the names of funding organizations, like ‘of’, ‘for’, etc.

To develop this very promising method further, we plan to leave these kind of words out of our regular expression search. Furthermore, we plan to lower our initial starting set to FOs with 30 instead of 80 publications per alias and possibly include grant number patterns in our query.

References

Costas, R. & Yegros-Yegros, A. (2013). Possibilities of funding acknowledgement analysis for the bibliometric study of research funding organizations: case study of the Austrian Science Fund (FWF). In *Proceedings of ISSI 2013* (pp. 1401–1408). Vienna, Austria.

Rigby, J. (2011a). Systematic Grant and Funding Body Acknowledgement Data for Publications: New Dimensions and New Controversies for Research Policy and Evaluation. *Research Evaluation, 20*(5), 365–375. doi:10.3152/095820211X13164389670392

Rigby, J. (2011b). Systematic Grant and Funding Body Acknowledgment Data for Publications: An Examination of New Dimensions and New Controversies for Bibliometrics. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1742845

Sirtes, D. (2013). Funding Acknowledgements for the German Research Foundation (DFG). The Dirty Data of the Web of Science Database and How to Clean It Up. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. Moed (Eds.), *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference* (Vol. 1, pp. 784–795). Vienna, Austria: AIT GmbH.

Wang, J. & Shapira, P. (2011). Funding Acknowledgement Analysis: An Enhanced Tool to Investigate Research Sponsorship Impacts: The Case of Nanotechnology. *Scientometrics, 87*(3), 563–586. doi:10.1007/s11192-011-0362-5

Yegros-Yegros, A. & Costas, R. (2013). Analysis of the web of science funding acknowledgement information for the design of indicators on “external funding attraction.” In *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference, 15th-19th July 2013* (pp. 84–95). Vienna, Austria.