FUNDING ACKNOWLEDGEMENTS FOR THE GERMAN RESEARCH FOUNDATION (DFG). THE DIRTY DATA OF THE WEB OF SCIENCE DATABASE AND HOW TO CLEAN IT UP.

Daniel Sirtes[1]

[1] *sirtes@forschungsinfo.de*
iFQ Institute for Research Information and Quality Assurance, Schützenstrasse 6a, D-10117 Berlin (Germany)

Abstract
Since August 2008 the Web of Science database includes funding acknowledgements information. To date no study has been conducted concerning the data quality of these entries. In this paper, we show the vast array of problems emerging if one wishes to unify all funding organization entries of a large and diverse funding body such as the German Research Foundation (DFG). After enumerating all possible sources of error found by manual sifting through all funding acknowledgement entries of German publications, we introduce a new semi-automated method, in order to facilitate the same cleaning task for future years. The method which uses regular expressions and Levenshtein distance algorithms as building blocks shows a rather good result with precision and recall of 96% and 94%, respectively. With the cleaned data set, two examples are shown of the new possibilities emerging of this kind of bibliometric data. Connecting this information with financial funding data opens up the path to new kind of input-output analysis in the realm of scientific research while corroborating the validity of the funding acknowledgement data.

Conference Topic
Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2) and Science Policy and Research Evaluation: Quantitative and Qualitative Approaches (Topic 3).

Introduction

Structure and content of the funding acknowledgement fields in the Web of Science database

Since August 2008 the Web of Science database (WoS) includes funding acknowledgements. Thomson Reuters is extracting this information from the journal articles and fills the fields of funding organization and grant number. Additionally, it includes the raw extracted acknowledgement text in a grant text field. In the relational database developed on the basis of the raw WoS database by the Competence Centre for Bibliometrics for the German Science System

(http://bibliometrie.info/en/home.html) the structure of these fields is as depicted in Figure 1.

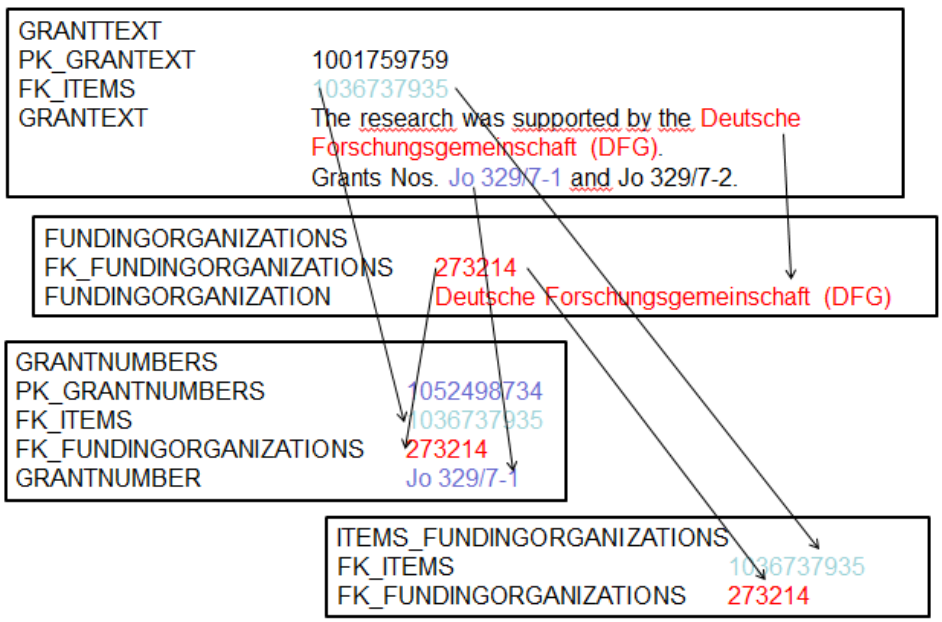

Figure 1. Structure and connections of the funding acknowledgement fields in the database of the Competence Centre for Bibliometrics for the German Science System.

Coverage of funding acknowledgements in the database

As the Competence Centre's database is frozen in week 17 of each year, it is possible to document the dynamics of the inclusion of the funding acknowledgement since its inception. From this information one can see that the amount of items with funding acknowledgements is growing far faster than the growth of the database for the most recent year, suggesting that the extraction methodology of Thomson Reuters is still changing substantially, although the journals' more standardized formatting of the acknowledgement field and more funding acknowledgements in general may also contribute to this growth. Figure 2 shows the count and percentage of journal articles with funding acknowledgements for all three full years of the funding field according to the past two years of the competence centre's database (called WOS2010 and WOS2011, respectively).

The overall coverage depicted above is only an average figure that does not represent the immense diversity in coverage in different disciplines. Table 1

shows that in certain disciplines (assigned by the WoS subject categories (SC)) the share of articles with funding acknowledgements (FA) is very high while in others it is only moderate or even hits zero. The worldwide coverage in these subject categories is juxtaposed with the coverage of articles with German contributions.

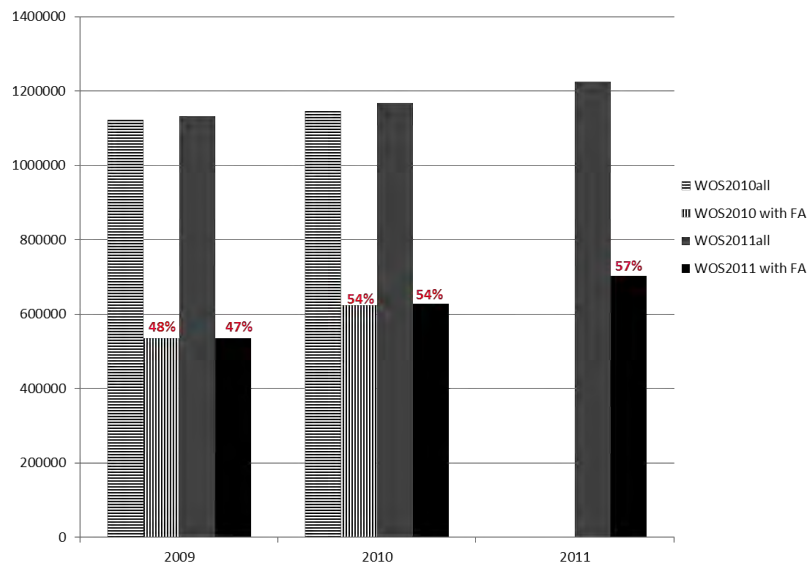

Figure 2. Count of all journal articles and those with funding acknowledgments and their share 2009-2011

Table 1. Coverage of articles in WOS2011 with funding acknowledgements (FA) worldwide and articles with German affiliation (representative selection)

WOS Subj. Cat.	All Articles	with FA	Percent of articles with FA	German articles	German articles with FA	Percent of of German articles with FA
Biology	14065	11524	82%	1082	972	90%
Biochemistry & Molecular Biology	44764	37517	84%	3734	3160	85%
Cell Biology	19558	16298	83%	1962	1669	85%
Ecology	14162	11332	80%	1106	893	81%
Physics, Atomic, Molecular & Chemical	15850	12065	76%	1958	1554	79%
Chemistry, Physical	42967	32165	75%	3459	2678	77%
Materials Science, Multidisciplinary	53242	35790	67%	3753	2542	68%

Physics, Applied	41464	25362	61%	3239	1982	61%
Mathematics	20450	11433	56%	1359	669	49%
Engineering, Chemical	21635	11513	53%	1159	469	40%
Medicine, General & Internal	16481	5777	35%	720	254	35%
Psychology, Experimental	5390	1665	31%	564	224	40%
Economics	14373	1081	8%	1147	67	6%
Humanities, Multidisciplinary	3037	0	0%	54	0	0%
Political Science	4908	0	0%	286	0	0%

This skewed distribution of articles with funding acknowledgements could be contributing to problems of data extraction, but is also consistent with an interpretation that certain disciplines do not have as much external funding as others. This is clearly the case when comparing biological sciences with humanities in general.

Finding Publications funded by the German Research Foundation (DFG)

A simple search and its problems

Finding all the publications funded by the German Research Foundation (DFG) is not a simple task. Thomson Reuters does not unify any of the entries in their funding organization field, which means that every different entry, even if it is only a one letter typo, gets its own identification number as a different funding organization[65]. This problem is multiplied enormously by the following problems.

a. The German Research Foundation has many funding programs (like *Sonderforschungsbereich*, *Emmy-Noether-Programm*, *Exzellenzinitiative*, etc. (for a full list see http://dfg.de/foerderung/index.html)). Very often these funding programs are entered in the grant text and therefore also into the funding organization field and thus is not subsumed under the DFG.
b. Not even the funding program, but rather the funded research facility or network are mentioned (e.g. 'Nanosystems Initiative Munich' or 'Ruhr University Research School').
c. As the name of the German Research Foundation and of its funding programs are originally in German, but many articles translate their name into English (sometimes with their official name, but to a substantial amount also with a creative translation) there are several name variants

[65] The problems of unification for a funding organization has been pointed out in (Rigby 2011) and exemplified for the Swiss National Science Foundation by (Van den Besselaar et al. 2012). However, the complexity of the problem, especially for such a big organization without a standardized system for funding acknowledgements in place, seems to be more daunting than expected (see footnote 3).

even for the same funding program. (Examples of 'creative' translations include 'German Society for the Advancement of Scientific Research' and 'German Academic Research Society').

d. There are a substantial amount of extraction errors which include:
    a. Substitution of the grant number for funding organization (i.e. funding organization 'SFB760')
    b. Co-funded papers appear in the database as a single funding organization (i.e. funding organization 'DFG and NIH')
    c. Severely incorrect extractions of funding organization from the grant text (e.g. from the grant text "…and funding by the GSC 203 for Carolin Schwarz" (which is a graduate school funded by the DFG) the funding organization assigned was 'Carolin Schwarz').

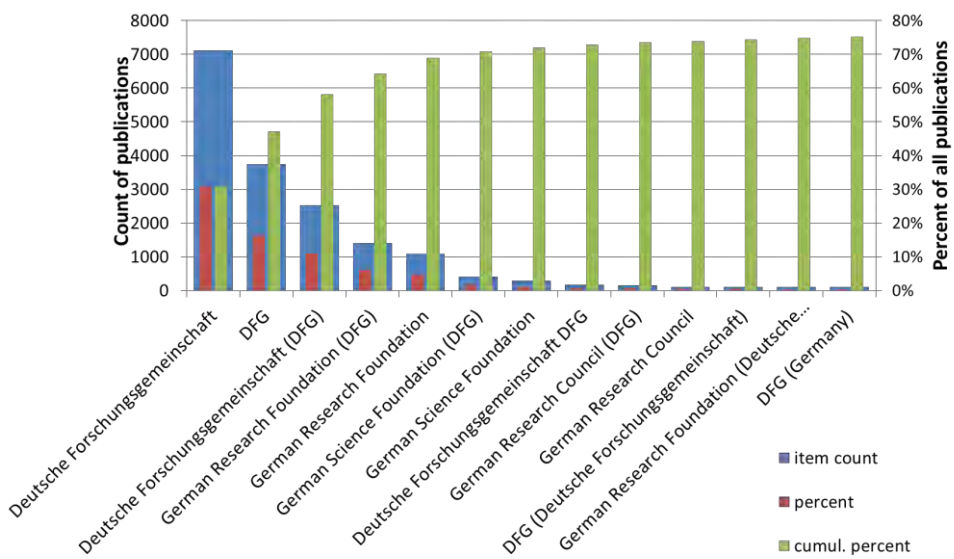

Figure 3. The 13 most common aliases for the German Research Foundation in the 2010 version of the database. Absolute item count, percentage of all publications and cumulative percentage of all publications are shown.

Manual sifting through all German publication

Because of these problems, a first step in finding the DFG funded publications cannot avoid sifting manually through all of German publications for entries in the funding acknowledgement field. (Although some DFG funded publications do not have contributions with a German affiliation, this methodology (restricting the publications to German ones) seems the only feasible one). Many hours of manually comparing the entries in the database with the list of programs funded by the DFG, in harder cases with the help of the grant text and wider internet

searches, has been executed (We would like to thank Simone Falk for her meticulous and excellent work conducting this laborious tasks).
Figure 3 shows the 13 most common entries for DFG funded publications and illustrates the problem with finding all of them. The first six aliases for the DFG cover around 60% of all publications. However, the additional amount of publications per alias flats out very fast and displays a typical power law distribution: Only the first 13 shown here have more than 100 publications per alias. Not more than 87 aliases have at least 10 publications each. Finally, 5747 aliases are associated with only one publication. Thus, the total number of DFG aliases amounts to an astonishing 6370 for the 2010 version of our database.

Development of a semi-automated method for finding aliases in subsequent years

In order to facilitate the search for DFG aliases in the database for subsequent years, a semi-automated method has been developed. With the help of a visual basic script, the results of the manual search has been reproduced. (We would like to thank Mathias Riechert for his help writing the script).
The method has three main components:

a. Regular expressions for the aliases found.
b. Calculation and definition of acceptable levels in Levenshtein distance in order to accommodate orthographical mistakes.
c. A false positive list of aliases that cannot be excluded with regular expressions.

Thus, the first step included finding appropriate regular expressions that are implemented in Oracle SQL in order to capture the aliases found in the manual search. (http://docs.oracle.com/cd/B12037_01/appdev.101/b10795/adfns_re.htm). Examples for these regular expressions can vary in their complexity from ‘for.*gr.*’ for ‘Forschergruppe’ to ‘em.*no\w+.([^ir]\w+)’ for ‘Emmy Noether’.
In a second step, the database entries found with these regular expressions are compared according to a Levenshtein distance algorithm (Levenshtein 1966) in order to calculate the amount of deletions, insertions and substitutions (single-character edits) needed in order to arrive from the found entry to the correct original alias. For example, ‘Forschargruppe’ would have a Levenshtein distance of 1 from ‘Forschergruppe’ as the first e was substituted for an a. In order to achieve uniformity in the algorithm, the Levenshtein distance was calculated as a share of the number of possible substitutions of a string of the same length as the correct entry (The so called ‘Hamming distance’). Thus, the relative Levenshtein distance of the above example is 1/14=0.07, as one out of 14 letters were substituted. The upper bound of acceptable Levenshtein distance was set relatively high with 0.4.
As some of the false positive results of this method were not eliminable with better regular expressions, a list of those entries was compiled in order to subtract

it automatically from the list of the entries found. For example, the California Department of Fish and Game (CDFG or California DFG) will appear in any searches for the DFG. Another example is the Austrian equivalent of the German 'Sonderforschungsbereich' (SFB) (collaborative research center), which uses the same name and abbreviation (e.g. 'Austrian SFB project IR-ON' or 'Austrian Science Fund (FWF) SFB17'). However, we maintained the goal of keeping this false positive list as short as possible which has reached 521 entries. Finally, at some point it did not seem viable to invent new regular expressions for singular entries; therefore 84 aliases were not included into the list for the reproduction of the manual results.
The lists and algorithm was then applied to the 2011 version of the database and yielded the results shown in table 2.

Table 2. Results of the semi-automated method for searching DFG-Aliases

Results		WOS2010	WOS2011
a.	**Levenshtein all**	6807	9550
b.	**Levenshtein *true positive***	6286	8655
c.	**total *false positive***	521	895
d.	**2010 false positive list**	521	521
e.	**false positive not in 2010 list (*new false positive*)**	0	374
f.	**Total true positives with method**	6370	8739
g.	**2010 *false negative* list**	84	84
h.	**Non-Levenshtein (*false negative*)**	84	659
i.	**Non-Levenshtein without 2010 (*new false negative*)**	0	575
j.	**Total DFG aliases**	**6370**	**9314**

Thus, the result of our 2010 method is composed by three lists

a. Levenshtein-list (all results obtained with the regular expression/ Levenshtein script).
c. False positive list (the list obtained by the script resulting in incorrect entries).
g. False negative list (The list of entries not entered into regular expressions).

The resulting list is therefore $a-c+g = f$= 6807-521+84= 6370. As the two false lists could be used for the 2011 application of the method the calculation of precision and recall of the method includes those lists as obtained by the method itself: True positive = $f = b+g$=8739, new false positive = e = 374, and new false negative = i= 575. The precision is therefore 8739/(8739+374) = 96% and the recall is 8739/(8739+575) = 94%. However, as 6370 entries were already set from 2010 one could alternatively calculate the precision and recall of the new entries

in the 2011 database. This yielded the following results: $\text{precision}_{(new)}$ = (8739-6370)/(8739-6370+374) = 86% and $\text{recall}_{(new)}$ = (8739-6370)/(8739-6370+575) = 80%. Considering that the method found 37% more entries in 2011 than in 2010, these results are quite promising.

Portrayal of the cleaned publications set with funding acknowledgements for the German Research Foundation

In order to exemplify the new possibilities of portrayal of the research funded by the DFG and in order to corroborate the validity of the funding acknowledgements data, two preliminary results are presented in the following:

Share of DFG funding by discipline

With the publication set obtained by our method it is now possible to study in which disciplines the German Research Foundation is more or less active. Figure 4 shows a selection of disciplines and the share of German publications with funding acknowledgements and with DFG funding in particular.

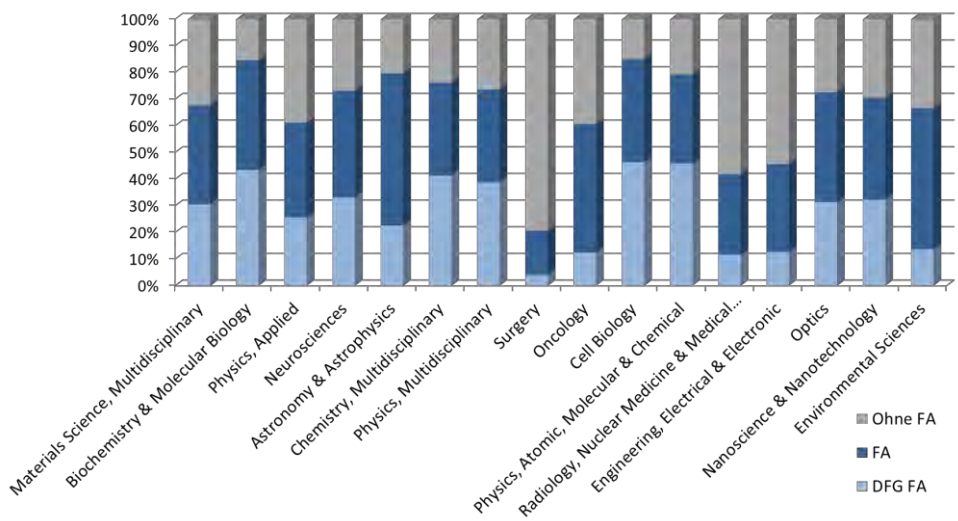

Figure 4. Share of 2010 German publications without, with no DFG, and with DFG, funding acknowledgements, accordingly.

The tendency of the German Research Foundation to fund basic and not applied research which is funded by other means can be directly observed.

Connecting DFG funding acknowledgements with DFG funding amounts

A more elaborate use of the cleaned data set can be obtained by connecting funding acknowledgments with other sources. With the data contained in the DFG issued 'Funding Atlas 2012' (http://www.dfg.de/en/dfg_profile/evaluation_

statistics/funding_atlas/index.html) the financial funding per university and discipline can be inferred. The amount of publications per discipline and German university in 2010 (subsumed in EFI SC super-categories) can then be compared to the funding received from the DFG in the years 2008-2010. Figure 5 shows all publications and all of funding in large German universities[66], while Figure 6 only shows publication and funding in the natural and life sciences. A remarkable correlation can be observed between the two. Although this cannot be considered conclusive evidence as other variables like the size of the universities were not controlled for, it is however noteworthy that in the natural and life science 83% of the variation can be explained by amount of funding received. The lower correlation in the overall picture (R^2=80%) could also be due to different coverage in different disciplines. A hint in this direction is the comparatively low output of Aachen TH, a technical university and the known lower coverage in technology and engineering publications in the WoS database.

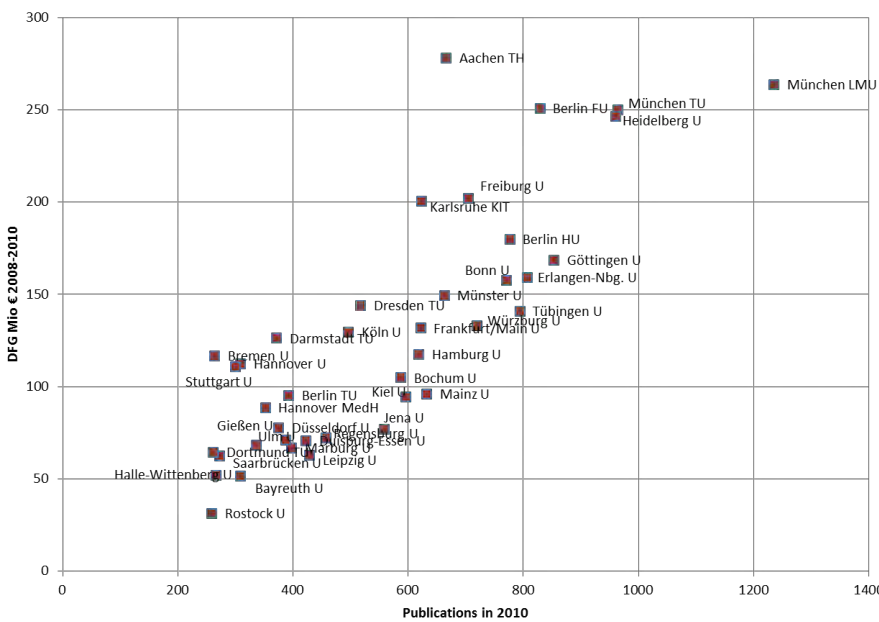

Figure 5. Comparison of all publications funded by the German Research Foundation in 2010 with the amount of funding by the DFG for the same university in the years 2008-2010.

[66] In Figure 5 and 6 only universities with at least 250 and 230 publications in the year 2010 are shown, respectively. However, the coefficient of determination is calculated with all universities that have received DFG funding.

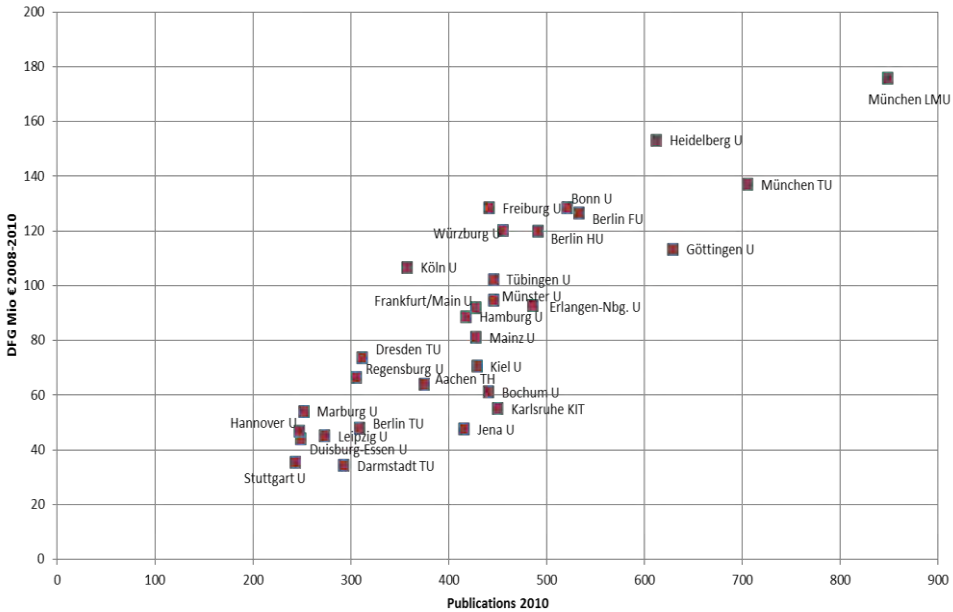

Figure 6. Comparison of all publications in the natural and life science funded by the German Research Foundation in 2010 with the amount of funding by the DFG for the same university in the years 2008-2010.

Discussion

Following the introduction of funding acknowledgements information in the Web of Science database in August 2008, this paper shows the necessary steps needed in order to make this information useful for further study. The growth of publications with funding acknowledgement between the years 2009, 2010 and 2011 shows that 2010 is probably the first year that can be used for further analysis. An analysis of the share of publications with funding acknowledgements in different disciplines shows that in some, like the life sciences the share is that high, that one could assume that most acknowledgements are processed in the database. Although in other disciplines the share is far lower, it is yet unclear whether this is due to less third party funding in these disciplines or due to problems with the extractions of the funding information in certain journals. However, the overall share of 57% for the 2011 shows that this information is usable for a new kind of analysis of the science system. The far more problematic part of this new information is the data quality. In this study we have looked at the German Research Foundation, a particularly large and diverse funding body with many different funding programs. Both on the side of the original funding text in the articles and in their extractions by Thomson Reuters immense problems emerge. Especially, the issue of funding programs being mistaken for funding organizations is particularly pressing and needs of a lot of man-hours in order to

be corrected. Further problems include many variations in translation of the German names of the funding organization and funding programs. In addition to the many orthographic mistakes occurring before and by the data extraction, more severe data extraction errors are apparent. Grant numbers are included in the funding organization field and several funding organization are treated as one combined one on several occasion. In conclusion, a first manual data cleaning step is unavoidable. This array of problems can however be sorted out if enough work is invested. The astonishing result is several thousand entries synonymous with funding given by the DFG[67]. In order to reduce this manual procedure for the subsequent years a new semi-automated method has been employed that uses the regular expression possibilities of the Oracle SQL and a visual basic script implementing a tolerance to typos with a Levenshtein distance algorithm. Using the replicated 2010 results with this method in order to identify new, but similar aliases the 6370 results for the 2010 version of the database could be expanded to include 8739 aliases in the 2011 version. Precision and recall of the method show promising results with 96% and 94%, respectively. In order to exemplify the potential of this cleaned data set two ways to use it in a broader context have been shown. First, with this data the amount of publications in different disciplines funded by the German Research Foundation can be demonstrated. This can be used to assess the disciplines in which the funding body is especially active and in which ones other funding organizations have a higher input. Second, putting the funding acknowledgement data in relation to the funding amounts given by the DFG, as they are included in the DFG Funding Atlas 2012, one can show an input-output relationship in funding. The high correlation between these two data sources shows on one side the validity of the funding acknowledgement information, on the other side opens up possibilities of assessment of funding result not known before. As said, this is only the beginning. The laborious task of data cleaning has now been completed for the German Research Foundation. Once all the major funding organizations are cleaned and unified, a new kind of bibliometric research is possible. Its limits are only set by our own imagination.

Acknowledgments

I would like to thank Simone Falk and Mathias Riechert for their help with the data cleaning and the writing of the visual basic script, respectively.

References

Levenshtein, V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 707–10.

Rigby, John. (2011) Systematic Grant and Funding Body Acknowledgement Data for Publications: New Dimensions and New Controversies for Research Policy and Evaluation. *Research Evaluation* 20, 365 -375.

[67] Grant Lewison, a pioneer in the study of funding acknowledgements was completely incredulous and flabbergasted, confronted with this finding (personal communication at the STI 2012)

Van den Besselaar, P., Inzelt A. & Reale, E. (2012) Measuring Internationalization of Funding Agencies. In: Archambault, Éric / Gingras, Yves / Larivière, Vincent (eds): Proceedings of 17th International Conference on Science and Technology Indicators, Montréal: Science-Metrix and OST, Volume 1, 121-130.