This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field

Fernanda Morillo (1, 3) and Belén Álvarez-Bornstein (1, 2)

(1) IFS, Spanish National Research Council (CSIC), Albasanz, 26-28, 28037 Madrid, Spain

(2) Library and Information Science Department, Faculty of Library and Information Sciences, Complutense University (UCM), Santísima Trinidad 37, 28010 Madrid, Spain

(3) Corresponding author.

E-mail addresses: fernanda.morillo@cchs.csic.es, belen.alvarez@cchs.csic.es

Abstract

In a context of increasingly limited resources, the demand for information from research funding bodies is growing. The exploitation of the funding acknowledgements collected in WoS publications can be useful for these sponsors, not only because it allows them to know the published results with their financial support, but also because it provides a framework to evaluate the efficiency of the different funding instruments. The present work adds to the knowledge of previous studies to offer a simple and efficient methodology that automatically identifies major sponsors, and their funded research, using keywords. To this end, articles with Spain in the address field and English in the language field are obtained (years 2010-2014), given that WoS only considers funding acknowledgements written in English. Subsequently, the *Funding Agency* (FA) field of these articles is treated, selecting funders' variants that will serve as keywords in the FTS (*Full-Text Search*) for the location of the research supported by major sponsors. In addition, a sample of reviewed documents is provided to evaluate the reliability of the proposed methodology, performing also some statistical tests. The results show a recall of 91.5% of the sample articles, with a precision of 99%. Notwithstanding, there are differences in the automatic identification of funders by institutional sector and/or area, being the *Government* sector the one with the highest precision and recall, and the area of *Agriculture, Biology & Environment* the one with the best degree of association between the automatic classification and the reviewed one. Finally, possible future developments are offered, paying special attention to increasing the automation of the standardisation of funders' names.

Keywords

WoS; Funding acknowledgements; Funders identification; Automatic procedures; Performance evaluation; Statistical Analyses

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

Introduction

Scientific and technological progress demands an adequate economic and human investment. However, resources are increasingly limited and thus the funding bodies' requirements are harder, especially those in which there is greater competitiveness (Wang and Shapira [2011](#)). They request to be informed about the scientific careers of applicants and their connections (including international collaboration), but also about the results obtained in previous projects, to be able to assess the feasibility of proposals (Oztaysi et al. [2017](#)). In this context, the proper allocation of research authorship is as important as the correct identification of the type of support received. For these and other reasons, the presence of acknowledgements in publications has been increasing over the years, becoming a further formal element of scientific writing and, in particular, in regard to the recognition of the funds received (Cronin et al. [2003](#); Giles and Councill [2004](#); Costas and Van Leeuwen [2012](#)).

Bibliographic databases such as the *Web of Science* (WoS) offer a reference for these results, because from 2008 onwards they include the funding acknowledgements collected in publications. The exploitation of these data may supplement the information available to funders, because they do not usually have the results produced after the projects fulfilment (Costas and Yegros-Yegros [2013](#)). Moreover, WoS databases provide additional information on the impact of research and a framework for each scientific discipline. The interest of funding bodies in these data is clear. As pointed by Grassano et al. ([2017](#)), some of the largest international sponsors, such as the *US National Institutes of Health*, have created their own databases to link scientific results with the given grants. Furthermore, as an initiative of the *Medical Research Council* (UK), *ResearchFish* was launched in 2012, which is a platform for researchers from around the world to be able to give an account of their scientific outputs to their funding agencies. Within this frame, it is possible to make comparisons, even at the international level, with results of projects supported by other sources of funding. All this reinforces the idea that the research system depends in large part on the ability of applicants to meet the funders expectations to achieve their goals. For this reason, agencies must also take into account the existing capabilities before promoting new funding instruments (Lepori [2011](#)).

With respect to the funding acknowledgements collected by WoS, previous studies have analysed their advantages and limitations. On the one hand, Wang and Shapira ([2011](#)) analyse a small sample of documents, stating that the acknowledgements information is erroneously transferred in a low percentage of publications (<1%). This corresponds to the *Funding Text* (FT) field, which in turn is subdivided into two fields, *Funding Agency* (FA) and *Grant Number* (GN). On the other hand, other authors examine the trustworthiness of the FA and GN fields,

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

since being easier to treat, are the main source of information used in research funding studies. In this sense, Rigby (2011) and Álvarez-Bornstein et al. (2017) advise not to carry out a direct count of sponsors and grants from these fields. The latter authors also conclude that not all data are correctly collected, although the reliability of the FA field (only 1% of publications do not contain information) is greater than that of the GN field (44% of variants are blank).

Additionally, funding acknowledgements count does not provide a complete picture of research, because monetary support has other results than document production and it is difficult to measure the influence of a given funding on a research result. Besides, Rigby (2011) believes that the same problems encountered with citations could be found with acknowledgements, because some authors may exaggerate the productivity of certain grants in order to enhance their reputation, especially from *high-status* funders. At the same time, authors could hide information from *lower-status* funding sources. Moreover, it is not easy to know the exact contribution of each sponsor reading the FA field, because, if collected, it only appears in the full section of acknowledgements. Nevertheless, more and more funding bodies include requirements for thanking their support in published documents, while citations imply a certain *freedom* when it comes to recognising the sources of information used. However, both citations and acknowledgements involve some interaction and influence (Díaz-Faes and Bordons 2014). In fact, Costas and van Leeuwen (2012) state that funding acknowledgements are an indicator of quality similar to that of citations, given that they contribute to increasing research impact. Furthermore, they think that acknowledgements can improve the understanding of the communicative process through network analysis of scientific interactions.

Challenges in the treatment of funding data

Apart from the above mentioned, the treatment of the FA field is complex and difficult to accomplish. On the one hand, it contains heterogeneous information on funders and/or specific programmes, projects or grants. On the other hand, multiple entities and/or multiple grants may appear in a single article. Indeed, WoS databases do not include standard names or codes for funding organisations, whereas Scopus does it with the *Crossref Funder Registry* (<https://www.crossref.org/services/funder-registry/>). However, Scopus extracts funding information from acknowledgements only when sponsors are already included in the *Crossref* taxonomy, therefore only their full text field includes all data. With respect to WoS, since there is no standardisation, it is possible to find a single funding agency under different names (e.g. *US National Science Foundation* or *NSF*), while a name may refer to several entities (e.g. *National Science Foundation of Switzerland* or *Chinese National Science Foundation*). The identification of one of these funders can be hampered if no geographic origin is included (according to the work by Álvarez-Bornstein et al. 2017, this happens in about 30% of the Spanish funding bodies).

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

Another issue in counting results from a particular sponsor is the change of name over time (Rigby [2011](#)). This is particularly complicated with the ministerial bodies, which show remarkable variations, as is the case of the Spanish ministry responsible for science management (with seven different names in the last 20 years). In fact, in a single FA record, the same agency can be found twice with two different names. For example, *Spanish Ministerio de Educacion y Ciencia y Ministerio de Ciencia e Innovacion* (sic) includes two funding bodies, although both were in charge of science management in Spain (the first from 2004 to 2008 and the second from 2008 to 2011). It is also common to find *creative* translations in English from agencies originally in other languages (Sirtes [2013](#)): *Spanish Scientific Research Council (CSIC)*, instead of the *Spanish National Research Council (CSIC)*, or *Spanish Ministry of Economy and Sustainability* instead of the *Spanish Ministry of Economy and Competitiveness* (MINECO). Moreover, sometimes the names of projects or programmes are mentioned rather than their funders (e.g. *DEMETER*, a research project, or *Juan de la Cierva*, a postdoctoral fellowship, both included in the *National Plan for Scientific and Technological Research and Innovation* from the Spanish Government). With this practice, taking into account that project information is not always easily found, locating funding sources is even more complex.

With regard to the full text of acknowledgements, there are several works with proposals for the treatment of this information. For instance, Giles and Council (2004) apply automatic methods to extract information from the acknowledgements included in documents of the *CiteSeer* digital library, differentiating between cases with clearly identified sections and those that do not, and dealing with all types of entities (both sponsors and individual researchers). In addition, they use regular expressions to extract the entities' names, merging the variants found (e.g. *National Science Foundation* and NSF). Although these authors obtain good results in terms of precision and recall, they also recognise the complexity of the task of entity name disambiguation. On the other hand, Grassano et al. (2017) analyse a sample of 7,510 UK cancer research publications in 2011 indexed in MEDLINE/PubMed. They extract the funding acknowledgements information from the full text of documents when authors explicitly declared it. Nevertheless, they also point out that this is an intricate task due to the ambiguous meaning of the term *support*, which can refer not only to financial aid but also to material or personal help (also stated by Cronin et al. 2003, Costas and van Leeuwen 2012, Díaz-Faes and Bordons 2014 or Álvarez-Bornstein et al. 2017). Indeed, some authors believe that personal acknowledgements may include some kind of *subauthorship*, offering additional information on collaboration (e.g. Cronin et al. 2003). The detection and extraction of funding data is not a simple job, given that in most cases it has to be searched in a general acknowledgements section and occasionally even in footnotes or at the beginning of papers (as in Giles and Council 2004). Nevertheless, the full text treatment of acknowledgements has the advantage to better distinguish potential conflicts of interest

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

(Lewison and Sullivan 2015). In this sense, Díaz-Faes and Bordons (2014) analyse the full texts of acknowledgements in a sample of documents and group them according to three typologies: *technical support*, *peer interactive communication* and *conflicts of interest*. The last case is mainly identified by terms such as *honorary*, *fees* and *consultants*, which greatly facilitates its location. However, in some cases, WoS considers some organisations as sponsors despite having been clearly defined as potential conflicts of interest (Grassano et al. 2017). In fact, these authors believe that their work provide a good basis for future studies that can develop natural language algorithms capable of automatically detecting and extracting document funding data.

For the supervised and/or semi-supervised identification of organisational names, a number of methodologies have been proposed, focusing chiefly on the unification of the authors' addresses (see for example Cuxac et al. 2013, Morillo et al. 2013a and 2013b, Huang et al. 2014, or Caron and Daniels 2016). All these proposals use data (the authors' addresses) that offer sufficient information, something that does not happen with funding acknowledgements, because sometimes they only provide acronyms or even grants' reference numbers. For funders' detection, apart from the one described by Giles and Council 2004, there is a more recent proposal by Wang and Shapira (2011), who use *VantagePoint* to clean and organise data from their set of documents extracted from WoS. This software applies approximate string matching techniques to this task, although, as these authors reveal, a considerable amount of validation and data cleaning work is necessary to carry out an accurate analysis. In addition, this study is based on the results of *Nanotechnology* and does not detail the methodology used to standardise funding agencies. In a similar way, Gök et al. (2016) use this text mining software to automate the process of extracting sponsors from the full text of acknowledgements, examining the scientific production of six European countries. Furthermore, they clean and merge the different denominations of funders included in their sample, and emphasise that various iterations and revisions are required to minimize ambiguity. For his part, Sirtes (2013) studies funding data from the *German Research Foundation* (DFG) and, after analysing the text to extract possible variants of the year 2010, proposes a new semi-automatic cleaning method. This method is applied to 2011 data, using regular expressions, the Levenshtein's distance algorithm that sorts possible misspellings, and lists of false positive to exclude from the results. Subsequently, Sirtes and Riechert (2014) present a new proposal for the unification of funding bodies, creating a thesaurus of variants from an in-house database. With this methodology, they reduce the manual data processing, as they also take advantage of the acronyms included in funding acknowledgements (combining them with full text and controlling the presence of homonyms).

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

Objectives and organisation

The present work adds to the knowledge of previous studies to offer a simple and efficient methodology that allows to find scientific output funded by major sponsors and collected in WoS databases. The main advantage of this procedure is that it can be applied to any data set, since it automatically selects a list of funders' variants as keywords that best identify the main sponsors of that set. Thanks to this list, it is possible to detect and recover the publications supported by these funding organisations. Two objectives are considered:

- To study the reliability of the automatic technique with regard to the recognition of funding sources
- To assess whether there are differences between institutional sectors and/or thematic areas, comparing a reviewed sample with the automatic retrieval of sponsors

The organisation of this paper is as follows: firstly, a section of [Materials and methodology](#) is presented, detailing the applied technique; secondly, a [Results and evaluation](#) section is offered, which analyses a reviewed sample of documents and provides statistical tests to evaluate the automatic detection of funders. Finally, a section of [Discussion and conclusions](#) is shown, which compares the results obtained with those of previous studies and focuses on possible future developments.

Materials and methodology

This paper benefits from the WoS *Funding Agency* (FA) field, collected for documents with acknowledgements written in English, to analyse the sponsors included therein. To that end, articles with *Spain* in the *address* field and *English* in the *language* field are retrieved (245,088 items in years 2010-2014). From these, the documents with data in the FA field are analysed (181,610), extracting and identifying the main funding bodies. Besides, in order to facilitate the analysis of sponsors, a master table is created, which includes different variants of geographical names (cities, countries and regions), their gentilics, and their corresponding ISO codes.

The automatic procedure follows the steps described below:

1. Firstly, the FA field is grouped and processed, removing non-alphanumeric characters and stop words (as in Morillo et al. [2013b](#)) and grouping again all possible variants to get a new field. Example:

→ *Spanish Ministry of Economy and Competitiveness (MINECO)* is transformed into *Spanish Ministry Economy Competitiveness MINECO*.

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

2. Afterwards, the new field is analysed, creating another field (FA_clean) from which all the possible geographical names are extracted while adding the corresponding ISO codes, based on the information provided by the master table previously created (51% of the variants include geographical names collected in the master table). Example:
→ *Spanish Ministry Economy Competitiveness MINECO* is transformed into *Ministry Economy Competitiveness MINECO* with the ISO code *ES*.
3. Once the FA_clean field is created, a word count is performed, concluding that this field varies from 1 to 26 words, although shorter variants (1 to 6 words) can be found in 98% of the articles. Therefore, these variants are taken as a starting point for the creation of keywords, calculating the frequency of appearance of each set of words to establish the selection threshold ($\geq 0.05\%$). That is, if there are 75,055 articles with variants of 2 words, each variant that appears in a minimum of 38 articles will be automatically taken as a keyword (and so on for each set of words). With this procedure, a list of 959 keywords are selected, excluding those that do not clearly refer to a single funding body.
4. Subsequently, each keyword is assigned a sponsor's name, creating a list of 573 standardised funders. Although manual revision is necessary, some of these tasks are performed automatically, grouping variants with a high level of similarity and taking advantage of acronyms (as in Giles and Council 2004 or in Sirtes and Riechert 2014). With regard to the latter case, for the creation of standardised sponsors, only those acronyms that match letter by letter with their full text are used (e.g. *ERDF* is allocated to the *European Regional Development Fund*).

The list of standardised funders includes organisations from different institutional sectors following the *Frascati Manual* (OECD 2015) in order to allow international comparisons. Four sectors are considered: *Business enterprise*, *Government*, *Higher education*, and *Private non-profit*. Each funding body is placed only in one sector, following the decision tree of the manual. Therefore, market producers are considered in the *Business enterprise* sector, as well as those that serve them (e.g. business associations), but not if they fit in the *Higher education* sector. Funders that provide formal tertiary education services are included in the latter sector (also those controlled by them), while the ones administered by governments are placed in the *Government* sector (except when already inserted in the previous sectors). Finally, the *Private non-profit* sector comprises all non-profit institutions serving society, unless they are considered in the other sectors.

In addition, in the case of the *Government* sector, it is necessary to take special care with the presence of funding instruments or programmes instead of their corresponding sponsors, the hierarchical structure of organisations and the name variants of entities. To solve some

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

of these problems, authors such as Wang and Shapira (2011) chose to group European programmes and funds (e.g. ERDF) under the name *European Union*. However, this decision does not ease the work of locating the different European instruments and bodies. For this reason, in the present study, this effort is used to establish hierarchical relationships between the different funding sources and provide detailed data at different levels, which also avoids duplication. This task has been possible thanks to the use of supplementary data from the analysed documents and from external sources such as the Internet. Example:

→ The *Spanish Ministry of Economy and Competitiveness (MINECO)* is under the *Spanish Government*, but above the *Spanish National Research Council (CSIC)*, or above the *Spanish National Plan for Scientific and Technical Research and Innovation*.

5. Finally, the keywords list is used to automatically locate funders in the FA_clean field, assigning the research results to their standardised sponsors. Normally, when trying to find the different variants of a funding body, it is necessary to consider the possible variants of its name and, at the same time, those of its place of origin. However, in this work, the methodology is simplified by processing the original FA field, assigning an ISO code for each geographical origin and looking for variants of the funder's name only in those records with the same ISO code as in the keywords list. Likewise, when the funder's name is meaningful enough, the search is done in all records without taking into account any ISO code (this happens with companies and with almost half of the acronyms).

The procedure executes a FTS (*Full-Text Search*), which finds all words regardless of their order, excluding terms with a single character, and establishing a match threshold of 50% (i.e. at least half of the keyword's terms are located in a FA variant). Therefore, for the shortest keywords (1-2 words) all the terms that compose them must be found. In very doubtful cases, a PS (*Phrase Search*) is performed, using the original FA field, and retrieving the words in a certain order. This search is necessary mainly with keywords that refer to funding bodies with very similar names, as happens with universities (e.g. *University of Washington* and *Washington State University*).

Once the process described is completed, the data are grouped by sponsor, which allows obtaining the number of publications funded by each of them. Example:

→ The FA variant *Spanish Ministry of Economy and Competitiveness (MINECO)* is found in 446 articles. Additionally, it can be allocated to the *Ministry of Economy and Competitiveness (MINECO)*, *Spain*, a standardised funder that, in turn, can be found in 81,323 documents.

Taking into account the original FA variants (160,677), 50% of them are automatically detected, with an additional 19% of which only their geographical origin can be identified. This 50% of variants is found in 167,259 articles (92% of the total articles with FA), reaching 174,612

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

articles (96%) if those variants with geographical origin are also considered. Therefore, despite the limited proportion of FA variants automatically recognised, as they correspond to large sponsors, the proportion of research results funded by them is much higher. Nonetheless, these data do not offer information on the precision of the automatic approach. For this reason, in order to evaluate the trustworthiness of the applied methodology, a sample of articles already reviewed is used. As described in Álvarez-Bornstein et al. (2017), this sample is obtained through a stratified random sampling, getting a total of 1,045 articles (year 2014) from eight areas: *Agriculture, Biology & Environment; Biomedicine; Chemistry; Clinical Medicine; Engineering & Technology; Mathematics; Multidisciplinary; and Physics*. Thanks to these data, the results of the manually reviewed funders can be compared with those automatically provided, as well as obtaining precision and recall indicators by institutional sector (as in Morillo et al. 2013a).

Moreover, several statistical tests are executed. On the one hand, non-parametric correlations are applied between the automatic list of sponsors and the manually reviewed one, including a comparison by thematic area. On the other hand, Cross-tabulation procedures are performed to assess whether there are differences between institutional sectors and/or thematic areas, considering the reviewed versus automatic retrieval of funding bodies. In addition, a Chi-square test is used to contrast the null hypothesis of independence between variables, applying also some symmetric measures, such as Cramer's V, to know the strength of the association. Finally, several directional measures are also employed, such as Goodman and Kruskal tau, which calculates the proportional reduction of the error when the values of the automatic classification are used to predict the values of the reviewed classification.

Results and evaluation

The selected sample (1,045 articles), previously reviewed, and its corresponding automatic identification of funders were analysed in order to assess the validity of the proposed methodology. As stated by Álvarez-Bornstein et al. (2017), there is some missing information in the FA field (1% of items in the sample). Therefore, funding bodies were found only in 1,032 articles, which included 1,958 FA variants that could be analysed. The automatic procedure allowed the recognition of sponsors in 944 articles of this sample (recall of 91.5%), with a precision of 99% (only 9 items were misclassified). Furthermore, 1,232 FA variants were automatically detected (62.9% of the total variants), with a precision of 97.2%, although there was an additional 2.1% partially identified. That is, there were variants that contained several funders and only some of them were located with the automatic technique.

Notwithstanding, it was possible to retrieve an important proportion of the research output funded by major sponsors with this automatic technique. [Table 1](#) shows the percentage of

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

funded research, and the share of it that can be detected automatically, for each reviewed sponsor with a participation in at least 1% of the articles in the sample. It should be noted that, in order to offer a certain homogeneity in results, a single hierarchical level of sponsors with their name in English is presented, assigning the most recent and officially recognised form (at the time of the study), whenever possible. As it can be seen, the largest contribution comes from the Spanish ministry that manages scientific research (MINECO), with a presence in the acknowledgements of 61% of the sample articles. The second position is held by the *European Commission* (EC), with 29% of the items (half of those of MINECO), and third is the *Spanish Ministry of Education* (MECD), with a still much lower participation, barely exceeding 5% of articles. Taking into account the part that is automatically detected, the share is very high for the first two funders, given that at least 86% of their sponsored output can be found. In fact, on average, 70% of the research funded by the organisations shown in the table could be detected automatically.

Table 1. Major reviewed funders with their percentages of research supported ($\geq 1\%$) and the share of them that can be automatically detected

Major funders	%	Share
Ministry of Economy and Competitiveness (MINECO), Spain	61.3%	87.8%
European Commission (EC), European Union	29.0%	86.0%
Ministry of Education, Culture and Sport (MECD), Spain	5.2%	42.6%
Regional Ministry of Business and Knowledge, Regional Government of Catalonia, Spain	4.1%	66.7%
Ministry of Science, Technology and Higher Education (MCTES), Portugal	3.3%	79.4%
Ministry of Agriculture, Food and Environment (MAGRAMA), Spain	2.9%	33.3%
Regional Ministry of Education, Research, Culture and Sport, Regional Government of Valencia, Spain	2.6%	59.3%
Regional Ministry of Economy and Knowledge, Regional Government of Andalusia, Spain	2.5%	65.4%
Interministerial Commission for Science and Technology (CICYT), Spain	2.4%	96.0%
Regional Ministry of Education, Youth and Sport, Regional Government of Madrid, Spain	2.4%	52.0%
Ministry of Health, Social Services and Equality (MSSSI), Spain	2.2%	100.0%
Regional Ministry of Employment, University and Business, Regional Government of Murcia, Spain	1.7%	77.8%
Ministry of Foreign Affairs and Cooperation (MAEC), Spain	1.6%	94.1%
United Nations (UN), United States of America	1.6%	82.4%
Ministry of Higher Education, Research and Innovation (MESRI), France	1.6%	68.8%
Regional Ministry of Education, Regional Government of the Basque Country, Spain	1.5%	100.0%
Regional Ministry of Economy, Employment and Industry, Regional Government of Galicia, Spain	1.4%	50.0%
Ministry of Science, Technology and Productive Innovation, Argentina	1.3%	100.0%
Regional Ministry of Education, Regional Government of Andalusia, Spain	1.2%	66.7%
Research Councils UK (RCUK), United Kingdom	1.2%	66.7%
Regional Ministry of Culture, Education and University Planning, Regional Government of Galicia, Spain	1.1%	45.5%
Secretariat of Public Education (SEP), Mexico	1.1%	45.5%
Ministry of Education, University and Research (MIUR), Italy	1.0%	50.0%

On the other hand, in order to assess the reliability of the automatic approach, the number of articles was distributed by standardised funding body and compared with the reviewed

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

distribution. The result of the statistical test shows a high Spearman's correlation coefficient (0.911) with significance level 0.01 (**Table 2**). Moreover, when comparing the reviewed with the automatic allocation of articles, by sponsor and thematic area, significant correlations were also observed in all cases at the 0.01 level. However, there were important variations in the Spearman's correlation coefficient, especially high for *Mathematics* (0.949) and moderate in the case of the *Multidisciplinary* area (0.774).

Table 2. Spearman's rho correlation between articles by reviewed funder (ArtR) and articles by automatically identified funder (ArtA)

		ArtR	ArtA
ArtR	Correlation Coefficient	1.000	0.911**
	Sig. (2-tailed)		0.000
ArtA	Correlation Coefficient	0.911**	1.000
	Sig. (2-tailed)	0.000	

**. Correlation is significant at the 0.01 level (2-tailed).

In addition, to evaluate the levels of recall and precision of the automatic method in the identification of funding sectors, the sample was grouped and analysed by institutional sector. Accordingly, the Cross-tabulation procedure was used with the percentages of articles of each reviewed sector and those of each automatically identified sector. As shown in **Table 3**, there is a precision of 100% in the automatic classification of all sectors (% within the SectorA), with the exception of the *Government* sector (although it has a precision of almost 100%). Conversely, the distribution of recall by sectors is much more irregular, reaching 91.1% for the *Government* sector and only 23.8% for the *Business enterprise* sector (% within SectorR). In any case, as it can also be seen in the table, the *Government* sector is the one with the greatest presence (almost 98% of the reviewed articles), while the *Business enterprise* sector does not even fund 8% of the research results, being the *Private non-profit* sector the one with the lowest participation (6.4%).

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

Table 3. Cross-tabulation between percentages of articles by reviewed institutional sector (SectorR) and by automatically identified institutional sector (SectorA), and total percentages of articles by SectorR

SectorR	SectorA				% SectorR
	Business enterprise	Government	Higher education	Private non-profit	
Business enterprise					7.8%
% within SectorR	23.8%	0.0%	0.0%	0.0%	
% within SectorA	100.0%	0.0%	0.0%	0.0%	
Government					97.9%
% within SectorR	0.0%	91.1%	0.0%	0.0%	
% within SectorA	0.0%	99.8%	0.0%	0.0%	
Higher education					14.4%
% within SectorR	0.0%	1.3%	57.7%	0.0%	
% within SectorA	0.0%	0.2%	100.0%	0.0%	
Private non-profit					6.4%
% within SectorR	0.0%	0.0%	0.0%	31.8%	
% within SectorA	0.0%	0.0%	0.0%	100.0%	

% within SectorR = Recall. % within SectorA = Precision.

Furthermore, several tests were used to check if both variables were associated and to know the strength of this relation, as well as the ability of the automatic sector variable to predict the reviewed sector variable. On the one hand, the Chi-square test shows that there is a significant linear association between both variables (**Table 4**). Additionally, there is a good degree of association calculated through various symmetric measures. For instance, the value of Cramer's V is 0.700 and the one of the Contingency Coefficient is 0.772, although being less limited than Cramer's V, its interpretation is more complex (**Table 5**). On the other hand, some directional measures are also offered, such as Goodman and Kruskal tau, which presents a reduction of 0.613 in the prediction error of the reviewed classification using the automatic classification, while the Uncertainty Coefficient is 0.638 (**Table 6**).

Table 4. Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	70535.537	12	0.000
Likelihood Ratio	47038.982	12	0.000

Table 5. Symmetric Measures

Nominal by Nominal	Value	Approximate Significance
Cramer's V	0.700	0.000
Contingency Coefficient	0.772	0.000

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

Table 6. Directional Measures

Nominal by Nominal	Value	Asymptotic Standard Error ^a	Approximate T ^b	Approximate Significance
Goodman and Kruskal tau				
SectorR Dependent	0.613	0.001		0.000 ^c
SectorA Dependent	0.504	0.003		0.000 ^c
Uncertainty Coefficient				
Symmetric	0.598	0.003	138.235	0.000 ^d
SectorR Dependent	0.638	0.003	138.235	0.000 ^d
SectorA Dependent	0.562	0.003	138.235	0.000 ^d

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on chi-square approximation.

d. Likelihood ratio chi-square probability.

Finally, to assess the trustworthiness of the automatic technique by thematic area and funding sector, several sets were created and Cross-tabulation procedures were performed. Once again, the results showed a precision of 100% in the automatic classification in all areas and sectors except for the *Government* sector of *Biomedicine*, although its precision was remarkable (99.3%), and *Mathematics*, with a precision of 97.6%. Conversely, as displayed in [Figure 1](#), recall is also very irregular, being better for the *Government* sector in all areas and especially in *Biomedicine* (91.8%), and worse for the sectors *Business enterprise* and *Private non-profit*, of which no article is retrieved in *Multidisciplinary*. In any case, these data refer again to institutional sectors with little presence in the funding acknowledgements and an area with the lowest scientific production in the analysed sample. With respect to the statistical analyses, Chi-square tests showed a significant linear association between both variables for all areas. Nevertheless, the degree of association varied, being the highest for *Agriculture, Biology & Environment* (with a Cramer's V of 0.775 and a Contingency Coefficient of 0.802), and the lowest for *Mathematics* (with a Cramer's V of 0.586) and *Multidisciplinary* (with a Contingency Coefficient of 0.668). Moreover, the directional measures also offered heterogeneous results, with the highest values of the Goodman and Kruskal tau and Uncertainty Coefficient for *Agriculture, Biology & Environment* (0.712 and 0.729 respectively) and the lowest for *Mathematics* (0.472 and 0.492 respectively), being this latter the second area with the lowest output in the analysed sample.

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

Figure 1. Measures of recall by institutional sector and area

Discussion and conclusions

This article offers a simple and efficient methodology to automatically identify major sponsors, and their funded research, through keywords. As pointed out before, the main advantage of this technique is that it can be applied to any set of documents, because the selection of funders' variants that will serve as keywords is automatic and based upon the chosen set, being also automatic the location of funded documents. Regarding the creation of a standardised list of sponsors, some manual intervention is required, which allows greater control of the terms selected and a greater precision to recover the supported research by each funder. Nevertheless, it is believed that accuracy is also ensured by the selection of short and full FA variants. These variants (from 1 to 6 words) may not only better identify the main sponsors, but also, since they are the most frequent, they can find a greater number of results funded by those sponsors. Notwithstanding, this work also provides a sample of reviewed documents to assess the reliability of the proposed methodology. The funding sources of these documents are detected and classified with standard names, and the reviewed classification is compared with the automatic one. Moreover, statistical tests are carried out by institutional sector and/or thematic area, obtaining remarkable results for all cases. The results of the evaluation show a recall of 91.5% of the sample articles, with a precision of 99%. If the FA variants are taken into account, there is a recall of 62.9% with a precision of 97.2%. These percentages are similar to those found by Sirtes and Riechert (2014) for the *German Research Foundation* (DFG) and, as these authors point out, the large differences in recall between the total number of documents and the FA

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

variants are explained because the unclassified variants are located in a smaller number of articles. Besides, in the present work, the precision of 97.2%, of the FA variants, increases if an additional 2.1% corresponding to those partially detected are considered. The reason for this incomplete identification is due to the study's restrictions, given that those cases do not reach the thresholds established for the creation of keywords (such as companies, with a lower contribution, or unusual/unclear variants of funders' names).

Concerning the results obtained, [Table 1](#) shows that funding bodies with the greatest presence in acknowledgements come from the *Government* sector, which is not strange bearing in mind that this sector appears in 97.9% of the sample articles. As explained above, this table offers only sponsors of the same hierarchical level (to avoid duplications), including contributions from the agencies that depend on them and excluding those ones from a higher hierarchical level. Therefore, although part of the production funded by some organisations cannot be located automatically (e.g. only 42.6% of the MECD items have been found), it should be borne in mind that in these cases the assignment is made to a higher hierarchical level (i.e. the Spanish Government), so it is less precise but not less correct. Similarly, the results funded by MINECO include the ones of its subordinated agencies. For instance, 9.6% of them have acknowledgements to the *Spanish National Institute of Agricultural and Food Research and Technology* (INIA), and 7.7%, to the *Spanish National Research Council* (CSIC). Regarding the *European Commission* (EC), most of its funding (55.2%) is in fact acknowledged to the *European Regional Development Fund* (ERDF). Nonetheless, it is evident that, in this sample of documents with Spanish addresses, national funding prevails, as happens with other data sets from other countries (e.g. Gök et al. [2016](#)).

In relation to the hierarchical structure described in this study, four different hierarchical levels have been established: governments, their corresponding ministerial departments, their main bodies and the funding instruments used by them. Although the need to establish standardised hierarchical levels depends largely on the type of the analysis carried out, it is desirable, however, to have a minimum hierarchical structure that considers at least the first three levels mentioned. Thus, if the study is based on general international comparisons, a higher hierarchical level will suffice. On the contrary, it is necessary to create an accurate hierarchical classification in order to be able to compare the results of scientific research of certain funding instruments. For instance, it may be interesting to contrast the results funded by the ERDF with those funded by the Framework Programme of the European Union (EU). This study might present differences between both sets, since there are harder requirements to get funding from the Framework Program than those needed for the ERDF. This is because the latter fund is

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

managed jointly between the EU and the Member States and is requested through national or regional calls for proposals.

With respect to the first objective considered, which is studying the reliability of the automatic detection of sponsors, it was observed a high and significant non-parametric correlation with their manually reviewed identification. However, some differences were found by thematic area, presenting *Mathematics* the greatest association between the automatic distribution of articles by funder and the one manually reviewed, while *Multidisciplinary* had the lowest correlation. These results can be explained by the average number of funders, being the lowest for *Mathematics* and the highest for *Multidisciplinary* (as described in Álvarez-Bornstein et al. [2017](#) for the FA-GN entries). For the second objective proposed, the evaluation of possible differences in the automatic identification of funders by institutional sector and/or thematic area, various Cross-tabulation procedures were executed. Therefore, articles with funding acknowledgements were grouped according to their institutional sectors, comparing those manually reviewed with those automatically assigned. As it can be seen in the [Results and evaluation](#) section, the precision of the automatic classification is 100%, with the exception of the *Government* sector, although it is almost 100%. When analysing the data by thematic area, it was observed that this classification error occurred in *Biomedicine* and *Mathematics*. However, this is only due to a couple of publications that include acronyms with homonymy and that should have been assigned to the *Higher education* sector. As for the recall obtained, the results were less promising, since only 24% of the *Business enterprise* sector and 32% of the *Private non-profit* sector are found. However, it should be taken into account that each of these sectors did not even represent 8% of the sample documents. Besides, when studying thematic areas, *Agriculture, Biology & Environment* showed the highest proportion of publications automatically identified for the *Business enterprise* sector (36%), while *Biomedicine* had the highest percentage of articles for the *Private non-profit sector* (56%). Both in the whole data set and in the subsets of thematic areas, the highest representation was observed for the *Government* sector, which was also the one most exhaustively collected. Consequently, a good degree of association was observed between the reviewed and the automatic classification (e.g. Contingency Coefficient of 0.772), being worse for thematic areas with lower presence (i.e. *Mathematics* and *Multidisciplinary*).

Although this methodology makes easier the process of gathering data from research output funded by a particular sponsor, not all contributions are known and not all have the same weight in a specific scientific result. In many cases, internal (implicit) funds provide certain results or have a partial influence on them. In other cases, part of the support received is not recognised, mainly due to author's negligence or funders' security reasons (as stated by Rigby [2011](#) or Wang

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

and Shapira [2011](#)). On the other hand, in this study, it was not possible to treat some data without enough information for the correct identification of a particular funding organisation. In any case, according to Rigby ([2011](#)), there is no direct input-output relationship between resources and publications, considering that the funded research has also other outcomes, although this does not mean that sponsors do not try to determine which financial instruments are the most effective. In this sense, the present work gives valuable information for funders, since the analysis of specific agencies and instruments is carried out. In spite of this, as stated by Álvarez-Bornstein et al. ([2017](#)) or Grassano et al. ([2017](#)), it is necessary to offer clearer norms for the inclusion of funding acknowledgements in publications, which will facilitate studies like this one that try to identify what funded results are obtained by each sponsor.

Given that there is still a great heterogeneity in the way in which the economic support is acknowledged, it is necessary to dedicate time and effort to the unification and standardisation of the different funding sources. The present work minimises this effort by removing non-alphanumeric characters and stop words, extracting all possible geographical origins of the FA variants and automatically selecting appropriate keywords that identify major sponsors. This system of previous extraction of geographical origins is possibly the main contribution of this study. It enables the detection of each funder's variants of a given place and those of a funder meaningful enough to not require any physical location. However, the automatic methodology focuses mainly on the location of funded research, while offering only a certain degree of automation for the assignment of standardised names to the different sponsors. For this reason, future research will pay special attention to increasing the automation of this assignment, based on the previous experiences referred to in various publications (e.g. Cuxac et al [2013](#) or Caron and Daniels [2016](#)). These works use clustering methods when there is no previous information about the analysed data, although in the case of funders, their particular characteristics must be taken into account. Finally, further efforts will be devoted to improve the detection of the funded research by each sponsor, with the aid of data already reviewed, which will undoubtedly serve to increase the efficiency of the whole procedure.

Acknowledgements

This work is supported by the *Spanish Ministry of Economy and Competitiveness* (Grant CS02014-57826-P and predoctoral contract BES-2015-073537).

References

Álvarez-Bornstein, B., Morillo, F.; & Bordons, M. (2017). Funding acknowledgements in the Web of Science: completeness and accuracy of collected data. *Scientometrics*, 112(3):1793–1812. [doi:10.1007/s11192-017-2453-4](https://doi.org/10.1007/s11192-017-2453-4)

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

Caron, E.A.M., & Daniels, H.A.M. (2016). Identification of Organization Name Variants in Large Databases using Rule-based Scoring and Clustering - With a Case Study on the Web of Science Database. In *Proceedings of the 18th International Conference on Enterprise Information Systems* (Vol. 1, pp. 182-187). [doi:10.5220/0005836701820187](https://doi.org/10.5220/0005836701820187)

Costas, R., & van Leeuwen, T. N. (2012). Approaching the “reward triangle”: General analysis of the presence of funding acknowledgments and “peer interactive communication” in scientific publications. *Journal of the American Society for Information Science and Technology*, 63(8), 1647-1661. [doi:10.1002/asi.22692](https://doi.org/10.1002/asi.22692)

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 the 14th International Conference of the International Society for Scientometrics and Informetrics* (Vol. 2, pp. 1401-1408). Retrieved from http://www.iss2013.org/Images/ISSI_Proceedings_Volume_II.pdf. Accessed April 21, 2017.

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for information Science and Technology*, 54(9), 855-871. [doi:10.1002/asi.10278](https://doi.org/10.1002/asi.10278)

Cuxac, P., Lamirel, J. C., & Bonvallot, V. (2013). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, 97(1), 47-58. [doi:10.1007/s11192-013-1025-5](https://doi.org/10.1007/s11192-013-1025-5)

Díaz-Faes, A. A., & Bordons, M. (2014). Acknowledgements in scientific publications: presence in Spanish science and text patterns across disciplines. *Journal of the Association for Information Science and Technology*, 65(9), 1834–1849. [doi:10.1002/asi.23081](https://doi.org/10.1002/asi.23081)

Giles, C. L., & Councill, I. G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51), 17599-17604. [doi:10.1073/pnas.0407743101](https://doi.org/10.1073/pnas.0407743101)

Gök, A., Rigby, J., & Shapira, P. (2016). The impact of research funding on scientific outputs: Evidence from six smaller European countries. *Journal of the Association for Information Science and Technology*, 67(3), 715-730. [doi:10.1002/asi.23406](https://doi.org/10.1002/asi.23406)

Grassano, N., Rotolo, D., Hutton, J., Lang, F., & Hopkins, M. M. (2017). Funding data from publication acknowledgements: Coverage, uses, and limitations. *Journal of the Association for Information Science and Technology*, 68(4), 999-1017. [doi:10.1002/asi.23737](https://doi.org/10.1002/asi.23737)

Huang, S., Yang, B., Yan, S., & Rousseau, R. (2014). Institution name disambiguation for research assessment. *Scientometrics*, 99(3), 823-838. [doi:10.1007/s11192-013-1214-2](https://doi.org/10.1007/s11192-013-1214-2)

Lepori, B. (2011). Coordination modes in public funding systems. *Research Policy*, 40(3), 355-367. [doi:10.1016/j.respol.2010.10.016](https://doi.org/10.1016/j.respol.2010.10.016)

Lewison, G., & Sullivan, R. (2015). Conflicts of interest statements on biomedical papers. *Scientometrics*, 102(3), 2151–2159. [doi:10.1007/s11192-014-1507-0](https://doi.org/10.1007/s11192-014-1507-0).

Morillo, F., Aparicio, J., González Albo, B., & Moreno, L. (2013a). Towards the automation of address identification. *Scientometrics*, 94(1), 207-224. [doi:10.1007/s11192-012-0733-6](https://doi.org/10.1007/s11192-012-0733-6)

Morillo, F., Santabárbara, I., & Aparicio, J. (2013b). The automatic normalisation challenge: Detailed addresses identification. *Scientometrics*, 95(3), 953–966. [doi:10.1007/s11192-013-0965-0](https://doi.org/10.1007/s11192-013-0965-0)

This is a **postprint version** of:

Morillo, F., & Álvarez-Bornstein, B. (2018). How to automatically identify major research sponsors selecting keywords from the WoS Funding Agency field. *Scientometrics*, 117(3):1755–1770.

The final publication is available at Springer via <https://doi.org/10.1007/s11192-018-2947-8>

OECD (2015). *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*. The Measurement of Scientific, Technological and Innovation Activities. OECD Publishing, Paris. [doi:10.1787/9789264239012-en](https://doi.org/10.1787/9789264239012-en)

Oztaysi, B., Onar, S. C., Goztepe, K., & Kahraman, C. (2017). Evaluation of research proposals for grant funding using interval-valued intuitionistic fuzzy sets. *Soft Computing*, 21(5), 1203-1218. [doi:10.1007/s00500-015-1853-8](https://doi.org/10.1007/s00500-015-1853-8)

Rigby, J. (2011). 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](https://doi.org/10.3152/095820211X13164389670392)

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 *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics* (Vol. 1, pp. 784-795). Retrieved from http://www.forschungsinfo.de/Publikationen/Download/SIRTES_ISSI_2013_DFG_FUND_ACK.pdf. Accessed April 11, 2017.

Sirtes, D., & Riechert, M. (2014). A Fully Automated Method for the Unification of Funding Organizations in the Web of Knowledge. In *Proceedings of the 19th International Conference on Science and Technology Indicators* (pp. 594-597). Retrieved from http://www.forschungsinfo.de/publikationen/Download/SirtesRiechert_2014_STI_fully_automatic_FO_unification.pdf. Accessed April 11, 2017.

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](https://doi.org/10.1007/s11192-011-0362-5)