This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

The automatic normalisation challenge: detailed addresses identification

Fernanda Morillo (1, 2), Ignacio Santabárbara (1) and Javier Aparicio (1)

fernanda.morillo@cchs.csic.es, ignacio.santabarbara@gmail.com, javier.aparicio@cchs.csic.es

(1) Instituto de Estudios Documentales sobre Ciencia y Tecnología (IEDCYT). Centro de Ciencias Humanas y Sociales (CCHS). Spanish National Research Council (CSIC), Madrid, Spain.
(2) Corresponding author. IEDCYT-CCHS, CSIC. Albasanz 26-28, 28037-Madrid (Spain).

Abstract

The correct attribution of scientific publications to their true owners is extremely important, considering the detailed evaluation processes and the future investments based upon them. This attribution is a hard job for bibliometricians because of the increasing amount of documents and the raise of collaboration. Nevertheless, there is no published work with a comprehensive solution of the problem. This article introduces a procedure for the detailed identification and normalisation of addresses to facilitate the correct allocation of the scientific production included in databases. Thanks to our long experience in the manual normalisation of addresses, we have created and maintained various master lists. We have already developed an application to detect institutional sectors (issued in a previous paper) and now we analyse the details of particular institutions, taking advantage of our master tables. To test our methodology we have implemented it in a Spanish data set already manually codified (95,314 unique addresses included in the year 2008 on the Web of Science databases). This data was analysed with a full text search (FTS) against our master lists, giving optional codes for each address and choosing which one could be automatically encoded and which one should be reviewed manually. The results of the implementation, comparing the automatic versus manual codes, showed 87% automatically codified records with 1.9% of error. We should review manually only 13%. Finally, we applied the Wilcoxon non-parametric test to show the validity of the methodology, comparing detailed codes of centres already encoded with the automatically encoded ones, and concluding that their distribution was similar with a significance of 0.078.

Keywords

Addresses normalisation; Automatic procedures; Bibliometric indicators; Web of Science databases.

Introduction

Bibliometric studies have become widespread in many forms of scientific analysis and research assessment. The importance of bibliometric indicators is supported by various international reports on the state of the art of science and technology. The Science and Engineering Indicators (National Science Board 2012) are mostly a record of the main quantitative high-quality data of science and

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

technology in USA and the world. These indicators are intended to contribute to the better understanding of the current environment and the best information for future policies. On the other hand, OST (2010) presents a report with various indicators to help to understand French and international science. Research evaluation at national level is very important, as a tool for encouraging scientific productivity, particularly if the results are used to inform the selective funding by the Government. At the same time, it is necessary to be very careful to avoid conceptual and methodological problems in the assessment of the institutions. It is very important the correctly assignation of each publication to the organization that really carries out that research in order to reward only appropriate behaviour and not underestimate valuable research contributions. Given that scientific indicators are increasingly used in policy management, they should be constantly reviewed to increase their validity and predictive ability. This is particularly important because the inadequate interpretation of data can lead to inappropriate identification of future investments (Butler 1999; Van Raan 2005; Feller and Gamota 2007; Almeida et al. 2009; Abramo et al. 2011).

Considering the easy access to bibliographic databases for Bibliometric purposes, it is necessary to take into account their possible coverage limitations and data errors, specially those related with the attribution of publications and citations to a wrong institution, although some databases unify organizational names to a certain extent. Some errors are caused by incorrect database information, inaccurate indication of affiliation by the author, variance in the name of the institutions (e.g. optional translation of foreign names or inconsistency in producing abbreviations for institutional names), identical names for different institutions, etc. At the national level, the study of publications offers fewer problems, but as soon as the approach down nationally, allocation problems grew. As bibliometric indicators are strongly dependent on the methodology, quality control is an essential requirement, but with no published solution. Institutional unification is a significant concern for many researchers and many studies have reported the need to "clean-up" databases in order to reflect institutional counts more accurately (De Bruin and Moed 1990; Butler 1999; Hood and Wilson 2003; García-Zorita et al. 2006; Abramo et al. 2008; Perianes-Rodríguez et al. 2009).

The interest in the normalisation of names of authors and institutions is an important issue for different organizations. In France, ADEST (Measurement of Science and Technology Association) carried out several initiatives to promote standardised signature formulas for personal and institutional names. In Spain, FECYT (Spanish Foundation for Science and Technology) organized a Workshop in 2007, with the participation of several Spanish research groups in Bibliometrics. It was presented a computer tool capable of managing the names of entities, a proposal for their classification and some recommendations to register the names of organizations. We can also find some advice by the very authors. Bador and Lafouge (2005) claim that the lack of standardisation in the affiliations of French universities in published articles penalise their scientific production in the international rankings. They present a case study of the Claude Bernard Lyon 1 University, with 300 articles and only 15% of them with a correct address. With these data, they try to raise awareness about the use of unambiguous official names.

Mallig (2010) presents a proposal and a review of the literature on relational databases for bibliometric analysis arguing that no comprehensive article describes such design. He thinks that similar to other computer processes necessary for data extraction, Bibliometrics seems to keep the secret of its results very jealously. In the same way, we consider that there is scarcely any description of updated methodologies for the standardisation of addresses reflected on journals' papers. This is

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

particularly unusual, considering the growing amount of works that search the most adequate method to identify documents' authors. Addresses are used in studies of disambiguation because an author is related with an affiliation and, if two authors with the same name share an address, the probability that they are the same person is high and it is possible to remove the synonymy and polysemy of the authors' names. Some researchers use external databases with full information of authors (including the affiliation address), because the use of this information allows a binary classification and select/exclude authors depending on which characteristics incorporated in the bibliographic database agree or not with those in the external one. Nevertheless, if authors and/or affiliations are not known in advance, this method cannot be applied (D'Angelo et al. 2011; Gurney et al. 2012; Wang et al. 2012).

Only a few researchers have offered different methods to try to solve this lack of normalisation throughout the time. De Bruin and Moed (1990) create a database with both variations and unified names of publishing organizations, as well as those of cities and countries in which they are located (master files). They use SciSearch database (now Web of Science), which divide addresses in parts of items beginning mainly with organizations and ending with city and country, making easier to label them. Although, this database not always observe the order of segments or the whole information, these authors decide to consider only the first segment of the address as the representation of the overall organization (what we could call institutional sector). In those cases out of the rule, they tested other possibilities to identify a certain organization, but the problem still unsolved when a joint organization was found (e.g. University vs. Research Organization). On the other hand, Katz and Hicks (1997) present a work on desktop Bibliometrics where they introduce a methodology for unification and identification of institutions by extracting information from database addresses, considering the whole information of addresses, but with a laborious manual process. In a similar way, Butler (1999) points out that manual data cleaning and standardisation needs much effort, but with the construction of a thesaurus, the identification of new addresses is always easier and faster. Thijs and Glänzel (2008) present an analysis of publication profiles for the classification of European research institutes. For each country, they compile a list of distinct names of research institutes considering all possible synonyms and spelling variance/errors of them and, if possible, they assign a unique name for each institute. Nevertheless, they do not explain the procedure, so it is implied that this method involves the exclusively manual processing of data.

There is a need for an automated approach and for algorithms designed to extract patterns of similarity from different variables, patterns that can link each item to its corresponding document. In the case of institutions, the growing amount of documents explains the increasing importance of standardisation and the need of some kind of automatic process to ensure the quality of the results in a cost-efficient way (Perianes-Rodríguez et al. 2009). Gálvez and Moya-Anegón (2006 and 2007) develop a methodology based on the finite state transducers taking the advantage of the particular segmentation of databases. Their results are reasonable, but the process is tested with University records and this institutional sector is the best normalised. Bornmann and Ozimek (2012) present a work with Stata commands for importing bibliometric data and processing author address information. Although this study is quite interesting, its main objective is the geographical representation of the data, so the authors consider only part of the addresses information. To perform an exhaustive study, the whole address information and an accurate methodology is required.

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

The basis of our current method is rooted in the process explained in Fernández et al. (1993). A semi-automatic system to standardise and codify the institutional corporate sources was developed creating a master list of addresses. New address were matched against the normalised file comparing the three first characters of all non-empty words linked with an "and" operator. For each new address, the system offered a set of possible candidates with their percentage of agreement and if one was manually accepted, the "institution code" was directly linked to the new address, together with its standardised name. Nevertheless, as the information grew, the response time slowed down, so we needed quicker and better techniques to standardise addresses.

Preliminary research

Throughout the time, our research group have created and maintained different master lists of previously encoded unique addresses from different databases (international and national ones). Thanks to those lists, we have been able to codify automatically new downloaded records with already existing addresses. Thanks the creation of other master tables, those addresses not previously included, were automatically encoded at the country level (located in the last part of the addresses) and, the Spanish ones, at the region/city level (normally located before the country). In a previous study, we developed a method to assign a general code for addresses at the institutional sector level (Morillo et al. 2013). We used data mining techniques to generate a list of key terms that identified sectors, and could be applied to encode new addresses in each category. This general allotment at the sector level produced high quality results that offered possible comparisons with input data and other countries, but offered only an overview of the scientific authorship. The problem of the normalisation job for detailed studies remained unsolved.

Although bibliographic databases usually include corporate sources data (company or organization name for author affiliation), the way of display varies among them. In most cases, they consist of country, region/city and institution. They generally separate information by commas, but the order may change because sometimes they begin with particular data and sometimes with general data. As an international database, like the Web of Science (WoS), divides each address identifying institution, organization and sub-organization, we took advantage of this feature developing a first computer program to identify detailed information. We downloaded and processed new records from this database, exporting documents to different tables with information about titles, publication sources, categories, authors, affiliations, etc. Afterwards, we updated our master lists with the new information, normalising and completing some data when needed. We excluded addresses already part of the WoS master table and identified country and region/city for the rest of the records, creating a new table with the Spanish affiliations. The program checked if there was a similarity equal or over 40% of characters between each new address and one or more of this master table. Similarity was tested contrasting segment by segment of the new record with the words of records in the WoS master table, between commas and in the same order, and discarding segments when not found, starting with the last ones of the address. Addresses, which fulfilled this condition, were considered automatically encoded and not reviewed manually, except new addresses with different region/city than those in the WoS master table. The rest of the records were left to be reviewed manually, although in some cases this coding procedure was easier thanks to the partial encoding information of some addresses. The results obtained with this method and different unique addresses had 1.7% error and 69% of automatic codified records. However, 31% records had to be reviewed manually.

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

In the case of Spanish databases, not previously segmented, we developed a variation of this software. New addresses were compared with the corresponding Spanish master tables and those addresses not already included in them were codified at the country and region/city levels. The Spanish addresses were divided, taking into account the punctuation marks (e.g., commas) and considering keywords[1]. The program checked if there was a similarity equal or over 70% of characters between each new address and one or more of the corresponding Spanish master table. Again, records that fulfilled this condition were considered automatically encoded and not reviewed manually, except new addresses with different region/city than those in the corresponding Spanish master table. This exception exceeded that one found in the WoS database because national databases were less standardise. For this reason, the automatic encoding threshold was also larger. Non-automatically encoded records were left to be reviewed manually, although in some cases this coding procedure was easier thanks to the partial encoding information of some addresses. The results obtained with this method were less satisfactory than those of the WoS database, and as the number of addresses were low (around 6,000 records for each Spanish database), this method was put aside and we concentrated in new and better software for the automation of WoS addresses identification. Besides, thanks to our previous study that identifies institutional sectors, it was possible to determine if these records could remain with a general code or if they could be codified in detail (depending on the type of bibliometric study to be carried out).

Based on results obtained in the first computer program, and the semi-automatic coding at the institutional sector level, we developed a new application to code addresses in steps, from a general to a particular point of view. Thanks to our master list of unique addresses, we could generate an automatic list of terms. As was done in the institutional sector level encoding application (Morillo et al. 2013), the program created master tables with the most frequent combinations of one or more words not considering the empty ones. At the encoding stage, after excluding records from the master list of addresses and identifying country/region/city, the process searched master terms in the new addresses assigning them the corresponding codes when founded. Once the records were encoded at the institutional sector level, the program tried to encode them in detail. Nevertheless, this encoding process from a general to a particular point of view was too broad and produced some contradictions; so we decided to do the opposite. In this way, it was possible to identify some addresses clearly, by assigning them a detailed code, and reducing the number of records encoded only at the institutional sector level. After several experiments, we could check that the combinations of one or more words slowed down the preliminary process and the encoding stage. We realized that we should try to solve the problem in another way, improving the software process.

Objectives and organization

Some Bibliometric studies seem reliable because they are based on objective data. Nevertheless, some automatic analysis do not distinguish among institutions and mixed or subdivide them relying only on the corporate source information, which is not at all standardised. We must identify and normalise addresses to avoid this mess, although the huge amount of information to be processed makes it very difficult. The need of faster techniques to standardise institutions should be a major concern, taking into account the systematic use of bibliometric indicators for national research

[1] E.g.: "Univ", "Dept", "Fac", etc.

assessment, and considering the increase of comparisons at an international level, without knowing the specifics characteristics of national institutions. This article offers a methodology for the detailed automatic identification of addresses. We have a new proposal that avoids the manual revision of a high percentage of addresses and we want to test our automatic methodology against the manual one in order to verify if it is enough efficient for the task.

The rest of this article is organized as follows. First, we present the "Material and methods" section, with a general description of the sources used and a detailed explanation of the software process for addresses identification: master database processing, encoding process and code assignment. Secondly, the "Results and evaluation" of this methodology are analyzed, contrasting the detailed addresses identification reliability and applying statistical tests. Finally, in the "Discussion and conclusions" section, we highlight the key findings and point out the strengths and weaknesses, ending with the future developments of the study.

Material and methods

For different bibliometric purposes our research group use different kind of bibliographic databases. Nevertheless, the most widely used is the Web of Science (WoS). For this study, we develop a new method to automatically encode addresses, and we use a set of Spanish records included in this database (95,314 unique addresses in the year 2008) to test the technique. After applying the method to this set, we perform a statistical analysis using the non-parametric Wilcoxon signed-rank test to contrast manual and automatic encoding, as we observed they did not present a standard normal distribution (Kolmogorov-Smirnov, $p < 0.01$). The confidence level is set at 95% with a margin of error of 0.5. Besides, we applied two non-parametric correlation coefficients, Kendall's tau_b and Spearman's rho, often used to establish whether two variables may be regarded as statistically dependent. Finally, for the assessment of the automatic encoding effect in documents, we compare manual and automatic distribution of Spanish documents, by institutional sector, in the year 2008 (48,224 documents, WoS database). As Spanish institutional sectors we consider: Public Administration (national, regional and local); CSIC (Spanish National Research Council, including joint centres with university and other sectors); Companies (public and private ones); Miscellaneous Sector (organizations with different institutional sectors involved); NPO (Non-profit Organizations); Other PRO (other Public Research Organizations excluded CSIC); Health Sector (including joint documents with universities); University and Others.

Thanks to our previous work of normalisation, we have a master table with a large amount of variations of affiliations and codes manually assigned. Besides the master table, we use a knowledge base that contains all the information needed in the various program processes. The content of this database was introduced manually and with the use of scripts. Keeping this database up-to-date is very important if we want to include any change of any master table. The program uses a very small database manager (SQLite) that includes a full text search (FTS) engine. By default, the FTS matches against whole words in the addresses, regardless their order and only records that contain every word or term found in the search term. Matching search is much faster if we locate all the words. For the rest of the searches, we have to perform all the possible combinations of searches and order them by number of matches. However, the overall result is much more efficient than other methodologies previously tested.

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

Master database processing

First, the program adds all our master records to a new relational database with addresses and codes. As it assigns detailed codes only to Spanish addresses, it selects the Spanish encoded addresses and foreigners with Spain or Spanish region/city segments, and copies them to another table removing country and region/city segments[2], punctuation marks, extra spaces, repeated and stop words[3]. This table is prepared for the full text search (FTS). The entire process lasts just have an hour (Dual-Core processor, 4GHz, 3 GB RAM).

To remove country and region/city, their segment position and words are analysed to establish what they are thanks to our knowledge base. This database includes tables of country words and Spanish region/city words created from data included in the addresses. The program searches the country and the region/city in the last four segments. When there is no exact match for the country "Spain", an approximate string matching is performed with Jaro-Winkler distance >= 0.85.

Examples: "Tech Univ Varna, Dept Math, Varna 9010, **Bulgaria**" (country segment)
"Univ Castilla La Mancha, Dept Comp Sci, Albacete, **Castilla Mancha**, Spain" (region segment)
"Univ Castilla La Mancha, Dept Comp Sci, **Albacete**, Castilla Mancha, Spain" (city segment)
"Univ Basque Country, Dept Paediat, Hosp Cruces, **Bilbao**, Spain" (city segment)

Commonly, it finds Spanish region/city segments included in addresses with "Spain" in the country segment. Nevertheless, occasionally it finds those segments in other countries because some of them have the same name of region/city as Spain, probably due to WoS errors. To solve this, the program checks those segments against our knowledge base, in order to make possible the double allocation (Spain and abroad).

Example: "Childrens Mercy Hosp, Dept Pediat, **Toledo**, OH **USA**" (Toledo, USA)
"Complejo Hosp Toledo, Dept Radiol, **Toledo**, OH USA" (Toledo, Spain)

From time to time, the process finds addresses with "Spain" in the country with a foreign code and checks them in order to evaluate similar addresses and encode them consequently.

Example: "Ctr Invest Opt, **Leon** 37000, Gto, Spain" (Leon, Mexico)
"Ctr Med Regenerat & Terapia Celular Castilla & Le, **Leon**, **Spain**" (Leon, Spain)

The rest of the addresses, those ones with no country or region/city segment assigned, are fully included to the table prepared for the full text search (FTS).

Encoding process

In a second process, the program takes a list of new addresses (collected in a table) and analyses one by one to offer optional codes for each country or centre/organization, considering those included in the master list. These options are added to a table with an identifier of the record for each address, assigned codes, its rank of "context" and its "permatch". "**Context**" is the percentage of words of the new address that matches the address/es of the master list. It is assigned a value based on the rank (Table 1).

[2] Segments are those parts of the records (addresses) between two commas.

[3] The list considers "&", "and", "de", "el", "lo", "las", "la" and "los". Optionally, we can update it with some other words.

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

Table 1. Examples of each rank of context for new addresses

Context	New address	Master address
9 when the address already existed	"Inst Catalana Oncol, Barcelona, Spain"	"Inst Catalana Oncol, Barcelona, Spain"
8 when there is an exact match	hosp 12 octubre madrid complutense univ dept pediat infect dis unit (all words matching)	univ complutense madrid hosp 12 octubre dept pediat infect dis unit (all words matching)
7 for a context of 100%	**univ pais vasco euskal herriko unibertsitatea escuela politecn dept ingn quim ambiente mat technol grp** (100% words matching)	**univ pais vasco euskal herriko unibertsitatea escuela** univ **politecn** politekniko unibertsitate **dept ingn quim ambiente mat technol grp**
6 between 90% and 99%	**csic inst ciencia mat barcelona** networking **ctr bioengn biomat nanomed ciber** (91% words matching)	**csic barcelona inst ciencia mat** dept nanociencia **ciber bioengn biomat nanomed ctr**
5 between 80% and 89%	**univ complutense inst pluridisciplinar unit** lasers (83% words matching)	**univ complutense** brain mapping **unit pluridisciplinar inst** (code for the city: Madrid)
4 between 70% and 79%	**autonomous univ barcelona hosp mar imas imim** reticef urfoa (78% words matching)	**autonomous univ barcelona hosp mar imas imim**
3 between 60% and 69%	**csic** cchs **hist inst** gea (60% words matching)	**csic hist inst**
2 between 50% and 59%	**ims hlth** sa barcelona (50% words matching)	**ims hlth**
1 between 40% and 49%	**hosp san agustin** fac especialista area endocrinol (43% words matching)	**hosp san agustin**
0 less than 40%	fdn publ **hosp** virxen **xunqueira** serv gallego salud (25% words matching)	fp **hosp** virxe **xunqueira** med interna

Although not added to the table, "**inverted context**" is calculated for each rank of context, which is the percentage of words of the address/es of the master list that matches the new address (a minimum threshold is assigned to add alternatives in the table).

Example: **univ barcelona hosp clin idibaps epidemiol** hlth **serv** res unit (new address) (70% words matching in context, rank 4)
univ autonoma **barcelona idibaps hosp clin serv** farmacol clin lab bioestadist **epidemiol** (master address) (58% words matching in inverted context, rank 2)

Finally, "**permatch**" is the percentage of addresses of the master list with a certain code found in each rank.

Example: univ hosp (new address) (code for the region: Vigo)
152 matches in the master list: 2 for miscellaneous codes (1% permatch), 13 for "Complejo Hospitalario Universitario Meixoeiro" (9% permatch), 68 for "Complejo Hospitalario Universitario Xeral Cies" (45% permatch) and 69 for "Complejo Hospitalario Universitario de Vigo" (45% permatch)

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

The program has a minimum threshold assigned by default (usually 70% or higher), but it is possible to change it. This threshold is not only for the inverted context, but also for subsequent automatic encoding based on the values of context and permatch.

1. Firstly, the program verifies the presence of the new addresses in the master list, and in this case, takes their codes and inserts them in the table with a context rank of 9 and a permatch of 100%.
2. Afterwards, the rest of the addresses are encoded by country if they are included in the table of countries. If they are not, they will be considered for manual coding. Those foreign countries with a region/city included in the table of exceptions will be checked and likewise the addresses with "Spain" in the country included in the other table of exceptions (last segments without numbers and words with less than 3 letters).
3. For those records with "Spain" in the country and a Spanish code assignment of region/city (located as it is or by using Jaro-Winkler distance >= 0.85), the program will perform a detailed encoding, removing country and region/city segments, punctuation marks, extra spaces, repeated and stop words. If no region/city can be assigned, it will consider all the segments for the encoding process, except the country one. Several options:

    - The program finds an exact match with the master list (rank 8) with the same code for the region/city. The new address words are the same than those of one or several addresses in the master list, and there are no repeated words either in the new address or in the master list. As there are sometimes records encoded in several ways in the master list, all the possibilities will be displayed specifying the permatch for each one of them.
    - If the program does not find any option with an exact match, it will look for an approximate string matching with the master list (ranks 7 or lower) with the same code for the region/city and the minimum inverted context assigned. All or some of the new address words are identical to those of one or several addresses in the master list, without taking into account the repeated words.
    - If the program does not find the region/city (located as it is or by using Jaro-Winkler), it will consider all the segments for the encoding process, except the country one, with the minimum inverted context assigned. Besides, it will not take into account the region/city code of the master list. Every option will be considered and added to the table with the corresponding code, context, and permatch and it will be excluded for the automatic encoding process.

In these various methods of the encoding process, the permatch is estimated for each rank of context, and the resulting information is added to the table of options. Only records with the minimum inverted context threshold assigned are added to the table.

Code assignment

Finally, the program processes this information to choose which addresses can be automatically encoded and which ones must be reviewed manually.

1. A new table is created to insert automatically encoded records with a suitable code. They will be removed from the original table.

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

2. In any case, the program will insert those records already existing in the master list (context with rank 9) and those codified as foreigners (considering exceptions indicated in the "Encoding process" section).
3. In the rest of cases, the program will consider both the information of the context and the permatch. Addresses will be grouped by ranks of context, starting from the highest rank to the minimum threshold assigned (usually 4 or higher). For each rank of context, there will be a permatch.
    - If there is only one possible code (permatch 100%) in the highest rank, that option is inserted into the new table with automatically encoded records, removing all the other options from the original table.
    - If there is more than one possible code for the rank in process, the permatch is checked to test if it reaches at least the minimum threshold assigned (usually 70% or higher). In that case, the address is inserted with its code into the new table, removing the corresponding records from the original table.
    - If the new address has a code with not enough permatch, the program will seek a lower rank of context until it reaches the minimum threshold assigned.
    - If there is no option to encode automatically a new address in detail, the program will try to encode it in part, at the institutional sector level. The encoding process is done in the same way as in the detailed one, but without taking into account the context, because the inverted context is considered enough. Addresses that reach the required thresholds are inserted into a new table with all the records codified at the institutional sector level, although is also possible to manually codify them in detail.
    - Those new addresses with no options to be automatically encoded will remain without an assigned code until their manual process.
4. A specific graphic interface of the computer program was developed for the manual encoding. Those records with no automatic code will be displayed to be reviewed and assigned their corresponding codes. In any case, the addresses partly encoded (at the institutional sector level) can be encoded in detail when desired.

Results and evaluation

We matched our automatic encoded method of the unique addresses against already encoded data and we compared bibliometric results obtained through both methods. The results of the implementation of the new methodology, with a downloaded Spanish data (WoS 2008) and 95,314 different unique addresses, presented 1.9% error and 82,591 (87%) automatic codified records (25% of them encoded using only partial matching). Records to be reviewed manually were 12,723 (13%). Of these last ones, 36% were also automatically codified at the institutional sector level, leaving 8,159 addresses with data with an under-threshold code. All the automatic process lasted one day and a half. For a person it would have taken more than seven months to codify all the addresses manually, but only one month to codify in detail the records not automatically encoded. Besides, thanks to the specific graphic interface developed for the manual encoding, the codes introduction is even easier and faster.

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

After applying our program, we performed a statistical analysis comparing records already encoded with the automatically encoded ones. Records were grouped by frequency of detailed codes of centres for each method. We use the Wilcoxon non-parametric test to compare their distributions. There were no significant differences between medians of both groups (p = 0.078), so we could conclude that their distribution was similar (Table 2). In addition, both variables were also highly correlated (Kendall's tau_b correlation coefficient is 0.853 and Spearman's rho is 0.924) (Table 3).

Table 2. Wilcoxon Signed Ranks Test

Test Statistics[a]	
	Auto - Manual
Z	-1.761[b]
Asymp. Sig. (2-tailed)	.078

a. Wilcoxon Signed Ranks Test
b. Based on positive ranks.

Table 3. Non-parametric Correlations

Correlations				
			Manual	Auto
Kendall's tau_b	Manual	Correlation Coefficient	1.000	.853**
		Sig. (2-tailed)	.	.000
		N	2004	2004
	Auto	Correlation Coefficient	.853**	1.000
		Sig. (2-tailed)	.000	.
		N	2004	2004
Spearman's rho	Manual	Correlation Coefficient	1.000	.924**
		Sig. (2-tailed)	.	.000
		N	2004	2004
	Auto	Correlation Coefficient	.924**	1.000
		Sig. (2-tailed)	.000	.
		N	2004	2004

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

We also compared manual and automatic distribution of WoS Spanish documents, by institutional sector, in the year 2008 (48,224 documents). The primary y-axis of the Figure 1 shows a column with the total number of documents for each sector in 2008 (dark column) along with another column with the number of automatically codified documents (light column). This light column has also a box with the relative percentage of each institutional sector. The secondary y-axis shows the percentage of error, which is the relative number of wrong assigned documents of each sector. The overall 1.9% error was distributed among institutional sectors unevenly, having more weight in NPO (0.67%) and Companies (0.65%), although each one represented only 4% of the total number of documents. The sector Others, instead, had no error at all, but there was only a 56% automatically encoded documents and this sector represented 1% of the total number of documents. On the other side, University was very well codified (0.01% error and 86% automatically encoded documents) and,

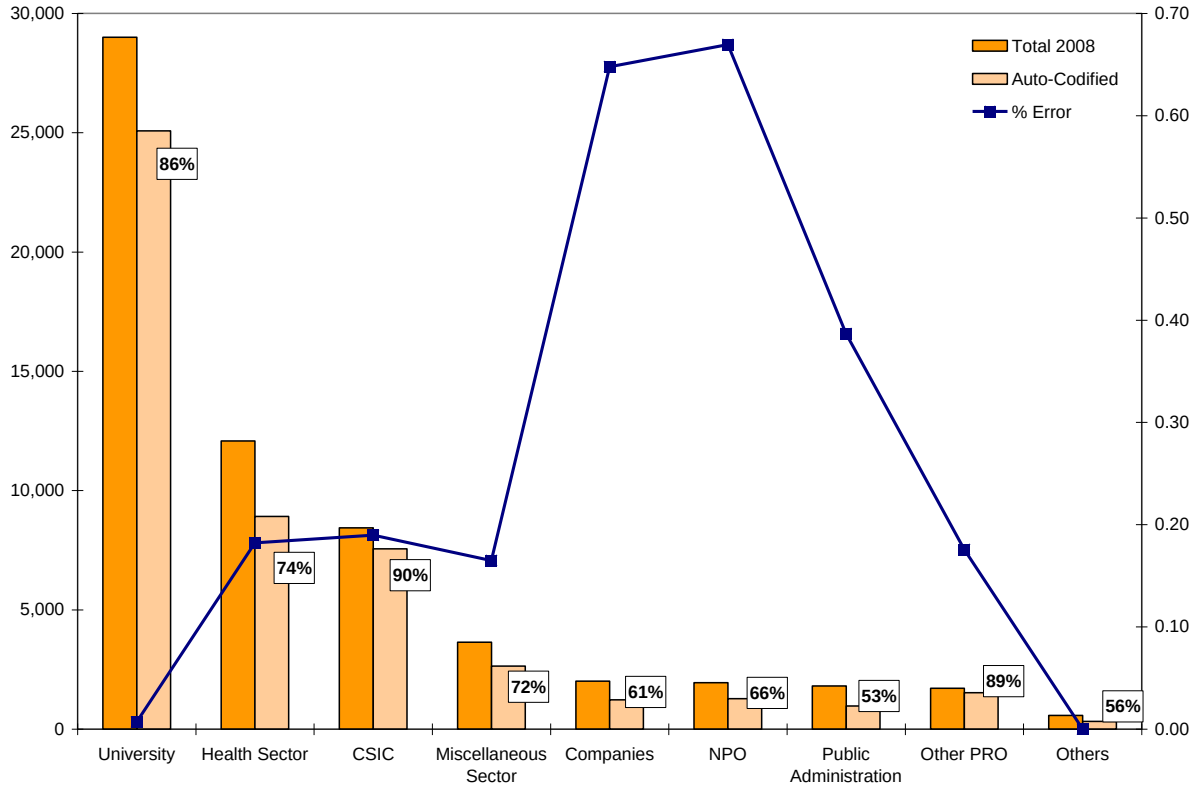
This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

besides, it represented 60% of the total number of documents. Finally, in the third position of output, CSIC showed an outstanding percentage of auto-codified documents (90%) with only 0.19% error.

Figure 1. Distribution of Spanish documents by institutional sector (WoS 2008)



Discussion and conclusions

As bibliometric indicators are increasingly used for evaluation purposes, the adequate identification of producers of scientific publications is very important, considering the lack of standardisation of this data. Some authors have pointed out that this drawback should be attributable mainly to journals (García-Zorita et al. 2006), but we have experienced that sometimes databases include misspellings and wrong data not found in the original documents. We believe that, despite the growing amount of information to deal with and homologate in bibliometric studies, the issue of normalisation of institutions has barely attracted interest in the scientific journals over the past years. This is particularly surprising, given that there are many studies trying to find the most suitable technique for the identification of the authors of the documents. Many of these techniques use the addresses for disambiguation. However, they imply some sort of addresses classification or do not explain how these techniques are carried out.

As it concerned the institutions, proposals for standardisations found in the literature are primarily structural in nature and most of them clearly outdated. Some authors have used the WoS segments to identify and group institutions. However, there are still unsolved cases in which the records have different partitions or elements that identify other addresses. Because of this drawback, those authors

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

insisted on the need to process manually the addresses when they do not fit into the expected structure, requiring subsequent adjustment (Gálvez and Moya-Anegón 2006 and 2007). In this article, we offer a method for the detailed identification and normalisation of addresses that does not require subsequent adjustments. Our proposal can solve most of these cases thanks to its flexibility and the speed of the full text search.

The results obtained by applying our methodology to a Spanish set have been statistically very reliable, allowing us to replace much of the manual encoding job. While it is necessary to start from a list previously encoded, with addresses from any country and/or database, the program uses that information to generate new codes that can quickly complete the initial list. As in the works of Gálvez and Moya-Anegón (2006 and 2007), University sector was very well represented with a minimum error. Besides, in our article, CSIC sector showed the highest percentage of automatically encoded documents. This could be striking, but we had included relatively more records of this sector in the master table, because we study it in detail periodically. However, we have still 13% of addresses to be reviewed manually; therefore, our procedure needs further improvement.

The methodology proposed in this study has meant a considerable upgrading on the techniques used by our group beforehand. Taking into account the recent interest in studies for the identification of the authors, mainly for evaluation purposes, we introduce this method based on the demand for standardisation in increasing information, which implies the need to review and update the current methods. Bibliometric researchers need a methodology for diminishing the output costs, ensuring the quality at the same time, because collaboration in science also influences the amount of data to deal with, making the process more expensive. The possibility of a fast achieving of results is also a requirement of science policy managers, who demand information as up-to-date as possible to make the appropriate decisions. Nevertheless, since these data can have an important impact on the distribution of funds or other resources, it is essential to offer not only quick results but also good results. The results offered in this article are part of a project that aims to identify complex institutions, in an efficient manner, using computer and bibliometric techniques. The next steps will deal with those corporate sources that include various institutions and the detailed analysis of records not automatically encoded using approximate string matching techniques.

Acknowledgements

We wish to thank Adrián Arias Díaz-Faes for his valuable statistical assistance and the anonymous reviewer of this paper for his/her comments and suggestions.

This work is supported by the Spanish Ministry of Science and Innovation (Grant CSO2011-25102).

References

- Abramo, G.; D'Angelo C. A. & Di Costa, F. (2011). National research assessment exercises: the effects of changing the rules of the game during the game. *Scientometrics*, 88 (1), 229-238.
- Abramo, G.; D'Angelo, C. A. & Pugini, F. (2008). The measurement of Italian Universities' research productivity by a non-parametric-bibliometric methodology. *Scientometrics*, 76 (2), 225–244.

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

- Almeida, J. A. S.; Pais, A. A. C. C. & Formosinho, S. J. (2009). Science indicators and science patterns in Europe. *Journal of Informetrics*, 3 (2), 134-142.
- Bador, P. & Lafouge, T. (2005). Rédaction des adresses sur les publications. Un manque de rigueur défavorable aux universités françaises dans les classements internationaux. *La Presse Médicale*, 34 (9), 633–636.
- Bornmann, L. & Ozimek, A. (2012). Stata commands for importing bibliometric data and processing author address information. *Journal of Informetrics*, 6(4):505– 512.
- Butler, L. (1999). Who "owns" this publication? Problems with assigning research publications on the basis of addresses. In: *Proceedings of the Seventh Conference of the International Society for Scientometrics and Informetrics* (pp. 87-96). Universidad de Colima, México.
- D'Angelo, C. A.; Giuffrida, C. & Abramo, G. (2011). A Heuristic Approach to Author Name Disambiguation in Bibliometrics Databases for Large-Scale Research Assessments. *Journal of the American Society for Information Science and Technology*, 62(2), 257–269.
- De Bruin, R. E., & Moed, H. F. (1990). The unification of addresses in scientific publications. In: L. Egghe & R. Rousseau (Eds.), Informetrics 89/90. Selection of papers submitted to the *2nd international conference on bibliometrics, scientometrics and Informetrics*, London, Ontario, Canada, July 5–7, 1989 (pp. 65–78). Amsterdam: Elsevier.
- FECYT (2007). Workshop on "Normalisation of institutions for Bibliometric uses", Barcelona.
- Feller, I. & Gamota, G. (2007). Science Indicators as Reliable Evidence. *Minerva*, 45 (1), 17–30.
- Fernández, M. T., Cabrero, A., Zulueta, M. A., & Gómez, I. (1993). Constructing a relational database for bibliometric analysis. *Research Evaluation*, 3(1), 55–62.
- Gálvez, C. & Moya-Anegón, F. (2006). The unification of institutional addresses applying parameterized finite state graphs (P FSG). *Scientometrics*, 69 (2), 323-345.
- Gálvez, C. & Moya-Anegón, F. (2007). Standardizing formats of corporate source data. *Scientometrics*, 70 (1), 3-26.
- García-Zorita, C., Martín-Moreno, C., Lascurain-Sánchez, M. L., & Sanz-Casado, E. (2006). Institutional addresses in the Web of Science: the effects on scientific evaluation. *Journal of Information Science*, 32(4), 378–383.
- Gurney, T.; Horlings, E. & van den Besselaar, P. (2012). Author disambiguation using multi-aspect similarity indicators. *Scientometrics*, 91 (2), 435–449.
- Hood, W. W. & Wilson, C. S. (2003). Informetric studies using databases: Opportunities and challenges. *Scientometrics*, 58 (3), 587-608.
- Katz, J. S. & Hicks, D. (1997). Desktop scientometrics. *Scientometrics*, 38 (1), 141-153.
- Mallig, N. (2010). A relational database for bibliometric analysis. *Journal of Informetrics*, 4 (4), 564–580.
- Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of address identification. *Scientometrics*, 94 (1), 207–224. DOI 10.1007/s11192-012-0733-6.
- National Science Board (2012). *Science and Engineering Indicators 2012*. Arlington VA: National Science Foundation (NSB 12-01).

This is a **postprint version** of:

Morillo, F.; Santabárbara, I. & Aparicio, J. (2013). The automatic normalisation challenge: detailed addresses identification. *Scientometrics*, 95 (3), 953-966.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-013-0965-0

- OST (2010) *Indicateurs de Sciences et de Technologies. Édition 2010.* Rapport de l'Observatoire des Sciences et des Techniques établi sous la direction de Ghislaine Filliatreau par l'équipe de l'Observatoire des Sciences et des Techniques (OST), Paris.
- Perianes-Rodríguez, A.; Chinchilla-Rodríguez, Z.; Vargas-Quesada, B.; Olmeda-Gómez, C. & Moya-Anegón, F. (2009). Synthetic hybrid indicators based on scientific collaboration to quantify and evaluate individual research results. *Journal of Informetrics*, 3 (2), 91–101.
- Thijs, B. & Glänzel, W. (2008). A structural analysis of publication profiles for the classification of European research institutes. *Scientometrics*, 74 (2), 223–236.
- Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62 (1), 133-143.
- Wang, J.; Berzins, K.; Hicks, D.; Melkers, J.; Xiao, F. & Pinheiro, D. (2012). A boosted-trees method for name disambiguation. *Scientometrics*, doi:10.1007/s11192-012-0681-1.