This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

Towards the automation of addresses identification

Fernanda Morillo (1), Javier Aparicio, Borja González-Albo and Luz Moreno
Instituto de Estudios Documentales sobre Ciencia y Tecnología (IEDCYT)
Centro de Ciencias Humanas y Sociales (CCHS)
Spanish National Research Council (CSIC), Madrid, Spain
fernanda.morillo@cchs.csic.es, javier.aparicio@cchs.csic.es, borja.gonzalezalbo@cchs.csic.es, luz.moreno@cchs.csic.es

(1) Corresponding author. IEDCYT-CCHS, CSIC. Albasanz 26-28, 28037-Madrid (Spain).

Abstract

A new semi-automatic method is presented to standardize or codify addresses, in order to produce bibliometric indicators from bibliographic databases. The hypothesis is that this new method is very trustworthy to normalize authors' addresses, easy and quick to obtain. As a way to test the method, a set of already hand-coded data is chosen to verify its reliability: 136,821 Spanish documents (2006-2008) downloaded previously from the Web of Science database. Unique addresses from this set were selected to produce a list of keywords representing various institutional sectors. Once the list of terms is obtained, addresses are standardized with this information and the result is compared to the previous hand-coded data. Some tests are done to analyze possible association between both systems (automatic and hand-coding), calculating measures of recall and precision, and some statistical directional and symmetric measures. The outcome shows a good relation between both methods. Although these results are quite general, this first overview of the address at the institutional sector level is a good way to develop a second approach for the selection of particular centers. This system has some new features because it provides a method based on the previous non-existence of master tables and it has a certain impact on the automation of tasks. The validity of the hypothesis has been proved taking into account not only the statistical measures, but also considering that the obtaining of general and detailed scientific output is less time-consuming and will be even less due to the feedback of the master tables reused for the same kind of data.

Keywords

Addresses identification; Data mining; Automatic standardization; Performance evaluation; Bibliographic databases.

1 Introduction

The need to promote scientific and technological research entails the establishment of the essential guarantees to ensure that investments are appropriate and meet criteria of scientific excellence and opportunity. Following these criteria, assessment has influenced the development of knowledge, producing pressures as it is an instrument to determine funds allocation and change in organizational structures. In recent years, these pressures have grown due to the exponential increase of science and the shortage of resources, producing additional forces to apply more restrictive funding policies. Bibliometrics aids in these processes providing indicators of scientific production included in bibliographic databases, serving as a complement to the traditional assessment (peer review) and other techniques based on quantitative indicators (eg. patents, R&D expenditures, human resources, etc.) or on qualitative indicators (eg. surveys).

In evaluative bibliometrics, reliable and accurate analysis of the output and impact of research in centers and institutions is very important, as these data can have an influence on the distribution of funds or other resources. Nowadays, bibliographic databases offer extensive information about their indexed documents. However, in these databases, there are a large number of variations for each address or center responsible for the authorship of a document (Sher et al., 1966; Hood and Wilson, 2003; Van Raan, 2005), so there is a risk for mistakes. A previous standardization is required to know the actual number of documents signed by each organization. This is essential in the assessment of

research performance and in the credit recognition. A thorough evaluation of scientific activity is possible if a dependable normalization of addresses is performed in order to ensure the adequate tracking of the publications from sectors and institutions, and to guarantee comparative studies (De Bruin and Moed, 1990; Moed, 1996; Katz and Hicks, 1997; Van Raan, 2005). As government agencies take into account bibliometric methods to assess research and to make decisions about the investment of funds, it is important that the output is attributed to the correct producer of the research, be an institutional sector or a particular organization (Butler, 1999). Besides, the location of the authorship from a miscellaneous sector is a very difficult task, given the interrelationship between institutional sectors and the huge variety of signatures for each one. This problem has been detected widely and some researchers have pointed out the risk of underestimating their participation in the scientific and technological progress (De Bruin and Moed, 1990; Moed, 1996; Hood and Wilson, 2003).

Since scientific production gathered in bibliographic databases begun to be a matter of study, to find a suitable method for the standardization of institutions and authors has been a steady concern. De Bruin and Moed (1990) analyzed addresses included in the SCISEARCH database (specifically University), studied the phenomenology of variations and proposed the unification of signatures under a same denominator, storing this information afterward in reusable master files for other studies. Butler (1999), on the other hand, examined problems encountered to properly determine who the author of a publication was at various levels: national, sectoral and institutional. For the second level, she showed the use of a flexible hierarchical address that enabled adjustment by sector when required, given that no research system remained static and that various sectoral distributions were done in different countries. García Zorita et al. (2006) also analyzed the variety of ways in which an institution may appear in databases. This interest reaches also some institutions like ADEST (Measurement of Science and Technology Association) in France and FECYT (Spanish Foundation for Science and Technology) in Spain. In recent years they have carried out several initiatives to promote standardized signature formulas for personal and institutional names.

Also, Gálvez and Moya Anegón (2006 and 2007) have published studies to unify and standardize corporate sources coming from different databases, applying finite-state techniques. Nevertheless, they pointed out that it was necessary to manually process addresses when they did not match the expected structure and therefore suggested the implementation of complementary methodologies as a possible solution. In the area of processes automation outside the field of Bibliometrics, there are other studies that develop algorithms for the identification of personal and institutional names, using data mining techniques (Patman and Thompson, 2003) or approximate string matching techniques (Navarro et al., 2003). Another work from Navarro & Baeza-Yates (1999) searched algorithms to find words accepting some error on some characters. From other point of view, Christen and Belacic (2005) analyzed addresses based on the statistical Hidden Markov Model (HMM), widely used in natural language processing and that is a probabilistic finite-state machine. The aim of this model is to extract hidden (unknown) information from a string of visible parameters. Particularly novel is the work of Guo et al. (2009), which analyzes postal addresses using a model of Latent Semantic Association (LaSA). LaSA model is built to minimize the human efforts and the size of the control data. This model captures the latent semantic association between words coming from a non-tagged corpus and tries to solve the problem of elements shortage to gather all the characteristics of a particular domain. This technique is based on the Latent Semantic Analysis (LSA) (Jorge Botana et al., 2010), which is a computational model, that exploits the fact that words of the same semantic field will appear together in similar contexts, successfully applied in natural language processing and a very effective tool for simulating human language acquisition and representation. In the context of digital information availability, Jiang et al. (2011) propose a clustering method based on normalized compression distance for the purpose of affiliation disambiguation, considering that affiliation metadata in publications is hard to convert into semantic web data because different authors often express the same affiliation in different ways.

In spite of all these efforts, none of them solve the problem of corporate sources standardization. Although this problem is of concern to many authors and it is possible that many are looking for solutions, very few papers have focused on detailing automated methods for "cleaning" addresses. Given that proper attribution of the scientific production is a prerequisite for correct evaluation of

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

science and for planning future research, this paper tries to offer more efficient methods for the identification of institutions proposing a future standardization by steps or levels, firstly locating general data of the address (institutional sectors involved) and secondly, which is an ongoing research, finding more specific data (organizations or centers themselves). The first goal, classification of addresses in institutional sectors, will allow international comparisons at this level (with data from e.g. OST, 2010) and input-output relations.

2 Objectives and Organization

Indicators based on scientific publications are now an essential tool to analyze and assess the output of research of different R&D organizations. The growth in the number of the organizations and their output increases the complexity in data processing for bibliometric studies and multiplies the difficulty for scientific assessment and planning tasks, as well as the complicated distribution of the scarce resources within the Science and Technology System. Therefore, the development of appropriate methodologies for the process automation to allocate correctly output authorship, can optimize human and economic resources employed in this task, and can speed up and simplify the bibliometric indicators production.

This paper aims to implement these methodologies for the identification of institutional sectors contrasting this output to that one codified by conventional techniques and offering an alternative and more reliable solution than others already proposed. Comparisons will be established by sectors and areas of specialization. This work intends to offer a general image of institutional sectors allowing the periodic follow-up through the proposed methodologies and a detailed future analysis of the scientific output of each of the organizations or centers. Besides, these bibliometric results obtained can guide future scientific research policies. The hypothesis established is that these methodologies and automated techniques proposed can serve to identify institutional sectors, and even the location of particular organizations or centers in the future in an effective and efficient way.

First of all, an overview of the encoding method used so far by our team is provided (Background). Secondly, the Methods and Materials used (General description) and their employment to a specific data set are described (General assignment) followed by an explanation of the Statistical Measures applied. Subsequently, the Results of this method are analyzed, evaluating their reliability to identify addresses from each institutional sector (Evaluation). Finally, results are discussed considering strengths and weaknesses, highlighting key findings and pointing out future developments (Discussion and Conclusions).

2.1 Background

Our research team has been working for many years in Bibliometrics, in the development of new indicators and methodologies as well as in their implementation. In order to produce up-to-date reliable and precise data, processing and standardizing information included in databases is essential. In this regard, the team has extensive experience in normalizing and codifying addresses data and therefore can provide increasingly accurate information on the evolution of the output from the Spanish centers and institutions and even from research teams. Up to now, with already developed techniques the manual processing of around 43% of organizations was required. However, it was considered essential to implement some kind of computer application for the development of this task, due to the growing volume of data to be analyzed (Gomez et al., 2010). It was expected that this implementation would increase the efficiency of the job and would avoid, as far as possible, human errors that could occur.

Some tests to solve the problem of normalization of addresses were done years ago, but the greatest difficulty was the necessary computing power, not very developed at that time (Fernández et al., 1993). Current processors overcome this difficulty and, besides, new studies emerge every day analyzing the feasibility of different algorithms for data management (Navarro et al., 2003; Patman and Thompson, 2003; Christen and Belacic, 2005; Guo et al., 2009). Although these studies are not applied to Bibliometrics, they employ different techniques that can be used for present and future

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

improvements. In this context, our team has developed new automatic procedures for the identification of organizations, which are expected to give answer to these problems and allow carrying out bibliometric studies with effectiveness and efficiency.

Up-to-now, the data processing from organizations was performed in several stages. First, all addresses' variations were unified in one single list. Thanks to previous encodings those records that did not need further manipulation were immediately codified (because there were codes previously assigned). In order to do this task a master table was used. In that table there were several addresses with codes with a translated unified name in another table.

For those records without code, it was carried out a step-by-step method. First of all, a code for the country was attached (very standardized in most of databases and located at the end of the address). For non-Spanish records only this code is assigned (eg. "US" for "United States"). Secondly, for Spanish institutions, a code for the province was added, based on postal codes and names of towns and cities from addresses (eg. "08" for "Barcelona"). Reliability was quite high given that the selected item was the closest one to the country name. For Web of Science (WoS) databases this task was easier because addresses are ordered in segments that identify several elements as organizations and sub-organizations, cities or regions and countries: eg. "Autonom Univ Barcelona, Fac Vet Sci, Bellaterra 08193, Spain" [*organization*: "Autonom Univ Barcelona"; *sub-organization*: "Fac Vet Sci"; *city*: "Bellaterra"; *country*: "Spain"].

In the final step, the non-codified Spanish data were allocated manually and fed back into the master table, in such a way that future instant codification was possible reducing manual processing. Sometimes, a manual review of records was performed to test that automatic coding was done correctly. The automatic system treated around 74% of the total Spanish addresses (which was around 43% of the total number of organizations). The allocation was very high because the master list had already more than 500,000 unique addresses.

3 Methods and Materials

3.1 General description

This methodology can be applied to any set of documents and databases, but the list of terms will vary depending on country of publication and source of data. However, it may be reused, modified and updated with the same kind of output. In this paper the analysis is focused on the Spanish scientific publications (period 2006-2008) included in the international and multidisciplinary Web of Science (WoS) database. A total of 136,821 documents are imported in relational home-made databases for subsequent exploitation. New automated techniques specifically designed for the location of institutional sectors within data are developed in order to identify and normalize addresses (232,981 Spanish addresses and 130,818 unique ones). Methods used in this proposal involve data mining techniques applied specifically to this situation and, also, some truncation techniques have been used to search the root or part of a keyword in the whole address.

In this work, the Spanish institutional sectors considered are: Public Administration (national, regional and local), CSIC (Spanish National Research Council, including joint centers with university and other sectors), Companies (public and private ones), Miscellaneous Sector (formed by organizations in which different institutional sectors are involved), NPO (Nonprofit Organizations), Other PRO (other Public Research Organizations excluded CSIC), Health Sector (including joint documents with universities), University and Others.

3.2 General assignment of codes at the institutional sector level.

Some new techniques are developed to help with the creation of unique terms extracted from the analyzed set in order to normalize or unify address data. First a transformation was done erasing commas, hyphens, etc. except blanks from the addresses. Secondly, a table of keywords has to be created. This involves a certain manual selection process and to produce the table of keywords some of which are filtered out: empty words, zip codes, cities or countries, or non-determinant terms. Given

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

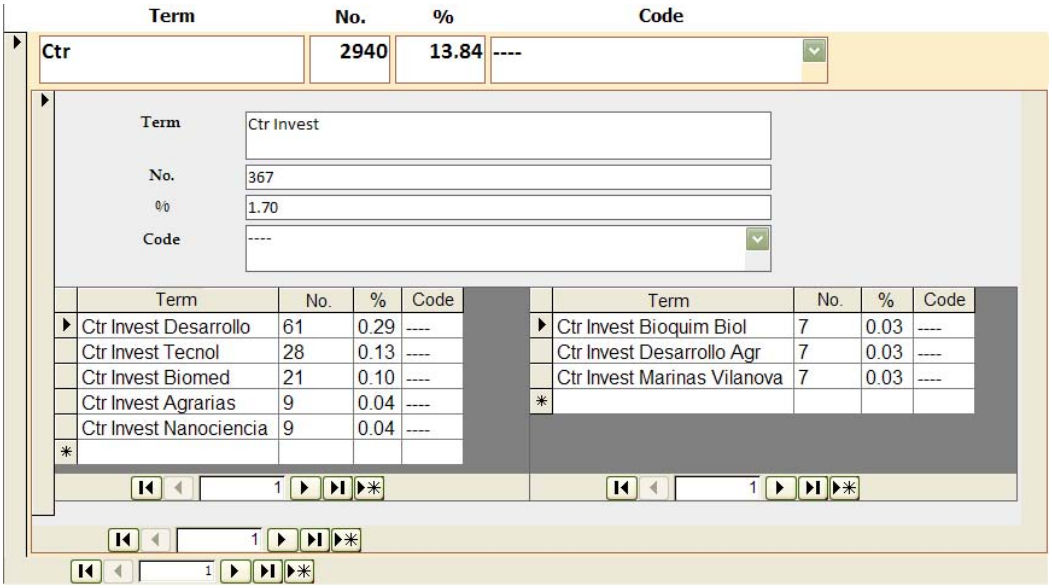
The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

that it is difficult to construct a priori a very comprehensive stop-word list, new terms can be added afterwards depending on the needs. Through the establishment of common criteria or an appropriate procedure manual, errors can be reduced and also possible variations among indexers or coders, like in the full hand-coding process. However, this subjective factor is hard to avoid.

From this table of unique keywords a new one is created showing the most frequent unique terms (found in more than 0.05% of documents of the set). For each case, a combination with another word is searched inside the addresses of the studied records. Then, other two lists are created joining together each two-word term with another word and with other two-words and forming a three and a four-word keyword in the same way (all the two, three or four-word terms must be found in more than 0.01% of documents). In the example shown in Figure 1, the unique keyword "Ctr" forms "Ctr Invest", and the two-word term "Ctr Invest" relates with five three-term connections (eg. "Ctr Invest Desarrollo"), and with three four-word combinations (eg. "Ctr Invest Bioquim Biol").

With these four tables, a form (Figure 1) to select keywords is proposed in order to assist the indexer to accept or reject words or their connections. For each unique term, sequences of two words and for each one of them, combinations of three and four words are presented. The form allows the choice of keywords, identifying a particular institutional sector, constituted by one, two, three or four words selected by the indexer. In each case, examples can be retrieved from the original database (the one to be encoded) as an aid for the decision about what sector must be considered for each keyword or each combination of terms or whether the keyword must be dropped.

Figure 1. Form example to select institutional sector keywords.



No.: number of unified addresses.

The goal is to choose the most significant keywords among the most frequent ones, but also to select those that avoid errors or problems with other institutional sectors (eg. "Inst" alone can lead to University: "Inst Vasco Educ Fis", to CSIC: "Inst Microelect Barcelona", to Health Sector: "Inst Municipal Invest Med", to NPO: "Inst Adv Studies Energy", etc.). To assist the task of the indexer:

- The application seeks and pre-selects unique terms frequently found in the first position (in more than 50% of their records), to review them initially as potential candidates. Only unique terms found in more than 0.05%[1] of records are considered. For example, the studied set contains 74,843 unique addresses with the keyword "Univ", and 61,887 begin with this

[1] This threshold has been established, after several tests, to ease manual work, ensuring that the sample will be significant with the minimum error.

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

keyword (83%). As this is more than 50% of "Univ" records, the application will pre-select this term as initial candidate.

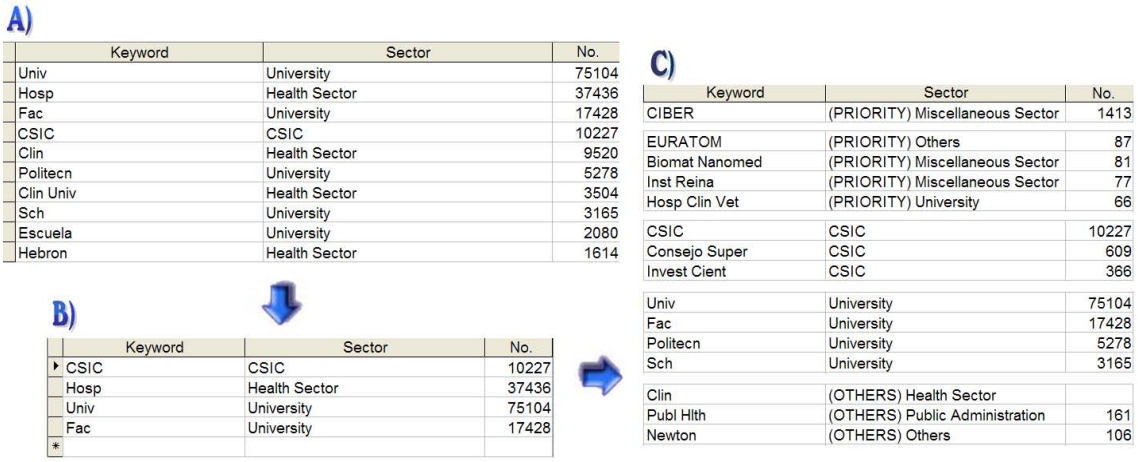
- In addition, the process pre-selects two-term keywords when they have the same frequency as the unique terms. Only keywords found in more than 0.01%[1] of records are considered. For example, the keyword "Irsi" is present in 9 of the unique addresses and the two-term "Irsi Caixa" also appears in these 9 records. As both have the same frequency, the application will pre-select and give priority to the two-term as initial candidate, because it can form a more specific keyword.

Once the first list of keywords is selected and produced by the indexer, it is sorted according to the most interesting criteria. In principle, the most frequent keywords prevail (Figure 2, A) but, to prevent errors, due to miscellaneous institutional sectors, a preference order by sector is given. In this paper, it is considered firstly CSIC, followed by Miscellaneous Sector, NPO, Health Sector, Companies, Other PRO, Public Administration, University and Others (see some examples in Figure 2, B). CSIC is in first place because it includes joint centers with university and other sectors and it is the main target of our work. However, such output can be considered in another sector if the order of priority is changed or if a more detailed identification within each sector is done.

Subsequently the indexer follows this order or changed it if there is a certain keyword that should prevail no matter its sector. For example, CIBER is a biomedical research network in which different institutional sectors participate. Therefore, a fictitious institutional sector known as "PRIORITY" is created. For keywords that indexer wants to postpone, the existing sector known as "OTHERS" is used. In both cases, in a final stage the real institutional sector is automatically assigned (Figure 2, C).

This ranking is formed in a semi-automatic way adjusting just what is required. To set the ranking for the coding of the institutional sectors it is necessary to make some prior checking to avoid most of the errors and to assess which order reflects better the purpose of a particular study.

Figure 2. Preference order example by sector and keyword.



A)

Keyword	Sector	No.
Univ	University	75104
Hosp	Health Sector	37436
Fac	University	17428
CSIC	CSIC	10227
Clin	Health Sector	9520
Politecn	University	5278
Clin Univ	Health Sector	3504
Sch	University	3165
Escuela	University	2080
Hebron	Health Sector	1614

B)

Keyword	Sector	No.
CSIC	CSIC	10227
Hosp	Health Sector	37436
Univ	University	75104
Fac	University	17428

C)

Keyword	Sector	No.
CIBER	(PRIORITY) Miscellaneous Sector	1413
EURATOM	(PRIORITY) Others	87
Biomat Nanomed	(PRIORITY) Miscellaneous Sector	81
Inst Reina	(PRIORITY) Miscellaneous Sector	77
Hosp Clin Vet	(PRIORITY) University	66
CSIC	CSIC	10227
Consejo Super	CSIC	609
Invest Cient	CSIC	366
Univ	University	75104
Fac	University	17428
Politecn	University	5278
Sch	University	3165
Clin	(OTHERS) Health Sector	
Publ Hlth	(OTHERS) Public Administration	161
Newton	(OTHERS) Others	106

No.: number of unified addresses.

Finally, this upgraded sectors' table is used to standardize or codify the original database. The process searches each keyword inside the addresses and assigns a code to each address when found. If it finds several keywords, one of them is selected first (based on the institutional sector and the frequency of the term). In the case of co-occurrence of institutional sectors it is advisable to make two analyses: a first selection to assign addresses to a given preference institutional sector and a second one to assign other institutional sectors among those records already standardized. In this study this second normalization is not made because a properly "miscellaneous sector" including those institutions belonging to two or more sectors was created.

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

3.3 Statistical measures

To codify each institutional sector, and the combination of various institutions (miscellaneous), a comparative study of the usefulness of different methods and algorithms is done. The results achieved are compared with our currently used methods and with other methodologies employed by other authors in terms of effectiveness and efficiency. The validation of these new techniques is tested using previously manually standardized data, contrasting the information from both sources to obtain precision and recall indicators, and some other statistical measures. In this study, "recall" is the percentage of the hand-coded records that have also been codified automatically, and "precision" is the percentage of the automatically coded records which match the hand-coded ones (similar measures are used by Gálvez and Moya Anegón, 2006).

Statistical measures are based upon a Crosstabulation, which is a combination of two frequency tables arranged so that each cell in the resulting table represents a unique combination of specific values of crosstabulated variables: the records hand-coded in each sector and those automatically coded. By examining this combination, we can identify relations between crosstabulated variables. In addition, this procedure provides a series of tests and measures of association. The structure of the table and whether categories are ordered determine what test or measure to use:

Pearson's Chi-square test is a nonparametric test that verifies the independence of two variables, through the presentation of data in cross tabulations. The contingency coefficient Chi-square is used to perform a formal contrast to the null hypothesis of independence of the variables A (hand-coding) and B (automatic coding), whose sample information is contained in the Crosstabulation. The alternative hypothesis is the existence of association between the two variables. The greater the value of Chi-square, the less plausible is that the hypothesis is correct. In the same way, the more the value of Chi-square approaches to zero the more adjusted are both distributions.

The Lambda coefficient (symmetric and asymmetric lambdas and Goodman and Kruskal's tau) no longer depends on Chi-square. It is a measure of the strength of association of the cross tabulations when the variables are measured at the nominal level. Assuming that A has been chosen as the explained factor and B as the explanatory one, the ability of B to predict A is evaluated by Lambda coefficient of A and vice versa (Lambda B). This value ranges from 0 to 1 and is designed for asymmetric measures. For this reason, when it is not possible to determine which of the two factors is the explanatory or the explained (A or B), the use of the symmetric version should be considered (though its drawback is that it is extremely sensitive to the presence of unbalanced marginal totals). A value close to 1 means that the independent variable perfectly predicts the dependent variable and a value close to 0 means the opposite. Measures of association can be interpreted in terms of the proportional reduction in error when values of the independent variable are used to predict values of the dependent variable.

The Uncertainty coefficient is another measure for variables at the nominal level and gives the degree of linear relationship between two factors or attributes. It provides data for A and B and also for both (symmetric version), if there is no known relationship of dependence between attributes. It indicates also the proportional reduction in error when values of one variable are used to predict values of the other variable. For example, a value of 0.83 indicates that knowledge of one variable reduces error in predicting values of the other variable by 83%.

Cohen's Kappa coefficient estimates the percentage of agreement between the evaluations of the two studied variables when both are rating the same object. It is generally thought to be a more robust measure than simple percent agreement calculation since Kappa takes into account the agreement occurring by chance. A value of 1 indicates perfect agreement and a value of 0 indicates that agreement is no better than chance (IBM SPSS Statistics Base 19, 2010; Pérez, 2005).

Finally, in order to test the reliability of this methodology by areas of specialization, a classification based upon the Current Contents one is used: Agriculture, Biology & Environmental Sciences; Arts &

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

Humanities; Chemistry; Clinical Medicine; Engineering, Computing & Technology; Life Sciences; Mathematics; Multidisciplinary Sciences; Physics; and Social & Behavioral Sciences.

4 Results

4.1 Evaluation of the institutional sector level.

Taking into account the analyzed set, all addresses are studied to explore possible association between both systems: the hand-coding and the automatic one. Crosstabulation shows a good relation between both methods. Only 3.3% (7,652) of the hand-coded records have no automatic code (not shown in Tables Ia and Ib).

Besides, as explained in the Method section, measures of recall and precision can be considered. For example, of 6,662 hand-coded records as Companies, 4,882 are likewise automatically coded (which is within hand or a recall of 73%). Considering the 5,068 automatically coded records as Companies, those 4,882 represents a precision of 96.3% (within auto). The rest of hand-coded Companies' records are automatically coded elsewhere. University has the highest matching with a recall of 98.5% and precision of 98.4%. Only four sectors have a recall or within hand lower than 90%, although those sectors represent a little over 8% of the total Spanish addresses (Tables Ia and Ib). On average, the percentage of precision is 98% and the percentage of recall is 83% (not shown in the table).

Table Ia. Crosstabulation between hand and automatic coding (total count).

		Automatic coding									Total Hand
		Companies	CSIC	Health Sector	Miscellaneous Sector	NPO	Other PRO	Others	Public Administration	University	
Hand coding	Companies	4882	2	116	15	7	0	0	10	33	6662
	CSIC	11	25818	15	31	14	3	0	4	399	26635
	Health Sector	16	15	62973	79	91	3	0	57	946	65317
	Miscellaneous Sector	13	29	122	8585	19	18	0	10	166	9277
	NPO	24	6	66	22	4163	0	1	33	81	5341
	Other PRO	3	14	7	2	7	4823	0	7	8	4974
	Others	22	12	37	36	3	26	823	13	38	1781
	Public Administration	17	13	124	7	24	18	0	4419	46	6089
	University	80	75	222	53	28	44	0	124	105256	106905
Total Auto		5068	25984	63682	8830	4356	4935	824	4677	106973	232981

Records with no automatic code are not shown in the table (3.3%), but they are considered in the column Total Hand.

This is a postprint version of:

Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

Table Ib. Crosstabulation between hand and automatic coding (percentages).

			Automatic coding								Total Hand
			Companies	CSIC	Health Sector	Miscellaneous Sector	NPO	Other PRO	Others	Public Administration	University	
Hand-coding	Companies	% within Hand	73.3%	0.0%	1.7%	0.2%	0.1%	0.0%	0.0%	0.2%	0.5%	100.0%
		% within Auto	96.3%	0.0%	0.2%	0.2%	0.2%	0.0%	0.0%	0.2%	0.0%	2.9%
	CSIC	% within Hand	0.0%	96.9%	0.1%	0.1%	0.1%	0.0%	0.0%	0.0%	1.5%	100.0%
		% within Auto	0.2%	99.4%	0.0%	0.4%	0.3%	0.1%	0.0%	0.1%	0.4%	11.4%
	Health Sector	% within Hand	0.0%	0.0%	96.4%	0.1%	0.1%	0.0%	0.0%	0.1%	1.4%	100.0%
		% within Auto	0.3%	0.1%	98.9%	0.9%	2.1%	0.1%	0.0%	1.2%	0.9%	28.0%
	Miscellaneous Sector	% within Hand	0.1%	0.3%	1.3%	92.5%	0.2%	0.2%	0.0%	0.1%	1.8%	100.0%
		% within Auto	0.3%	0.1%	0.2%	97.2%	0.4%	0.4%	0.0%	0.2%	0.2%	4.0%
	NPO	% within Hand	0.4%	0.1%	1.2%	0.4%	77.9%	0.0%	0.0%	0.6%	1.5%	100.0%
		% within Auto	0.5%	0.0%	0.1%	0.2%	95.6%	0.0%	0.1%	0.7%	0.1%	2.3%
	Other PRO	% within Hand	0.1%	0.3%	0.1%	0.0%	0.1%	97.0%	0.0%	0.1%	0.2%	100.0%
		% within Auto	0.1%	0.1%	0.0%	0.0%	0.2%	97.7%	0.0%	0.1%	0.0%	2.1%
	Others	% within Hand	1.2%	0.7%	2.1%	2.0%	0.2%	1.5%	46.2%	0.7%	2.1%	100.0%
		% within Auto	0.4%	0.0%	0.1%	0.4%	0.1%	0.5%	99.9%	0.3%	0.0%	0.8%
	Public Administration	% within Hand	0.3%	0.2%	2.0%	0.1%	0.4%	0.3%	0.0%	72.6%	0.8%	100.0%
		% within Auto	0.3%	0.1%	0.2%	0.1%	0.6%	0.4%	0.0%	94.5%	0.0%	2.6%
	University	% within Hand	0.1%	0.1%	0.2%	0.0%	0.0%	0.0%	0.0%	0.1%	98.5%	100.0%
		% within Auto	1.6%	0.3%	0.3%	0.6%	0.6%	0.9%	0.0%	2.7%	98.4%	45.9%
Total Auto		% within Hand	2.2%	11.2%	27.3%	3.8%	1.9%	2.1%	.4%	2.0%	45.9%	100.0%
		% within Auto	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

% within Hand dark shaded = RECALL and % within Auto light shaded = PRECISION.
Percentages of records with no automatic code are not shown in the table (3.3%), but they are considered in the column Total Hand.

The validity of the automatic coding is checked against some statistical measures. Chi-Square tests show a significant linear association between both variables (Table II). According to the directional measure Lambda, there is a reduction of the error of 0.911 to forecast the hand-coding. Given the Uncertainty Coefficient, the reduction of the error would be 0.867. In both cases, the null hypothesis is not assumed with a 99% confidence coefficient (Table III). Also symmetric measures show good results. The closer to 1 are the values of Kappa, the greater relationship between two variables. In this case the value is 0.931 with a 99% confidence coefficient (Table IV). All these results demonstrate the strength of the proposed method.

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

Table II. Chi-Square Tests.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1513569.643	72	0.000
Likelihood Ratio	618251.172	72	0.000
N of Valid Cases	232981		

Table III. Directional Measures.

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	0.917	00.001	470.504	0.000
		Hand Dependent	0.924	0.001	469.342	0.000
		Auto Dependent	0.911	0.001	462.168	0.000
	Goodman and Kruskal tau	Hand Dependent	0.916	0.001		0.000[c]
		Auto Dependent	0.881	0.001		0.000[c]
	Uncertainty Coefficient	Symmetric	0.878	0.001	574.352	0.000[d]
		Hand Dependent	0.889	0.001	574.352	0.000[d]
		Auto Dependent	0.867	0.001	574.352	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.

Table IV. Symmetric Measures.

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	0.931	0.001	789.816	0.000
N of Valid Cases		232981			

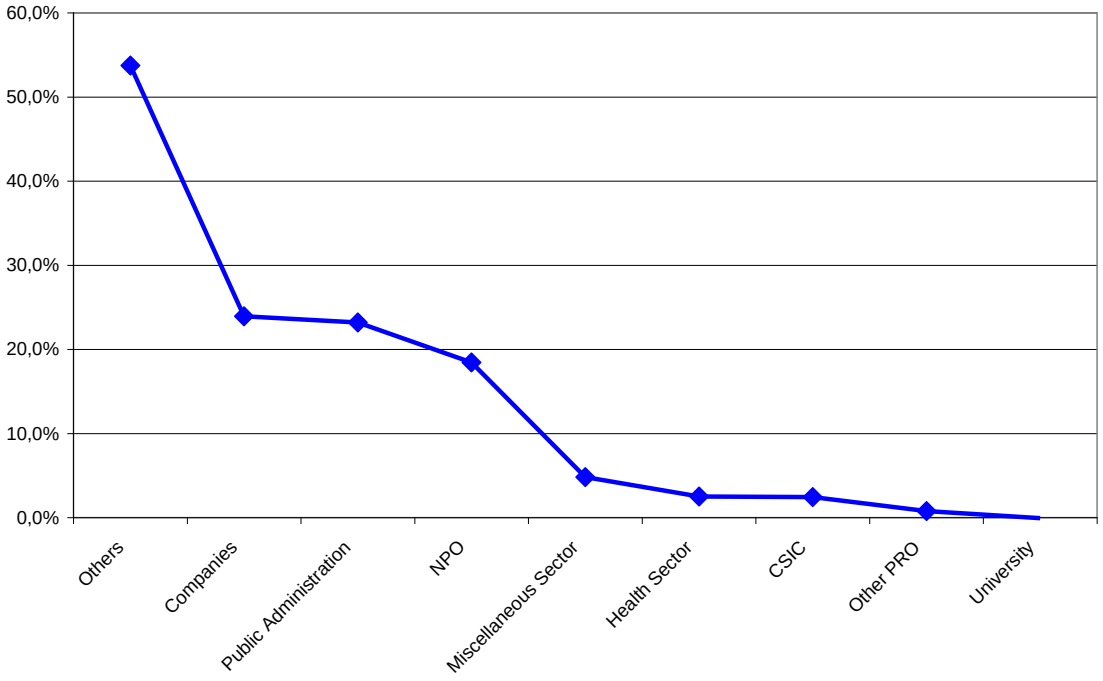
a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

A comparison between this sector coding procedure and the previous hand-coding method was established studying differences of percentages with both methods by sector. In figure 3 we observed practically no difference for Universities, Other PRO, CSIC, Health Sector and Miscellaneous Sector. These institutional sectors represent more than 91% of all addresses variations and 95-89% of documents (depending on whether the total number or the summation is considered). The automatic coding is less accurate for Companies, Public Administration and NPO sectors and, of course, with those considered as Others (Tables Ia, Ib and Figure 3).

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

Figure 3. Differences of percentages between hand and automatic coding by institutional sector.

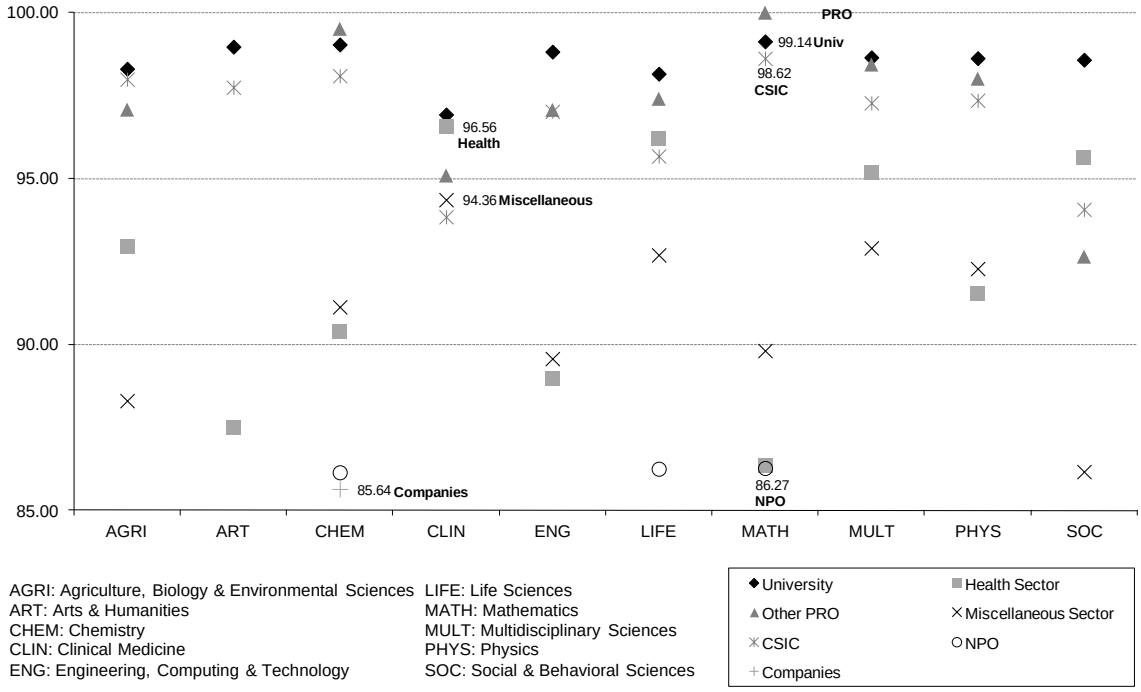


The test of the automatic coding method by thematic areas of specialization shows that the best recall results for Companies is Chemistry (86%); Mathematics has the highest percentages of recall for Other PRO, University, CSIC and NPO (> 86% in all cases); and Clinical Medicine is the best coded area for Health and Miscellaneous Sectors (> 94%). On the other hand, in general, the best automatically coded sector is University, with percentages higher than 96% for almost all areas, which is normal if we consider the good results in the general Crosstabulation (over 98% of recall and precision, Table 1). Only Chemistry and Mathematics present the highest scores for another institutional sector: Other PRO (> 99%) (Figure 4). Public Administration is not shown in the figure due to its lower percentages of recall and precision, although in Agriculture, Biology & Environmental Sciences has a recall of almost 80%. The results with less than 85% represent only a little over 8% of the total data (including sector "Others"), so that its impact is very low.

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

Figure 4. Percentages of recall by thematic area and institutional sector (only cases with recall over 85% are shown).



AGRI: Agriculture, Biology & Environmental Sciences
ART: Arts & Humanities
CHEM: Chemistry
CLIN: Clinical Medicine
ENG: Engineering, Computing & Technology
LIFE: Life Sciences
MATH: Mathematics
MULT: Multidisciplinary Sciences
PHYS: Physics
SOC: Social & Behavioral Sciences

Percentages indicated in the Figure correspond to the highest values of each sector.

Due to the low accuracy for "Others" sector, some tests were made to calculate directional and symmetric measures excluding it for the Crosstabulation. However, small differences were found: Lambda shows a reduction of the error of 0.917 instead of 0.911 to forecast the hand-coding, in the Uncertainty Coefficient, the reduction was of 0.870 instead of 0.867 and, finally Kappa shows a relationship between the two variables of 0.936 instead of 0.931. Considering that this sector includes only 0.8% of the Spanish addresses, its influence in the total statistical measures is logically limited.

The database used is WoS 2006-2008 (136,821 documents and 232,981 addresses), with 130,818 Spanish addresses after the unification of signatures under a same denomination. With this set of analysis, it takes around one person/month (24 days) to encode records at the sector level, based upon our own experience. With the method proposed in this paper only 3 days are needed to obtain sector keywords, and less than 1 hour for the encoding process. Time saving is remarkable. Obviously, the time spent will diminish once a list of some keywords is already produced, since the table will need only an update.

5 Discussion

The application presented in this paper was specifically developed for the task of identifying and codifying addresses for bibliometric purposes, and can facilitate the job avoiding mistakes made otherwise. This approach can be applied to those situations where no information is available (like master lists of organizations). Besides, institutional sectors concerned will depend on the different criteria applied as it is a very flexible method. The validity of the hypothesis has been proved taking into account the time-saving and the percentage of errors accepted (non assigned addresses at the institutional sector level represent only 3.3% of the total Spanish data). Regarding bibliometric indicators obtained with these data, very little differences can be observed. Furthermore, taking into account that University is the best identified sector, it is also the best represented because of its 60% of the Spanish production (Gómez et al., 2011). Also, a future time-saving will be possible as the keywords list produced is very easy to complete. Nevertheless, a drawback for the semi-automatic coding processes is that institutions are living bodies in constant change and sometimes the

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

assignation could need some updating. Anyway, it is also possible to do that quite fast and easily, being a very effective and efficient method.

As compared to other works as those of Gálvez and Moya Anegón (2006 and 2007) this method could be considered quite a good alternative based upon data mining techniques and some truncation searching. These authors study a way to unify addresses through finite state techniques (FST) establishing equivalence relations between variants with a semi-automatic method to facilitate this task. In their first work (Gálvez and Moya Anegón, 2006) they evaluate the method analyzing documents from the University of Granada. They get a fairly high recall (87%), with 719 number of variants, an under-standardization of 12% and a precision of 100%. In the other approach, dealing with University samples of databases INSPEC, MEDLINE, and CAB Abstracts (Gálvez and Moya Anegón, 2007), the authors generate a list of variants without duplicates with ProCite, transforming and segmenting addresses and using FST. Measures of recall for INSPEC are 99% (1,192 possible address patterns), for CAB 98% (1,307) and for MEDLINE 94% (1,416) with an under-standardization, in this last database, of 4.2%.

In this paper, although it is done only at the sectoral level, 130,818 Spanish unique addresses are used. Besides, the University sector, analyzed in those other works, is less affected in this work by errors or under-standardization because of a lower number of different addresses' variants. Only 1% of University records are not automatically coded, with a recall of 98.5% and a precision of 98.4%. On the other hand, a drawback of Gálvez's method is that it is necessary to know a priori the types of addresses and it does not solve the problems of ambiguity which can be found, so complementary techniques must be applied in such cases. In terms of hand labor, they talk about a large initial investment repaid by the fact that the tools can be used repeatedly, but they do not offer any additional information about the time spent.

Moreover, future improvements are expected in order to increase the application performance. Among the methods and algorithms that will be analyzed for their possible application to our task, the most interesting are those that maximize automation of work, minimizing human effort with an optimal result. We are currently working on the elaboration of a new application that will identify not only sectors, but also specific organizations. Given the need for quick and reliable results, we will take advantage of the information collected on our master tables to develop automatic lists of terms and use them in the standardization of new records. For this ongoing research the WoS database is used. Besides, we will try to work the other way round: considering the information inside new addresses to look them up in our master tables. In this database, each address is divided into segments that identify different elements (organizations and sub organizations, cities or regions and countries). For each record, each segment, word or combinations of them can be looked up in the master list to give a code at the organization level before looking for the next one. As exact match (100%) were already allocated, as explained in the Background section, in this approach partial correspondence can be searched, based on the location of the different possibilities within the master table.

6 Conclusions

The findings of this study and the ongoing and future improvements provide useful tools for the bibliometric analysis of corporate sources making the obtaining of different kinds of indicators more simple. The advantage of this method is that all the steps can feedback a master table (that can be created during the process for each specific purpose) and locate an increasingly accurate information about the addresses thanks to the ongoing research with the organization unification, considering techniques based upon segments or words searching. This precise bibliometric indicators will contribute as a starting point to analyze and assess the current situation and future actions to respond better to the existing socio-economic needs. For bibliometric researchers, the possibility of developing this methodology of addresses normalization will help in the production of indicators, providing them with reliable results. Usual bibliometric tasks are becoming increasingly more expensive, due to the raise in the number of signatures through increasing collaboration, and for this reason, it is necessary to optimize resources, taking into account that bibliometric studies affect scientific assessment processes.

This is a postprint version of:
Morillo, F.; Aparicio, J.; González-Albo, B. & Moreno, L. (2013). Towards the automation of addresses identification. *Scientometrics*, 94 (1), 207-224.

The final publication is available at Springer via http://dx.doi.org/10.1007/s11192-012-0733-6

The methodologies proposed and the tools developed, open the door to other areas that also require streamlining and improving of their processes. Bibliometric studies may also provide useful information for the evaluation and research planning, guiding and supporting decisions about the distribution of economic, human and material resources and, thanks to the proposed automated techniques, bibliometric data will be updated faster and more easily.

7 References

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. Universidad de Colima, México. P. 87-96.

Christen, P. and Belacic, D. (2005) Automated Probabilistic Address Standardization and Verification. 4th Australasian Data Mining Conference AUSDM'05.

De Bruin, R.E. and Moed, H.F. (1990) The unification of addresses in scientific publications. In: Egghe, L. and Rousseau, R (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. Amsterdam, Elsevier Science Publishers, P. 65-78.

Fernández, M.T.; Cabrero, A.; Zulueta, M.A. and Gómez, I. (1993) Constructing a relational database for bibliometric analysis. Research Evaluation, 3 (1): 55-62.

Gálvez, C. and Moya Anegón, F. (2006) The unification of institutional addresses applying parameterized finite state graphs (P FSG). Scientometrics, 69 (2): 323-345.

Gálvez, C. and 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.

Gómez, I.; Bordons, M.; Morillo, F.; Moreno, L.; Aparicio, J.; Díaz-Faes, A.A. and González-Albo, B. (2011) La actividad científica del CSIC a través de indicadores bibliométricos (Web of Science, 2006-2010). Instituto de Estudios Documentales sobre Ciencia y Tecnología (IEDCYT-CCHS), Madrid.

Guo, H.; Zhu, H.; Guo, Z.; Zhang, X. & Su, Z. (2009) Address Standardization with Latent Semantic Association. 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD'09, June 28–July 1, Paris, France. P. 1155-1163.

Hood, W.W. and Wilson, C.S. (2003) Informetric studies using databases: Opportunities and challenges. Scientometrics, 58 (3): 587-608.

IBM SPSS Statistics Base 19 (2010). © Copyright SPSS Inc. 1989, 2010.

Jiang, Y.; Zheng, H.T.; Wang, X.; Lu, B. and KaihuaWu (2011) Affiliation disambiguation for constructing semantic digital libraries. Journal of the American Society for Information Science and Technology, 62 (6): 1029-1041.

Jorge Botana, G.; León, J.A.; Olmos, R. and Hassan Montero, Y. (2010) Visualizing polysemy using LSA and the predication algorithm. Journal of the American Society for Information Science and Technology, 61 (8): 1706-1724.

Katz, J.S. and Hicks, D. (1997) Desktop Scientometrics. Scientometrics, 38 (1): 141-153.

Moed, H.F. (1996) Differences in the Construction of SCI Based Bibliometric Indicators Among Various Producers: A First Overview. Scientometrics, 35 (2): 177-191.

Navarro, G. and Baeza-Yates, R. (1999) Very fast and simple approximate string matching. Information Processing Letters, 72: 65–70.

Navarro, G.; Baeza Yates, R. and Azevedo Arcoverde, J.M. (2003) Matchsimile: A Flexible Approximate Matching Tool for Searching Proper Names. Journal of the American Society for Information Science and Technology, 54 (1): 3–15.

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.

Patman, F. and Thompson, P. (2003) Names: A New Frontier in Text Mining. Lectures Notes in Computer Science, 2665: 27-38.

Pérez, C. (2005) Técnicas Estadísticas con SPSS © 12. Aplicaciones al análisis de datos. Pearson Educación, S.A. Madrid.

Sher, I.H.; Garfield, E. and Elias, A.W. (1966) Control and Elimination of Errors in ISI Services. Journal of Chemical Documentation, 6 (3): 132.

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.