The citation from patents to scientific output revisited: A new approach to *Patstat /Scopus* matching

Vicente P. Guerrero-Bote; Rodrigo Sánchez-Jiménez; Félix De-Moya-Anegón

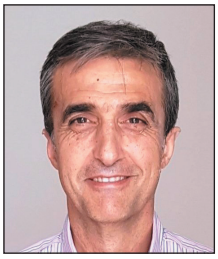
Nota: Este artículo se puede leer en español en:
http://www.elprofesionaldelainformacion.com/contenidos/2019/jul/guerrero-sanchez-de-moya_es.pdf

How to quote this article:

Guerrero-Bote, Vicente P.; **Sánchez-Jiménez, Rodrigo**; **De-Moya-Anegón, Félix** (2019). "The citation from patents to scientific output revisited: a new approach to *Patstat / Scopus* matching". *El profesional de la información*, v. 28, n. 4, e280401.

https://doi.org/10.3145/epi.2019.jul.01

Manuscript received on May, 03rd 2019
Accepted on May, 27th 2019



Vicente P. Guerrero-Bote ✉
https://orcid.org/0000-0003-4821-9768

SCImago Research Group, Spain
Universidad de Extremadura. Facultad de Ciencias de la Documentación y la Comunicación
Plazuela Ibn Marwam, s/n.
06071 Badajoz, Spain
guerrero@unex.es

Rodrigo Sánchez-Jiménez
https://orcid.org/0000-0002-3685-7060

SCImago Research Group, Spain
Universidad Complutense de Madrid, Facultad de Ciencias de la Documentación
Santísima Trinidad, 37. 28010 Madrid, Spain
rodsanch@ucm.es

Félix De-Moya-Anegón
https://orcid.org/0000-0002-0255-8628

SCImago Research Group, Spain
felix.moya@scimago.es

Abstract

Patents include citations, both to other patents and to documents that are not patents (NPL, Non-patent literature). Non-patent literature (NPL) includes articles published in scientific journals. The technological impact of scientific works can be studied through the citations they receive from patents, just like the scientific impact of articles can be analyzed through the citations. The NPL references included in patents are far from being standardized, so determining which scientific article they refer to is not a trivial task. This paper presents a procedure for linking the NPL references of the patents collected in the *Patstat* database and the scientific works indexed in the *Scopus* bibliographic database. This procedure consists of two phases: a broad generation of candidate couples and another phase of validation of couples, and it has been implemented with reasonably good results at a low cost.

Keywords

Citation; Quotes; Bibliographic references; Patents; Articles; Scientific production; Pairing; Databases; *Patstat*; *Scopus*; Methods; Methodology; Bibliometrics; Informetrics; Statistics; Analysis; Journals; Impact; Mapping; Name game.

Financing

This work has been funded by the *State Plan for Scientific and Technical Research and Innovation 2013-2016* and the *European Regional Development Fund (ERDF)* as part of the CSO2016-75031-R project.

1. Introduction

To their classic missions of teaching and research, the universities added the transfer of knowledge to the industry, what constituted their third mission (**Etzkowitz**; **Leydesdorff**, 2000). Since then, the demand for patent data has increased in academic work.

In this sense *Patstat* will become, or already is, a standard among researchers (**Kang**; **Tarasconi**, 2016). However, it is not a perfect database –it does not have a user-oriented interface; it has a European bias, it lacks standardization in the data of the applicants and inventors; the patent families are not clearly defined; and the classification is technological, lacking an industrial classification.

Due to *Patstat*'s orientation to patent applications and to the process of examining them, there is a need for debugging and normalizing the rest of the data. For example, the relationship of applicants and inventors with the data available in the company databases has been a problem for a long time due to the lack of standardization.

> Due to *Patstat*'s orientation to patent applications and to the process of examining them, there is a need for debugging and normalizing the rest of the data

The first attempts to normalize names were with the *Thomson Scientific*'s *Derwent World Patent Index* standardization tables (2002) and *United States Patent and Trademark Office*'s *Coname* file. Subsequently, a group of researchers from the *Katholieke Universiteit Leuven* (**Magerman**; **Van Looy**; **Song**, 2006) developed another method of normalizing names. **Thoma** and **Torrisi** (2007) elaborated an approximate matching method with the data of the Leuven and obtained a significant improvement of the completeness, although at the expense of the precision, for the *Crios*[1] database.

Raffo and **Lhuillery** (2009) studied a method of automatically recovering inventors in *Patstat*. However, due to the normalization problems the method is negatively referred to as the "Names game". Theis method established that a *name matching procedure* can be divided into three sequential phases:

- the parsing stage;
- the matching stage; and
- the filtering stage.

Lotti and **Marin** (2013) also made a matching system using the data found in *AIDA* (*Analisi Informatizzata delle Aziende*), a database marketed by *Bureau van Dijk* that includes data about Italian companies.

Coffano and **Tarasconi** (2014) carried out a cleanup and standardization of *Patstat* data, including the names of applicants and inventors, and they completed the information with other data in their *BD Patstat-Crios*.

Attempts have also been made to match inventors' names with university professors (**Lissoni**, 2012); his study proved that the message that European academic science does not contribute to technological advancement was incorrect. **Maraut** and **Martínez** (2014) developed a system to work specifically with Spanish names using natural language processing techniques. They establish four phases:

- text structuration;
- name matching;
- person disambiguation and clustering; and
- quality control and recursive validation.

Schoen, **Heinisch**, and **Buenstorf** (2014) played the "Names game" by applying it to a German case. In this case they established 5 phases:

- cleaning;
- professor-inventor name matching;
- inventor-inventor filtering;
- professor-inventor filtering; and
- manual control

It is clear that when a scientific advance is patented it is because it may be productive, both socially and economically. But not only technological progress is made when a patent is requested, that is, when a new product has been obtained. In fact, a large part of the patented inventions is based on scientific advances, often published in scientific journals. Patent documents include citations to previous patents and also to scientific articles (what is generically called non-patent literature, or NPL). In some countries the legislation requires that such citations be made by the applicant, while in others it requires the examiners do so.

> Because we determine scientific impact from scientific publications by analyzing citations, we can assume that we can determine technological impact by analyzing patent citations

Therefore, because we determine scientific impact from scientific publications by analyzing citations, we can assume that we can determine technological impact by analyzing patent citations.

In order to do this, it is necessary to identify which scientific publications correspond to the existing citations in the patents. At this point we encounter the same problem as in the case of the names of the applicants or the inventors: the lack of standardization. In this respect, fewer studies have been done. The only one that we have found has been in the development of *Lens influence mapping* (**Jefferson** *et al.*, 2018). With respect to the pairing with the papers, in their study it is only said that *PubMed* and *Crossref* are used, and it is not indicated how the cases in which more than one doi are retrieved are resolved, or the certainty that the retrieved document corresponds to the citation.

The aim of this paper is to present a methodology of matching the incomplete and unstructured references of the NPL (non-patent literature) section of *Patstat* with the references of the *Scopus* bibliographic database (2003-2017).

2. Data

Patstat (*EPO worldwide PAtent STATistical Database*) is a global patent database created by the *European Patent Office* (*EPO*), released for the first time in 2008 to assist patent statistical research at the request of a working group on patent statistics led by the *Organization for Economic Cooperation and Development* (*OECD*). Other members of this working group are: *World Intellectual Property Organization* (*WIPO*), *Japanese Patent Office* (*JPO*), *US Patent and Trademark Office* (*USPTO*), *Korean Intellectual Property Office* (*KIPO*), *US National Science Foundation* (*NSF*), and the *European Commission* (*EC*).

> The main *Patstat* advantages over other databases is its worldwide coverage, the inclusion of more information, and the existence of some auxiliary products that solve some problems

The main *Patstat* advantages over other databases such as *NBER* (from the United States) or *IIP* (from Japan) is its worldwide coverage, the inclusion of more information, and the existence of some auxiliary products that solve some problems, which has made it a *de facto* standard (**Kang**; **Tarasconi**, 2016). Its disadvantages are its orientation to Europe (data from national offices are exchanged with the *EPO* on the basis of agreements that change over time and may leave gaps) and its orientation to the examination process (data that are not necessary in the process of patent examination have a lower quality).

Patstat consists of 2 products:

- *Patstat global*: which has a worldwide coverage, and contains bibliographic information about applications and publications, as well as legal information about patents.
- *Patstat EP register*: which contains detailed bibliographic, procedural, and legal information on European and Euro-*PCT* (*Patent cooperation treaty*) patent applications.

Patstat is a relational database defined in the scheme of Figure 2. It can be used online or purchased on DVD to be installed on a local computer, and can be searched using SQL (**De-Rassenfosse**; **Dernis**; **Boedt**, 2014). The *EPO* publishes two

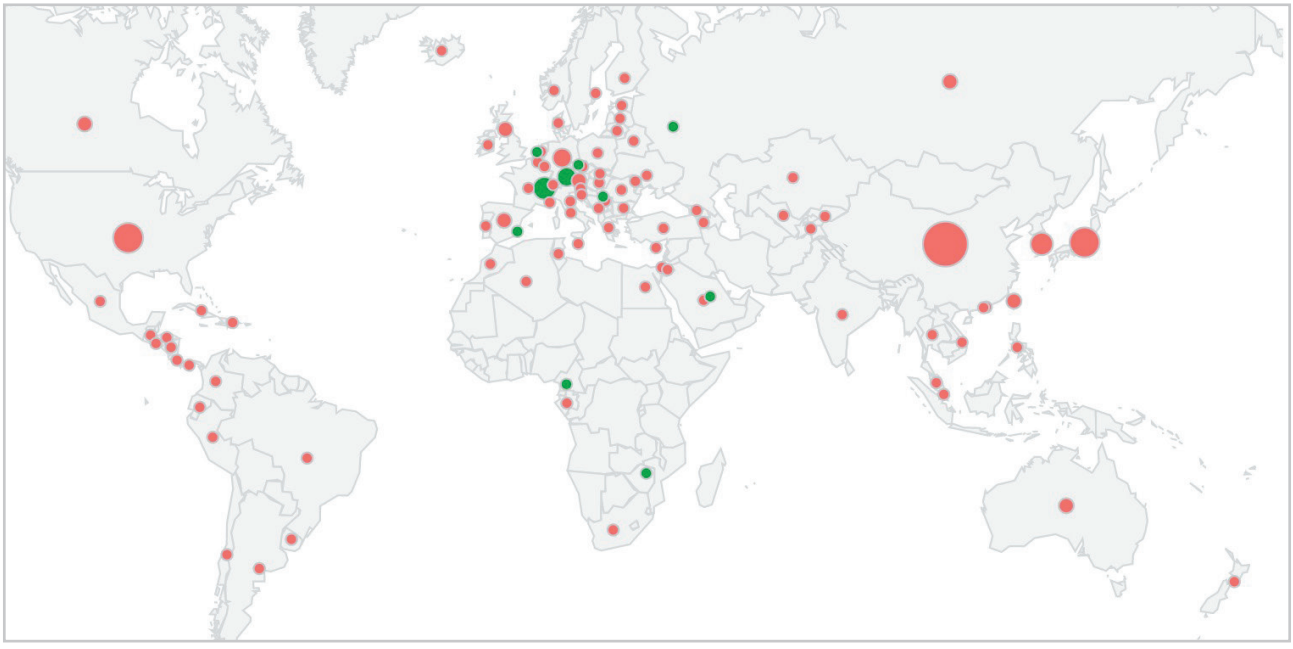


Figure 1. Number of applications submitted between 2003 and 2017 by each national office (in green the international and historical offices). Map based on *Patstat data* (2018).

editions a year of *Patstat*: *Spring* and *Autumn*. The *Spring* edition of 2018 (*Patstat - 2018 Spring edition*) is a snapshot of the data present in *Docdb EPO*, a global bibliographic database that includes data from more than 90 patent offices around the world, and *Inpadoc EPO*, a global database of the legal status, taken in the 5th week of 2018 (Figure 3). The data of people are taken from:

- *EP Patent register* for *EPO* applicants.
- *USPTO* for US data from patents published since 1976, and from patents applied as of November 29, 2005. The previous ones are taken from the *Docdb EPO*.

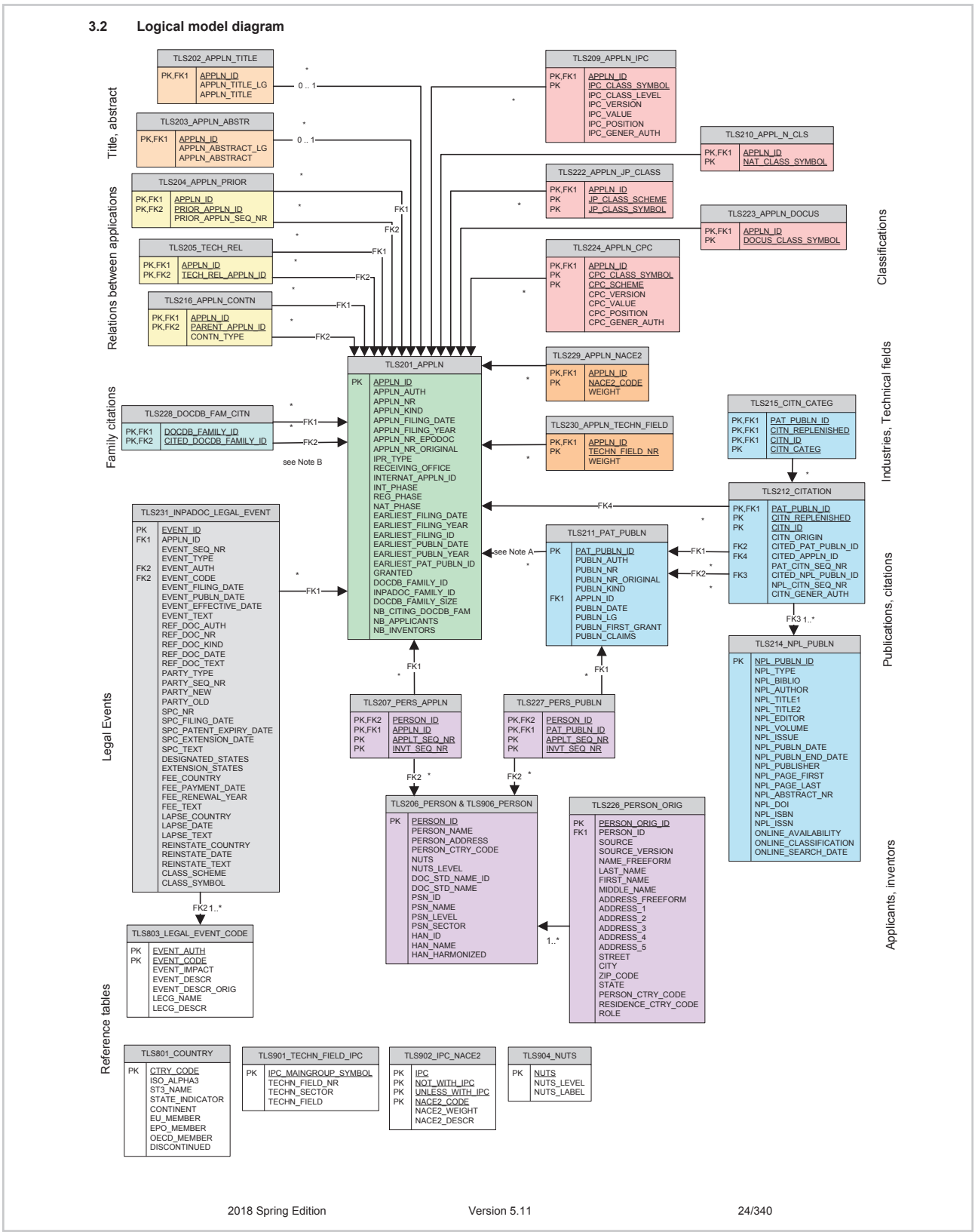


Figure 2. *Patstat - 2018 Spring edition* relational schema. Source: *European Patent Office* (2018)

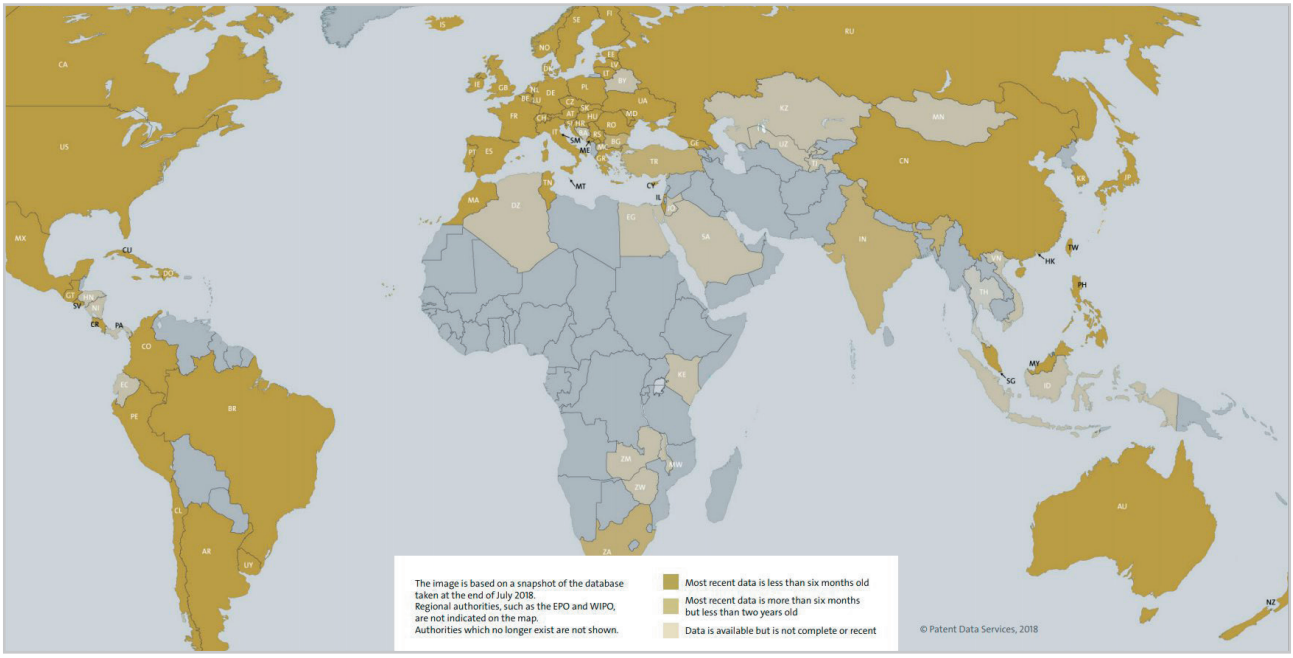


Figure 3. *Patstat* coverage. Source: *http://documents.epo.org/projects/babylon/eponet.nsf/0/73C531E61E437E8BC1258345005975AB/$File/Coverage_of_EPO_bibliographic_data_(DOCDB)_map_en.pdf*

The table TLS214_NPL_PUBLN is the one that includes the data of the NPL references. At first glance it may seem that it has a very rich structure, however, only the first three fields are complete in all registers:

- NPL_PUBLN_ID: Numeric key of the table.
- NPL_TYPE: Type of reference NPL (Table 1). The percentage of each type varies little from one edition to another.
- NPL_BIBLIO: Complete reference (as it appears in the patent, but that does not follow a fixed standard, nor is it complete).

Table 1. NPL_TYPE values, description and cardinality

NPL_TYPE	Description	No of NPL references	% of NPL references
a	Abstract citation of no specific kind	29,340,241	80.09
b	Book citation	748,515	2.04
c	*Chemical abstracts* citation	27,357	0.07
d	*Derwent* citation	118,033	0.32
e	Database citation	124,387	0.34
i	*Biological abstracts* citation	728	0.002
j	*Patent Abstracts of Japan* citation	392,819	1.07
s	Serial / Journal / Periodical citation	5,649,995	15.42
w	World Wide Web / Internet search citation	232,101	0.63

The rest of the fields (18) were incorporated in the *Spring 2017* version, and they are complete in a small percentage of cases (<20%); the percentage did not increase significantly in the *Spring 2018* version; additionally, the content is not always correct.

The percentage of fields that are completed depends on the type of reference (Table 2). Type "a" references are poor quality (*European Patent Office*, 2018) and have only the first three fields complete, accounting for 80% of the references. As we indicated previously, these percentages did not change significantly in the *Spring 2017 Patstat*. Nor does the percentage of filled fields for each reference type (Table 2) vary significantly from one edition to another.

There are many records in this table that are repeated, around a third, with the key NPL_PUBLN_ID varying exclusively, and this key is not maintained from one version to another, so the only way to relate them is through the field NPL_BIBLIO.

Scopus is a bibliographic database of *Elsevier* (**Hanne**, 2004; **Pickering**, 2004), which indexes 23,700 scientific journals. Although it is not the longest on the market, several studies have tried to characterize it (**Archambault** *et al.*, 2009; **Leydesdorff** *et al.*, 2010; **De-Moya Anegón** *et al.*, 2007), and it has been used in several scientometric studies (**Gorraiz**; **Gumpenberger**; **Wieland**, 2011; **Jacsó**, 2011; **Guerrero-Bote**; **De-Moya-Anegón**, 2015; **De-Moya-Anegón** *et al.*, 2018).

In *Scopus* the documents are classified by *Subject areas* and by *Specific subject areas* or *Categories*. There are more than 300 *Specific subject areas* that are grouped into 26 *Subject areas*. In addition, there is the *General subject area* that contains multidisciplinary journals such as *Nature* or *Science*.

Table 2. Filled fields of the table TLS214_NPL_PUBLN in *Patstat - 2018 Spring edition*

		Poor citations	Articles					Online		
Attributes		a	b	c	i	j	s	d	e	w
Amounts in thousands		29,340	749	27	1	393	5,650	118	124	232
NPL_BIBLIO	%, rounded	100	100	100	100	100	100	100	100	100
NPL_AUTHOR			2	66	81		95	2	54	85
NPL_TITLE1			24	67	82		61	5	72	95
NPL_TITLE2			100	100	100	100	100			66
NPL_EDITOR			78							
NPL_VOLUME			11	92	80	98	76	90		32
NPL_ISSUE				88	23	98	37	90		28
NPL_PUBLN_DATE			93	91	56	97	89	4	62	95
NPL_PUBLN_END_DATE										2
NPL_PUBLISHER			60						99	
NPL_PAGE_FIRST			30				80			54
NPL_PAGE_LAST			17				69			49
NPL_ABSTRACT_NR				96	95	59		99	82	
NPL_DOI							6			16
NPL_ISBN			3				2			1
NPL_ISSN			1				9			22
ONLINE_AVAILABILITY									38	77
ONLINE_CLASSIFICATION								51		
ONLINE_SEARCH_DATE										82

3. Methodology

Although the structured information present in 20% of the records must be used, it is also necessary to use the standardized textual reference (NPL_BIBLIO). Within it one can look for some patterns, to locate, for example, the year or the DOI. For all this, and following the contributions of the "*Names game*" (**Raffo**; **Lhuillery**, 2009) we have designed a procedure that is divided into four phases:

1. Preprocessing of data: Preparation of data to facilitate and streamline subsequent processes.

2. Pre-selection of candidate couples: From some coincidences of the elements of the references, pairs are preselected (NPL from *Patstat*, *Scopus* reference) candidates for the match.

3. Automatic evaluation of the candidate couples:

- The matching elements of each candidate pair are evaluated.
- A score is assigned to every couple; for each NPL we get an ordered list of the *Scopus* references that could fit.
- The overall score is the product of the scores obtained by each element of the reference (as a way of probability).

4. Human validation:

- From the top, from a certain score, the couples with the highest score can be validated.
- The lowest score can be discarded.
- For each NPL, only the *Scopus* reference with the highest score can be considered a fit (although there are also duplicate records in *Scopus*).

3.1. Preprocessing the data

This first phase includes preparing the data for the subsequent process. Much of this preprocessing is about solving some of the data normalization problems. It is carried out through SQL queries with the following steps:

- Unify the records: As indicated above, approximately one third of the tuples in the table are repeated. To reduce the IT burden, the first thing to do is unify the records by generating a new key.
- Identify the records evaluated in some previous phase: In this case we worked with the previous edition, so as to avoid starting from scratch the first task was to identify the records of previous editions already processed. In each edition of *Patstat*, the primary key of the table TLS214_NPL_PUBLN (NPL_PUBLN_ID) changes, so that the identification is made

by the field NPL_BIBLIO that contains the complete reference. This will also be necessary in later years.

- Assign a new numerical key that allows us to make slices in the table: In some of the processes that are carried out it is convenient to divide the table into equal parts. The fastest way to do this is to create a numeric key, where the already assigned records can be easily located (for example, assigning them a key from a certain number).
- Locate patterns corresponding to DOIs: It is possible to design regular expressions to locate DOIs. If a DOI is located in the reference the problem is solved, however, the number of references that include the DOI is very small.
- Assign years of publication: We can also look for patterns that match the 2003 to 2017 figures that correspond to the *Scopus* reference period with which we will match them. Since a reference can include several similar figures, in case the NPL_PUBLN_DATE field contains a correct value, this will be used. Otherwise we have to take into account that there will be references that contain more than one year and others that do not contain any years.
- All the textual fields are normalized both in the *Patstat* TLS214_NPL_PUBLN table and in the *Scopus* references, eliminating the special characters and reducing all the words to the root. In this way, the different lexical variants of a word are unified.
- Locate the texts in quotes, as candidates to be titles. These texts are stored, standardized, and reduced to the root.
- Extract the first word of the NPL_AUTHOR field or, failing that, from the NPL_BIBLIO field as a candidate to be the last name of the first author of the paper. Some exceptions are eliminated (van, der, von, etc.). The surname of the first author of the *Scopus* reference is also extracted. Both are stored once normalized and reduced to the root.
- Generate an inverted index with the roots extracted from the NPL_BIBLIO field, another with those extracted from the *Scopus* magazine titles, and another with the titles of the *Scopus* references.
- Try to assign to each reference in table TLS214_NPL_PUBLN one of the 23,000 *Scopus* journals. For this, the ISSN, the NPL_TITLE2 field, the title and the abbreviated title of the journals in *Scopus* are used in order of priority. In case textual comparisons are necessary, the inverted indexes are used to avoid a brute force comparison. For each assignment it is noted:
    - How the pairing has been done (if with the ISSN, with the title, with the abbreviated title, reduced to the root, etc.).
    - The number of characters of the match (the coincidence of three characters of an abbreviated title is not the same as forty of a raw title).

After performing the preprocess we have 24,046,625 records of the TLS214_NPL_PUBLN table without repeating, of the 36,634,177 that were originally in the table TLS214_NPL_PUBLN. Of them 2,604,437 we had them paired by the same procedure of the *Spring edition of 2017*.

Likewise, we have 37,792,849 *Scopus* references from the period 2003-2017 of all the document types present in *Scopus*.

3.2. Preselection of candidate couples

With the above data we can deduce that there are $9x10^{14}$ possible pairs formed by an NPL reference from *Patstat* and a *Scopus* reference. Due to the lack of normalization a direct comparison is necessary that is impossible to approach manually with such a large number of couples.

For that reason, this phase aims to reduce the number of couples to a more manageable number, but at the same time is large enough to minimize the possibility of a real couple being left out.

To that end, a series of rules are used that are applied in the form of SQL statements on the data obtained from the previous phase. The rules used are those corresponding to the following coincidences:

- DOI
- Journal, volume (NPL_VOLUME) and first page (NPL_PAGE_FIRST)
- Journal, volume (NPL_VOLUME) and number (NPL_ISSUE)
- Journal and last name of the first author
- Journal, volume (included in NPL_BIBLIO) and first page (included in NPL_BIBLIO)
- Journal, volume (included in NPL_BIBLIO) and number (included in NPL_BIBLIO)
- Journal, volume (included in NPL_BIBLIO) and last name of the first author
- Journal, year and first page (included in NPL_BIBLIO)
- Journal, year and last page (included in NPL_BIBLIO)
- Journal, first page (included in NPL_BIBLIO) and last page (included in NPL_BIBLIO)
- First author and first page (NPL_PAGE_FIRST)
- First author and last page (NPL_PAGE_LAST)
- First author, first page (included in NPL_BIBLIO) and last page (included in NPL_BIBLIO)
- Journal, first page (NPL_PAGE_FIRST) and last page (NPL_PAGE_LAST)
- Title of the paper reduced to the root (NPL_Title1)
- Title of the paper reduced to the root (first quotation mark of NPL_BIBLIO)

- Title of the paper reduced to the root (second quotation mark of NPL_BIBLIO)
- Title of the paper reduced to the root (third quoted from NPL_BIBLIO)
- Inclusion in NPL_BIBLIO of the year and two of the 4 least frequent roots of the *Scopus* reference title
- Inclusion in NPL_BIBLIO of the year and a term of the title of the *Scopus* reference that appears in less than 1,000 NPL references
- Surname of the first author (the candidate word to be the normalized surname that was extracted in the previous phase) and two of the 4 least frequent roots of the reference title of *Scopus*
- Surname of the first author (the candidate word to be the normalized surname that was extracted in the previous phase) and a term of the title of the *Scopus* reference that appears in less than 1,000 NPL references

As one can see, the specific fields of the TLS214_NPL_PUBLN table were used whenever possible, but since they are not filled in a large percentage, the corresponding terms or numbers were also searched in the textual reference (NPL_BIBLIO).

Candidate pairs have been generated for 14,758,096 NPL references of the 24 million we started from. Many references do not link to NPLs, while others point to bibliography not covered in *Scopus* or that was published outside the period studied. The pre-selection procedure generates 2,280,503,246 candidate pairs, once those that correspond to those assigned in the *Spring edition of 2017* have been eliminated. This means that in many cases the same NPL reference has many candidate *Scopus* references. There are 307,901 NPL references that each have more than 1,000 candidate *Scopus* references, while there are only 1,389,571 NPL references that have a single *Scopus* candidate reference. The distribution is power law, as can be seen in Figure 4.

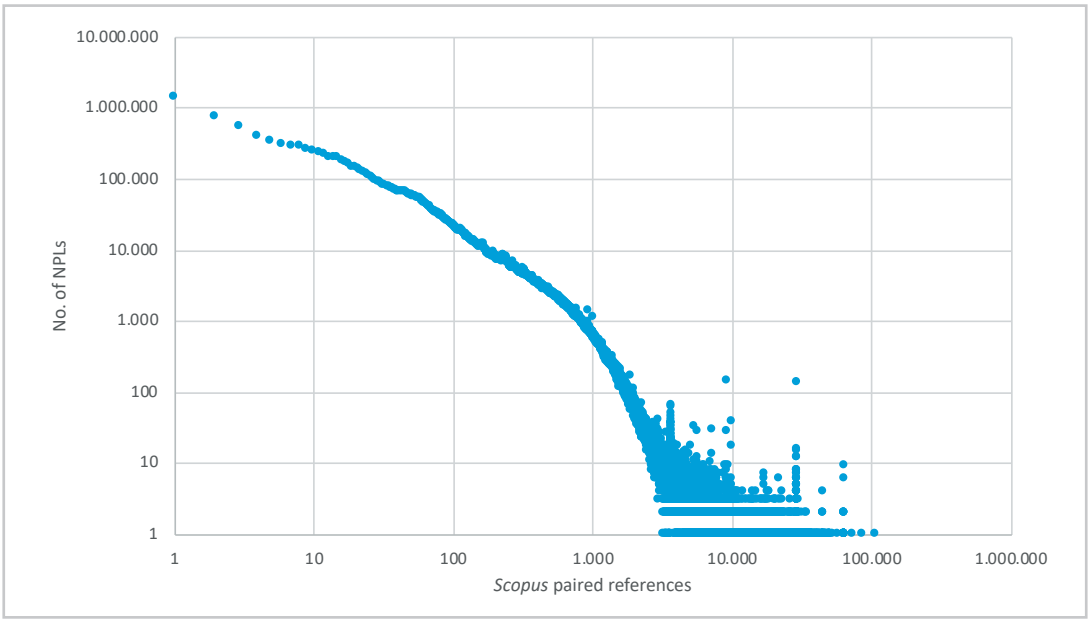


Figure 4. Scatter chart showing the number of *Scopus* references paired with each NPL reference

3.3. Automatic evaluation of candidate couples

The objective of this phase is to assign a score that allows selecting for each NPL reference the *Scopus* reference that is most likely to refer to the same document. To this end, a series of routines have been designed that look for the most important elements of the *Scopus* reference in the table record TLS214_NPL_PUBLN. The elements sought, for each of which an independent routine has been designed, are the following:

- Year of publication
- Last name of the first author
- Document title
- Journal
- Volume
- Issue
- Pages

Depending on the quality and the importance of the match, each routine assigns a score:

- In the event that an element is not contained, it is assigned a value less than one, except in the case of the title or first author, which is assigned one (some NPL references do not contain the title or first author, but are fully specified).
- The score is a function of the matching size, although it is multiplied by a factor based on the quality of the fit (this factor is greater if the fit is without reduction to the root, or in the specific fields of the TLS214_NPL_PUBLN table).
- The total matching score for each candidate pair is obtained by multiplying the value assigned by all the routines.

3.4. Human validation

As seen in the previous sections, an NPL reference may not have any *Scopus* candidate reference, it may have one or it may have several. Logically, if any of the candidates corresponds to the NPL reference, this should be the highest score obtained, but none of the assigned ones may be valid. For this reason, a manual validation is necessary.

To this end, an application has been developed that facilitates cooperation between many people in human validation (Figure 5).

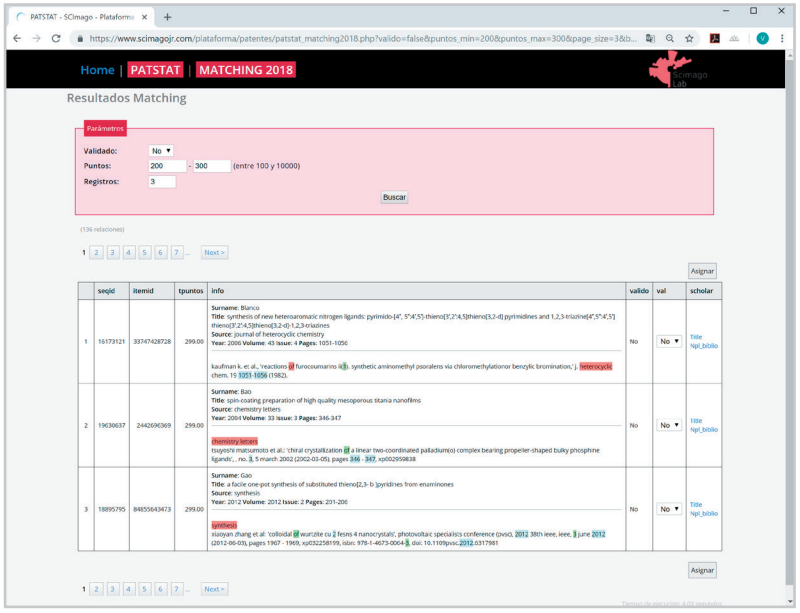


Figure 5. Screenshot of the application that allows to manually validate the pairings between NPL references of *Patstat* and *Scopus*.

4. Results

Table 3 shows the results of the pairing process after human validation, with the corresponding error percentages in each interval. These error percentages (up to 1,000 points) are absolute, and in the rest of the intervals a sampling of 100 pairs has been made. References with more than 10,000 points are incorporated automatically without the need for human validation. As shown, the success is 100% for these pairings. Finally, the references of the *Spring 2018* edition should be added to the 2,604,437 that obtained 10,000 or more points with the *Spring 2017* version.

Table 3. Distribution of the NPL references by scoring intervals received with the *Scopus* reference that best fits, and percentage of error

Min. points	Max. points	References NPL	% error	Errors	Success	% accumulated success	% processed references
0	100	12,139,055	99.60	12,090,499	48,556	100	100
100	150	901,443	94.00	847,356	54,087	98.74	30.08
150	200	312,131	67.68	211,250	100,881	97.34	24.89
200	250	215,041	76.00	163,431	51,610	94.72	23.10
250	300	119,397	60.00	71,638	47,759	93.38	21.86
300	350	81,606	20.00	16,321	65,285	92.14	21.17
350	400	65,211	35.00	22,824	42,387	90.45	20.70
400	450	57,945	36.00	20,860	37,085	89.35	20.32
450	500	43,734	21.82	9,542	34,192	88.38	19.99
500	600	75,555	19.05	14,393	61,162	87.50	19.74
600	700	50,646	26.67	13,507	37,139	85.91	19.30
700	800	37,539	27.27	10,237	27,302	84.95	19.01
800	900	31,509	8.54	2,691	28,818	84.24	18.79
900	1,000	25,571	11.62	2,971	22,600	83.49	18.61
1.000	2,000	144,735	7.34	10,622	134,113	82.90	18.47
2.000	3,000	26,663	2.20	587	26,076	79.42	17.63
3,000	4,000	17,620	0.81	142	17,478	78.75	17.48
4,000	5,000	11,133	0.24	27	11,106	78.29	17.38
5,000	6,000	9,435	0.10	9	9,426	78.00	17.31
6,000	7,000	8,403	0.00	0	8,403	77.76	17.26
7,000	8,000	7,227	0.01	1	7,226	77.54	17.21
8,000	9,000	6,451	0.00	0	6,451	77.35	17.17
9,000	10,000	6,156	0.08	5	6,151	77.19	17.13
10,000	-	363,890	0.00	0	363,890	77.03	17.10
2017		2,604,437		0	2,604,437	67.58	15.00
Total		17,362,533		13,508,915	3,853,618		

In Table 3 we can see four groups of references processed with their minimum and maximum points. The two lower groups (in blue and green) have been completely processed and for them the percentage of successfully matched NPL references is very high (99.6% of the processed references). The size of these two groups is 3,206,150 references, which constitutes about 18.5% of the total references, but we estimate that they represent around 83% of the references to which an article can be assigned. The group of references with minimum scores between 1,000 and 10,000 (in green) has been validated manually and constitutes only 1.4% of the references, which includes 5.9% of the references to which an article can be assigned. In this group of references the success percentage is very high (95.2%), which greatly reduces the human effort required.

The group of references with between 300 and 1,000 points still shows moderately low error rates (on average less than 25%) and includes 9.2% of the correctly matched references in a volume of data equivalent to 2.7% of the total. Although the relationship between references to process and correctly matched references is good, the net effort to be made is still high, so it is in this region where there is a greater margin for improvement in automatic procedures. Nevertheless, the last group of references offers a very uninspiring balance between the effort that must be made (processing about 79% of the data) and the expected advantage (7.9% of the matching references correctly).

The total number of references incorporated from the new version of the database by this procedure was 590,410, these were then added to those detected by the same procedure in the previous edition, with an aggregate total of about 3.2 million references to documents indexed in *Scopus*. In the near future it is estimated that around 350,000 more references will be linked manually by checking the references with between 300 and 1,000 points.

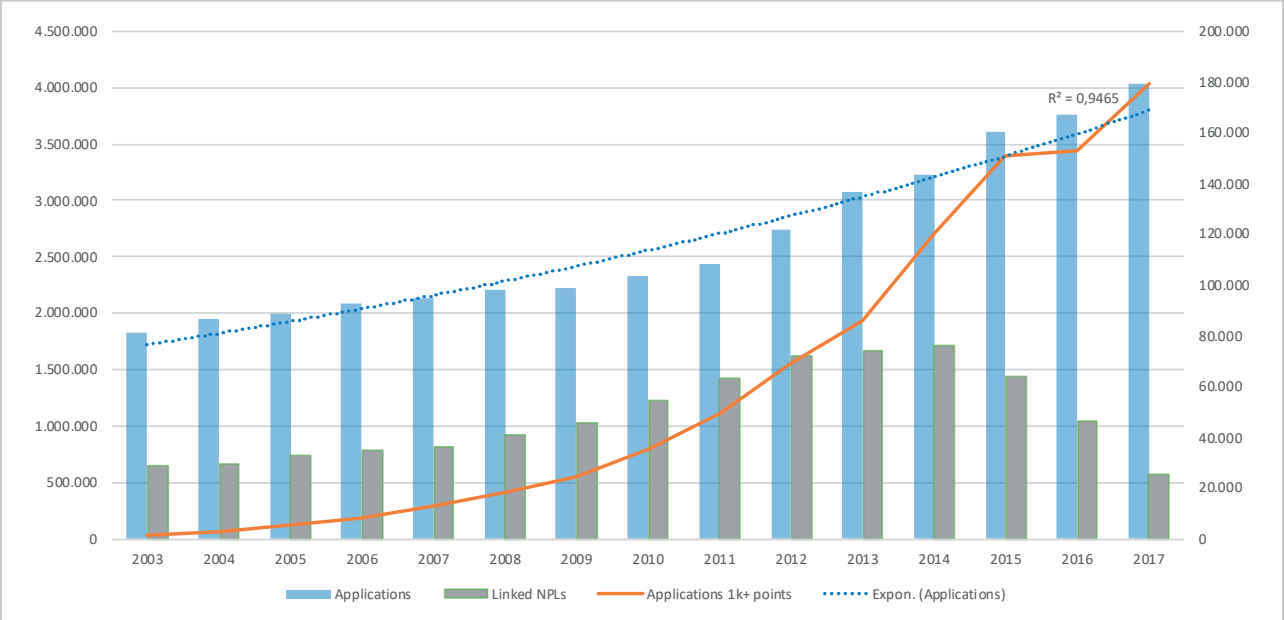


Figure 6. Evolution of the number of applications, linked NPL and linked NPL with more than 1,000 points (secondary axis on the right).

As can be seen in Figure 6, the number of applications has increased exponentially over the last few years, but the growth in the number of NPL references linked with more than 1,000 points has been even more pronounced. It seems reasonable to think that the strong rise in linked references implies that data is becoming increasingly robust. This figure also describes the rhythm of incorporation of NPL references. After the first publication date, other publications will be made in many cases that will enrich the bibliography cited in the patents, a bibliography that is not yet available in the patent data published for the first time in recent years.

5. Conclusions

The percentage of correctly matched references can still be improved with respect to the total available references. We believe that this situation will evolve positively in future editions of *Patstat*, since the above-mentioned data indicate a visible improvement in the scores of the pairings in recent years. This trend will most likely continue, therefore, the percentage of correctly matched references will increase. However, it is necessary to bear in mind that of the 24 million references processed, there are 6.7 million references that have no relation with any *Scopus* record.

On the other hand, there is room for improvement in the method we have used to carry out the pairings. The difference between the number of references with 10,000 or more points and the maximum number that we can theoretically match with this procedure is still important. Continued work to improve the automatic

> The automatic evaluation phase can be used to process pairings that until now were not reviewed and about which we only had an estimate of their quality

evaluation phase should result in an effective increase in the number of well-matched references. This would reduce the human effort required for validation. The automatic evaluation phase can be used to process pairings that until now were not reviewed and about which we only had an estimate of their quality.

6. Note

1. *Crios-Patstat* is a patent database created by a team of researchers from *Centro di Ricerca su Innovazione, Organizzazione e Strategia* (*Crios*), of *Università Bocconi*, in Milan.

In this database the user can find, for applications of the *European Patent Office*, names of disambiguated inventors and applicants, as well as other data that are often difficult to find in other patent databases.

7. References

Archambault, Éric; **Campbell, David**; **Gingras, Yves**; **Larivière, Vincent** (2009). "Comparing bibliometric statistics obtained from the Web of Science and Scopus". *Journal of the American Society for Information Science and Technology* (*Jasist*), v. 60, n. 7, pp. 1320-1326.
https://doi.org/10.1002/asi.21062

Coffano, Monica; **Tarasconi, Gianluca** (2014). *Crios - Patstat database: Sources, contents and access rules*. Center for Research on Innovation, Organization and Strategy, Crios Working Paper n. 1.
https://ssrn.com/abstract=2404344
https://doi.org/10.2139/ssrn.2404344

De-Moya-Anegón, Félix; **Chinchilla-Rodríguez, Zaida**; **Vargas-Quesada, Benjamín**; **Corera-Álvarez, Elena**; **Muñoz-Fernández, Francisco-José**; **González-Molina, Antonio**; **Herrero-Solana, Víctor** (2007). "Coverage analysis of Scopus: A journal metric approach". *Scientometrics*, v. 73, n. 1, pp. 53-78.
https://doi.org/10.1007/s11192-007-1681-4

De-Moya-Anegón, Félix; **Guerrero-Bote, Vicente P.**; **López-Illescas, Carmen**; **Moed, Henk F.** (2018). "Statistical relationships between corresponding authorship, international co-authorship and citation impact of national research systems". *Journal of informetrics*, v. 12, n. 4, pp. 1251-1262.
https://doi.org/10.1016/j.joi.2018.10.004

De-Rassenfosse, Gaétan; **Dernis, Hélène**; **Boedt, Geert** (2014). "An introduction to the Patstat database with example queries". *Australian economic review*, v. 47, n. 3, pp. 395-408.
https://doi.org/10.1111/1467-8462.12073

Derwent (2000). *World Patents Index - Derwent patentee codes, Revised edition 8*. Thomson Corporation. Leuven Manual. ISBN: 0 901157 38 4
http://ips.clarivate.com/m/pdfs/mgr/patenteecodes.pdf

Etzkowitz, Henry; **Leydesdorff, Loet** (2000). "The dynamics of innovation: from National Systems and 'Mode 2' to a Triple Helix of university–industry–government relations". *Research policy*, v. 29, n. 2, pp. 109-123.
https://doi.org/10.1016/S0048-7333(99)00055-4

European Patent Office (2018). *Data catalog Patstat global*. Versión 5.11. EPO Patstat customers.
https://www.epo.org

Gorraiz, Juan; **Gumpenberger, Christian**; **Wieland, Martin** (2011). "Galton 2011 revisited: a bibliometric journey in the footprints of a universal genius". *Scientometrics*, v. 88, n. 2, pp. 627-652.
https://doi.org/10.1007/s11192-011-0393-y

Guerrero-Bote, Vicente P.; **De-Moya-Anegón, Félix** (2015). "Analysis of scientific production in food science from 2003 to 2013". *Journal of food science*, v. 80, n. 12, R2619-R2626.
https://doi.org/10.1111/1750-3841.13108

Hane, Paula J. (2004). "Elsevier announces Scopus service". *Information today*. *http://newsbreaks.infotoday.com/nbreader.asp?ArticleID=16494*

Jacsó, Péter (2011). "The h-index, h-core citation rate and the bibliometric profile of the Scopus database". *Online information review*, v. 35, n. 3, pp. 492-501.
https://doi.org/10.1108/14684521111151487

Jefferson, Osmat A.; **Jaffe, Adam**; **Ashton, Doug**; **Warren, Ben**; **Koellhofer, Deniz**; **Dulleck, Uwe**; **Bilder, G.**; **Ballagh, Aaron**; **Moe, John**; **DiCuccio, Michael**; **Ward, Karl**; **Bilder, Geoff**; **Dolby, Kevin**; **Jefferson, Richard A.** (2018). "Mapping the global influence of published research on industry and innovation". *Nature biotechnology*, v. 36, n. 1, pp. 31-39.
https://doi.org/10.1038/nbt0818-772a

Kang, Byeongwoo; **Tarasconi, Gianluca** (2016). "*Patstat* revisited: Suggestions for better usage". *World patent information*, v. 46, pp. 56-63.
https://doi.org/10.1016/j.wpi.2016.06.001

Leydesdorff, Loet; **De-Moya Anegón, Félix**; **Guerrero-Bote, Vicente P.** (2010). "Journal maps on the basis of Scopus data: A comparison with the Journal Citation Reports of the ISI". *Journal of the American Society for Information Science and Technology*, v. 61, n. 2, pp. 352-369.
https://doi.org/10.1002/asi.21250

Lissoni, Francesco (2012). "Academic patenting in Europe: an overview of recent research and new perspectives". *World patent information*, v. 34, n. 3, pp. 197-205.
https://doi.org/10.1016/j.wpi.2012.03.002

Lotti, Francesca; **Marin, Giovanni** (2013). "Matching of Patstat applications to AIDA firms: Discussion of the methodology and results". *Bank of Italy occasional paper*, n. 166.
https://ssrn.com/abstract=2283111 https://doi.org/10.2139/ssrn.2283111

Magerman, Tom; **Van-Looy, Bart**; **Song, Xiaoyan** (2006). *Data production methods for harmonized patent statistics: Patentee name standardization*. Technical report, K.U. Leuven.
https://ec.europa.eu/eurostat/documents/3888793/5836029/KS-AV-06-002-EN.PDF

Maraut, Stéphane; **Martínez, Catalina** (2014). "Identifying author–inventors from Spain: methods and a first insight into results". *Scientometrics*, v. 101, n. 1, pp. 445-476.
https://doi.org/10.1007/s11192-014-1409-1

Pickering, Bobby (2004). "Elsevier prepares Scopus to rival ISI Web of science". *Information world review*, n. 8.

Raffo, Julio D.; **Lhuillery, Stéphane** (2009). "How to play the '*Names game*': Patent retrieval comparing different heuristics". *Research policy*, v. 38, n. 10, pp. 1617-1627.
https://doi.org/10.2139/ssrn.1441172

Schoen, Anja; **Heinisch, Dominik**; **Buenstorf, Guido** (2014). "Playing the 'Name game' to identify academic patents in Germany". *Scientometrics*, v. 101, n. 1, pp. 527-545.
https://doi.org/10.1007/s11192-014-1400-x

Thoma, Grid; **Torrisi, Salvatore** (2007). *Creating powerful indicators for innovation studies with approximate matching algorithms. A test based on Patstat and Amadeus databases* (No. 211). KITeS, Centre for Knowledge, Internationalization and Technology Studies, Università Bocconi, Milano, Italy.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.573.8107&rep=rep1&type=pdf