Revisión de la citación de las patentes a la producción científica: un nuevo enfoque para el emparejamiento *Patstat / Scopus*

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

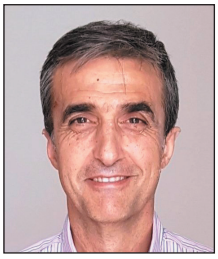
Note: This article can be read in English on:
http://www.elprofesionaldelainformacion.com/contenidos/2019/jul/guerrero-sanchez-de-moya.pdf

Cómo citar este artículo:

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

Artículo recibido el 03-05-2019
Aceptación definitiva: 17-05-2019



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

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

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

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

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

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

Resumen

Las patentes incluyen citas, tanto a otras patentes como a documentos que no son patentes (NPL, *Non-patent literature*). Entre estas últimas se incluyen citas a artículos publicados en revistas científicas. Igual que se estudia el impacto científico a través la citación de artículos y otros trabajos científicos, también se puede estudiar el impacto tecnológico de los trabajos científicos a través de la citación que reciben de las patentes. Las referencias NPL incluidas en las patentes están lejos de estar normalizadas, por lo que determinar a qué artículo científico se refieren no es trivial. En este trabajo se presenta un procedimiento de enlazado de las referencias NPL de las patentes recogidas en la base de datos *Patstat* y los trabajos científicos indexados en la base de datos bibliográfica *Scopus*. Dicho procedimiento se compone de dos fases: una generación amplia de parejas candidatas y otra fase de validación de las parejas. Ha sido implementado con resultados razonables y costes asumibles.

Financiación

Este trabajo ha sido financiado por el *Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016* y el *Fondo Europeo de Desarrollo Regional (Feder)* como parte del proyecto CSO2016-75031-R.

Palabras clave

Citación; Citas; Referencias bibliográficas; Patentes; Artículos; Producción científica; Emparejamiento; Bases de datos; *Patstat*; *Scopus*; Métodos; Metodología; Bibliometría; Informetría; Estadísticas; Análisis; Revistas; Impacto; Mapeado; *Name game*.

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.

1. Introducción

A sus misiones clásicas de enseñar e investigar, las universidades añadieron la transferencia de conocimientos a la industria, lo que constituyó su tercera misión (**Etzkowitz**; **Leydesdorff**, 2000). Desde entonces la demanda de los datos de las patentes se ha incrementado en los trabajos académicos.

En ese sentido *Patstat* ha llegado o llegará a ser un estándar entre los investigadores (**Kang**; **Tarasconi**, 2016), aunque no es una base de datos perfecta puesto que no tiene una interfaz orientada al usuario, tiene un sesgo europeo, falta normalización en los datos de los solicitantes e inventores, las familias de patentes no están definidas claramente, y la clasificación es tecnológica, echándose en falta una clasificación industrial.

> Debido a la orientación de *Patstat* a la solicitud de patentes y al proceso examinador de la misma, falta depuración y normalización en el resto de los datos

Debido a su orientación a la solicitud de patentes y al proceso examinador de la misma, falta depuración y normalización en el resto de los datos. Por ejemplo, la relación de los solicitantes e inventores con los datos disponibles en las bases de datos de empresas ha sido un problema desde hace mucho tiempo por la falta de normalización. Los primeros intentos de normalización de nombres fueron las tablas de normalización *Derwent World Patent Index de Thomson Scientific* (2002) y el fichero *Coname* de la *United States Patent and Trademark Office*. Posteriormente un grupo de investigadores de la *Katholieke Universiteit Leuven* (**Magerman**; **Van Looy**; **Song**, 2006) desarrolló otro método de normalización de los nombres. **Thoma** y **Torrisi** (2007) elaboraron un método de emparejamiento aproximado con los datos de la *Leuven* obteniendo una mejora importante de la exhaustividad, aunque a costa de la precisión, para la base de datos *Crios*[1].

Raffo y **Lhuillery** (2009) estudiaron un método de recuperar automáticamente inventores en *Patstat*, que debido a los problemas de normalización que tiene lo llamaron "*Names game*". Establecieron que genéricamente un *name matching procedure* se puede dividir en tres fases secuenciales:

- *the parsing stage*;
- *the matching stage*; y
- *the filtering stage*.

Lotti y **Marin** (2013) también realizaron un matching con los datos de *AIDA* (*Analisi Informatizzata delle Aziende*), base datos comercializada por *Bureau van Dijk* con los datos de las empresas italianas.

Coffano y **Tarasconi** (2014) llevaron a cabo una limpieza y normalización de los datos de *Patstat*, sobre todo de los nombres de solicitantes e inventores, a la vez que completaron con otros datos en su *BD Patstat-Crios*.

También se ha intentado emparejar los nombres de los inventores con los profesores de la universidad (**Lissoni**, 2012) que demostró que era incorrecto el mensaje de que la ciencia académica europea no contribuye al avance tecnológico. **Maraut** y **Martínez** (2014) elaboraron un sistema específico para trabajar con los nombres españoles utilizando técnicas de proceso del lenguaje natural. Ellos establecen cuatro fases:

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

Schoen, **Heinisch** y **Buenstorf** (2014) juegan al "*Names game*" aplicándolo al caso alemán. En este caso establecen 5 fases:

- *cleaning*;
- *professor-inventor name matching*;
- *inventor-inventor filtering*;
- *professor-inventor filtering*; y
- *manual control*.

> Si para determinar el impacto científico de una publicación científica utilizamos las citas recibidas en revistas científicas, para determinar el impacto tecnológico se podrán utilizar las citas recibidas en patentes

Está claro que cuando se patenta un avance es porque se considera que éste puede ser productivo, tanto a nivel social como económico. Pero no sólo se contribuye al avance tecnológico cuando se solicita una patente, es decir cuando se ha obtenido un producto nuevo. En realidad, gran parte de los inventos patentados se basan en avances científicos, muchas veces publicados en revistas científicas. En los documentos de las patentes se incluyen citas a patentes anteriores y también a artículos científicos (lo que genéricamente se denomina bibliografía no-patente, o *non-patent literature*, NPL). En algunos países la legislación exige que dichas citas sean introducidas por el solicitante, mientras que en otros exige que lo hagan los examinadores.

De modo que, si para determinar el impacto científico de una publicación científica utilizamos las citas recibidas en revistas científicas, para determinar el impacto tecnológico se podrán utilizar las citas recibidas en patentes.

Para ello es necesario identificar a qué publicaciones científicas corresponden las citas existentes en las patentes. En este punto nos encontramos con el mismo problema que en el caso de los nombres de los solicitantes o los inventores: la falta de normalización. A este respecto se han hecho menos estudios. El único que hemos encontrado ha sido en el desarrollo de *Lens influence mapping* (**Jefferson** *et al.*, 2018) donde respecto al emparejamiento con los *papers* solamente dice que se utilizan *PubMed* y *Crossref*, y no se indica cómo se resuelven los casos en los que se recupera más de un doi, o la seguridad que se tiene de que el documento recuperado corresponda a la cita.

El objetivo de este trabajo es presentar una metodología de emparejamiento de las referencias incompletas y no estructuradas de la sección NPL (*non-patent literature*) de *Patstat* con las referencias de la base de datos bibliográfica *Scopus* (2003-2017).

2. Datos

> Las principales ventajas de *Patstat* sobre otras bases de datos son la cobertura mundial, la inclusión de más información, y la existencia de algunos productos auxiliares que solventan algunos problemas

Patstat (*EPO worldwide PATent STATistical Database*) es una base de datos mundial de patentes creada por la *Oficina Europea de Patentes* (*EPO*), liberada por primera vez en 2008 para ayudar a la investigación estadística de patentes a petición de un grupo de trabajo de estadística de patentes liderado por la *Organización para la Cooperación y el Desarrollo Económicos* (*OCDE*). Otros miembros de dicho grupo de trabajo son: *World Intellectual Property Organisation* (*WIPO*), *Japanese Patent Office* (*JPO*), *US Patent and Trademark Office* (*USPTO*), *Korean Intellectual Property Office* (*KIPO*), *US National Science Foundation* (*NSF*) y la *Comisión Europea* (CE).

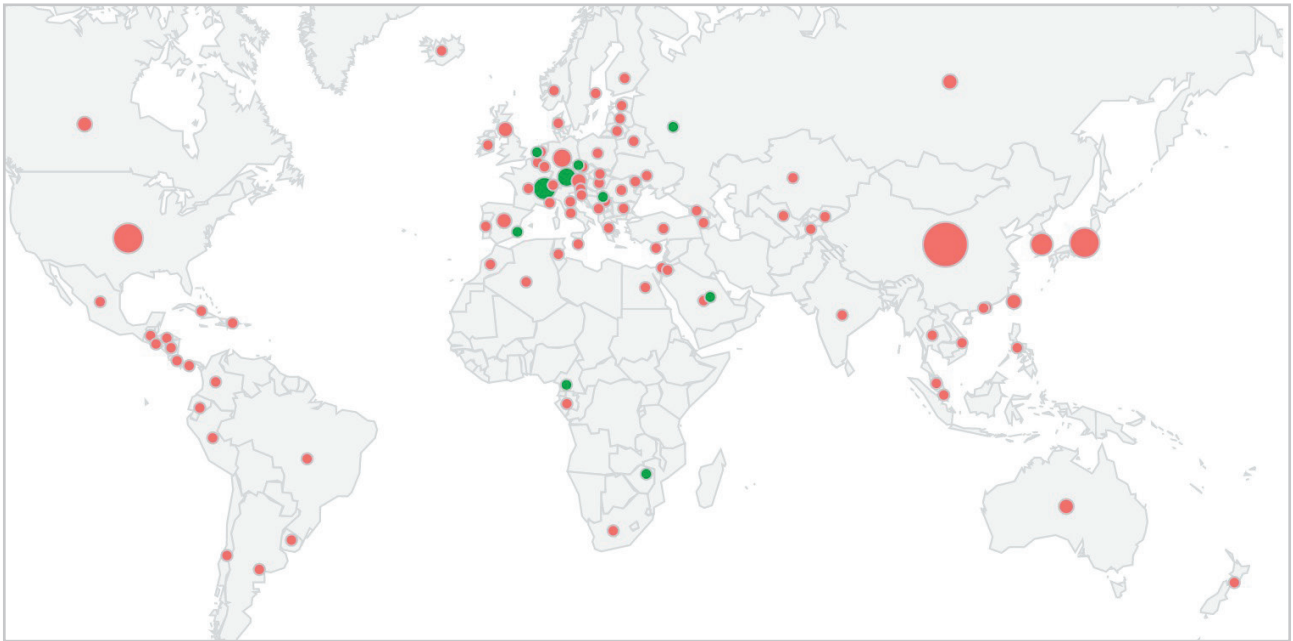


Figura 1. Número de solicitudes presentadas entre 2003 y 2017 por cada oficina nacional (en verde las oficinas internacionales e históricas). Elaborado a partir de datos de *Patstat* (2018).

Como principales ventajas sobre otras bases de datos como *NBER* (de Estados Unidos) o *IIP* (de Japón) tiene la cobertura mundial, la inclusión de más información y la existencia de algunos productos auxiliares que solventan algunos problemas, lo que la ha convertido en un estándar *de facto* (**Kang**; **Tarasconi**, 2016). Como desventajas tiene su orientación a Europa (los datos de las oficinas nacionales se intercambian con la *EPO* en base a convenios que cambian con el tiempo pudiendo dejar lagunas) y su orientación al proceso de examinación (los datos que no son vitales en el proceso de examen de la patente tienen una menor calidad).

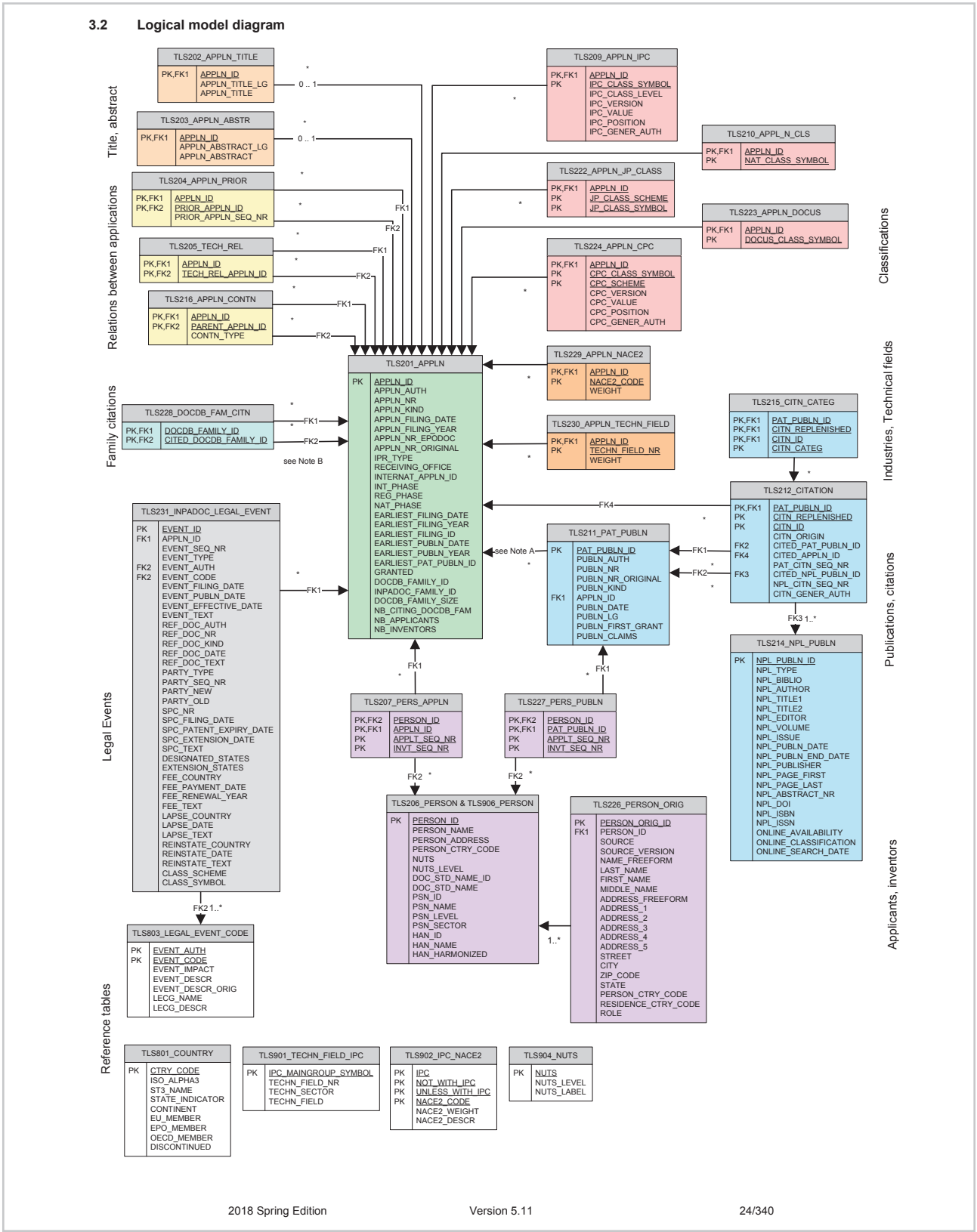


Figura 2. Esquema relacional de *Patstat* - 2018 Spring edition. Fuente: European Patent Office (2018)

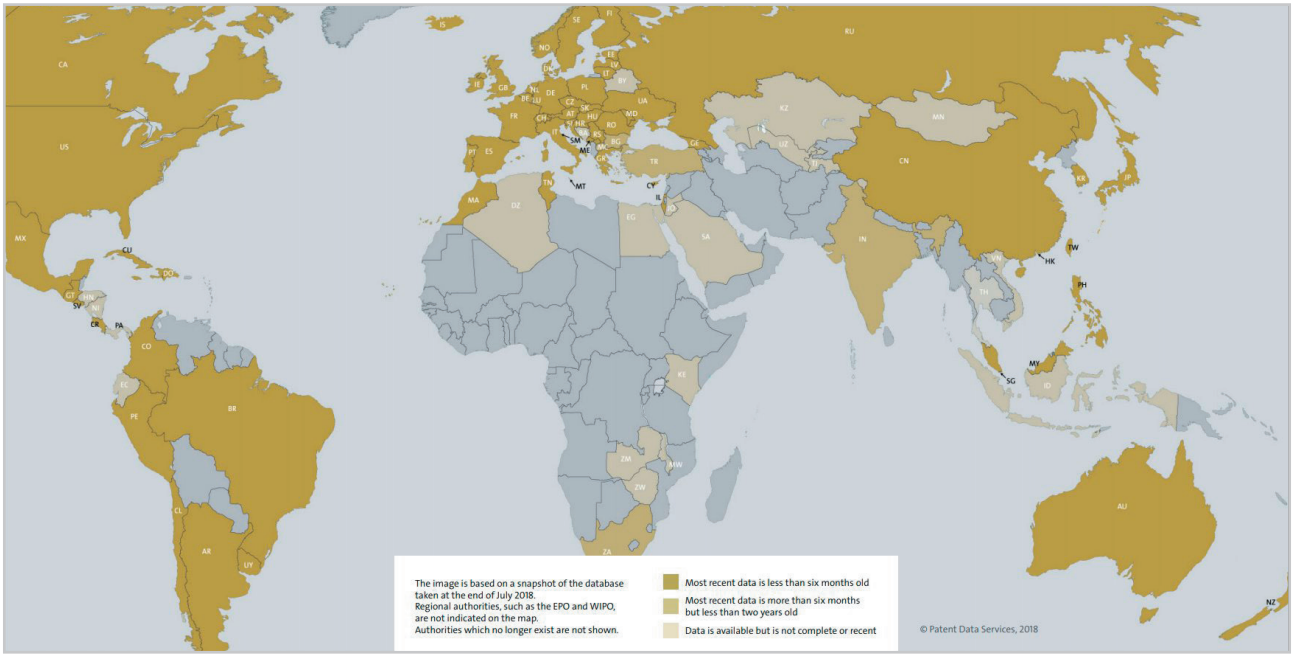


Figura 3. Cobertura de *Patstat*. Fuente: *http://documents.epo.org/projects/babylon/eponet.nsf/0/73C531E61E437E8BC1258345005975AB/$File/Coverage_of_EPO_bibliographic_data_(DOCDB)_map_en.pdf*

Patstat consiste en 2 productos:

- *Patstat global*: Tiene una cobertura mundial y contiene información bibliográfica sobre solicitudes y publicaciones, así como información legal sobre patentes.
- *Patstat EP register*: contiene información bibliográfica, procesal y legal detallada sobre solicitudes de patentes europeas y euro-*PCT* (*Patent cooperation treaty*).

Patstat es una base de datos relacional definida en el esquema de la figura 2. Se puede usar online o adquirir en DVD para instalar en un ordenador local, pudiéndose consultar mediante SQL (**De-Rassenfosse**; **Dernis**; **Boedt**, 2014). La *EPO* publica dos ediciones al año de *Patstat*, la de *Spring* y la de *Autumn*. La edición *Spring* de 2018 de *Patstat* (*Patstat - 2018 Spring edition*) es una instantánea de los datos presentes en *Docdb EPO*, base bibliográfica mundial que incluye los datos de más de 90 oficinas de patentes de todo el mundo, e *Inpadoc EPO*, base de datos mundial del status legal, tomada en la 5ª semana de 2018 (figura 3). Los datos de personas se toman de:

- *EP Patent register* para los solicitantes de la *EPO*.
- *USPTO* para datos de EUA de las patentes publicadas a partir de 1976, y de las solicitadas a partir del 29 de noviembre de 2005. Las previas se toman de la *Docdb EPO*.

La tabla TLS214_NPL_PUBLN es la que incluye los datos de las referencias NPL. A simple vista puede parecer que tiene una estructura muy rica, sin embargo, sólo están completos en todos los registros los tres primeros campos:

- NPL_PUBLN_ID: Clave numérica de la tabla.
- NPL_TYPE: Tipo de referencia NPL (tabla 1). El porcentaje de cada tipo varía poco de unas ediciones a otras.
- NPL_BIBLIO: Referencia completa (tal y como aparece en la patente, pero que no sigue una norma fija, ni tampoco está completa necesariamente).

El resto de los campos (18) se incorporaron en la versión de *Spring 2017*, y están rellenos en un pequeño porcentaje (< 20%), pero éste no aumenta significativamente en la versión de *Spring 2018* y su contenido tampoco es siempre correcto.

El porcentaje de registros con cada campo relleno depende del tipo de referencia (tabla 2). Las que son de tipo "a", son referencias pobres (*European Patent Office*, 2018) que sola-

Tabla 1. Valores de NPL_TYPE, descripción y cardinalidad

NPL_TYPE	Descripción	Nº de referencias NPL	% de referencias NPL
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

Tabla 2. Campos rellenos de la tabla TLS214_NPL_PUBLN en Patstat - 2018 Spring edition

		Citaciones pobres	Artículos					Online		
Atributos		a	b	c	i	j	s	d	e	w
Cantidades en miles		29,340	749	27	1	393	5,650	118	124	232
NPL_BIBLIO	%, redondeado	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

mente tienen poblados los tres primeros campos, y éstos suponen el 80% de las referencias. Como hemos indicado anteriormente, estos porcentajes no están variando significativamente en las versiones de *Patstat* posteriores a *Spring 2017*. Tampoco varía significativamente el porcentaje de cada campo relleno para cada tipo de referencia (tabla 2) de una edición a otra.

Existen muchos registros de esta tabla que están repetidos, en torno a una tercera parte, variando exclusivamente la clave NPL_PUBLN_ID, y dicha clave no se mantiene de una versión a otra, de modo que la única forma de relacionarlos es mediante el campo NPL_BIBLIO.

Scopus es una base de datos bibliográfica de *Elsevier* (**Hanne**, 2004; **Pickering**, 2004), que indexa 23.700 revistas científicas. Aunque no es la que mayor tiempo lleva en el mercado, varios estudios la han intentado caracterizar (**Archambault** *et al.*, 2009; **Leydesdorff** *et al.*, 2010; **De-Moya Anegón** *et al.*, 2007), y ha sido utilizada en varios estudios cienciométricos (**Gorraiz**; **Gumpenberger**; **Wieland**, 2011; **Jacsó**, 2011; **Guerrero-Bote**; **De-Moya-Anegón**, 2015; **De-Moya-Anegón** *et al.*, 2018).

En *Scopus* los documentos están clasificados por Áreas temáticas (*Subject areas*) y por Áreas temáticas específicas (*Specific subject areas* o *Categories*). Hay más de 300 *categories* que están agrupadas en 26 *subject areas*. Además, está el área *Multidisciplinary* que contiene revistas como *Nature* o *Science*.

3. Metodología

Aunque se debe aprovechar la información estructurada presente en el 20% de los registros, resulta necesario utilizar también la referencia textual normalizada (NPL_BIBLIO). Dentro de ella se pueden buscar algunos patrones, para localizar por ejemplo el año o el DOI. Por todo ello, y siguiendo las aportaciones del "*Names game*" (**Raffo**; **Lhuillery**, 2009) hemos diseñado un procedimiento divido en cuatro fases:

1. Preproceso de los datos: Preparación de los datos para facilitar y agilizar los procesos posteriores.

2. Preselección de parejas candidatas: A partir de algunas coincidencias de los elementos de las referencias, se preselecionan parejas (NPL de *Patstat*, referencia de *Scopus*) candidatas al match.

3. Evaluación automática de los las parejas candidatas:

- Se evalúan los elementos coincidentes de cada pareja candidata.
- A cada pareja se le asigna una puntuación, de modo que para cada NPL podamos obtener una lista ordenada de las referencias de *Scopus* que podrían encajar.

- La puntuación global se obtiene mediante el producto de las puntuaciones obtenidas por cada elemento de la referencia (a modo de probabilidad).

4. Validación humana:

- Por la parte alta, a partir de una determinada puntuación, las parejas con mayor puntuación pueden resultar validadas.
- Por la parte baja, a partir de una puntuación se puede descartar el encaje.
- Para cada NPL solamente se puede considerar encajada la referencia de *Scopus* con mayor puntuación (aunque también hay registros duplicados en *Scopus*).

3.1. Preproceso de los datos

En esta primera fase se trata de preparar los datos para el proceso posterior. Gran parte de este preproceso trata de solucionar algunos de los problemas de normalización de los datos. Se lleva a cabo mediante consultas SQL con los siguientes pasos:

- Unificar los registros: Como se ha indicado anteriormente, aproximadamente una tercera parte de las tuplas de la tabla están repetidas. Para reducir la carga informática lo primero que hacemos es unificar los registros generando una nueva clave.
- Identificar los registros evaluados en alguna fase anterior: En este caso ya se trabajó con la edición anterior, de modo que para evitar comenzar de cero la primera tarea era identificar los registros de ediciones anteriores ya procesados. En cada edición de *Patstat* cambia la clave primaria de la tabla TLS214_NPL_PUBLN (NPL_PUBLN_ID), de modo que la identificación se hace por el campo NPL_BIBLIO que contiene la referencia completa. Esto también será necesario hacerlo en años posteriores.
- Asignar una nueva clave numérica que nos permita hacer rodajas la tabla: En algunos de los procesos que se llevan a cabo resulta conveniente dividir la tabla en partes iguales. La forma más rápida es creando una clave numérica, donde además los registros ya asignados puedan localizarse fácilmente (por ejemplo, asignándoles una clave a partir de un determinado número).
- Localizar patrones correspondientes a DOIs: Se pueden diseñar expresiones regulares para localizar DOIs. Si se localiza un DOI en la referencia el problema está resuelto, sin embargo es muy pequeño el número de referencias que incluyen el DOI.
- Asignar años de publicación: Igualmente se pueden buscar patrones que coincidan con las cifras 2003 a 2017 que corresponde al período de referencias de *Scopus* con las que los vamos a emparejar. Como una referencia puede incluir varias cifras similares, en el caso de que el campo NPL_PUBLN_DATE contenga un valor correcto se utilizará éste. En otro caso, tenemos que tener en cuenta que habrá referencias que contengan más de un año y otras que no contengan ninguno.
- Se normalizan todos los campos textuales tanto de la tabla TLS214_NPL_PUBLN de *Patstat* como de las referencias de *Scopus*, eliminando los caracteres especiales y reduciendo a la raíz todas las palabras. De esta forma se unifican las distintas variantes léxicas de una palabra.
- Localizar los textos entre comillas, como candidatos a ser títulos. Se almacenan dichos textos normalizados y reducidos a la raíz.
- Extraer la primera palabra del campo NPL_AUTHOR o en su defecto del campo NPL_BIBLIO como candidata a ser el apellido del primer autor del paper. Se eliminan algunas excepciones (van, der, von, etc.). También se extrae el apellido del primer autor de la referencia *Scopus*. Ambos se almacenan una vez normalizados y reducidos a la raíz.
- Generar un índice invertido con las raíces extraídas del campo NPL_BIBLIO, otro con las extraídas de los títulos de revista de *Scopus* y otro con los títulos de las referencias de *Scopus*.
- Intentar asignar a cada referencia de la tabla TLS214_NPL_PUBLN una de las 23.000 revistas de *Scopus*. Para ello se utiliza por orden de prioridad el ISSN, el campo NPL_TITLE2, el título y el título abreviado de las revistas en *Scopus*. En caso de que sean necesarias comparaciones textuales se utilizan los índices invertidos para evitar una comparación a fuerza bruta. Para cada asignación se anota:
    - Cómo se ha hecho el emparejamiento (si con el ISSN, con el título, con el título abreviado, reducidos a la raíz, etc.).
    - El número de caracteres de la coincidencia (no es lo mismo la coincidencia de tres caracteres de un título abreviado que cuarenta de un título sin procesar).

Tras realizar el preproceso tenemos 24.046.625 registros de la tabla TLS214_NPL_PUBLN sin repetir, de los 36.634.177 que había originalmente en la tabla TLS214_NPL_PUBLN. De ellos 2.604.437 ya los teníamos emparejados por el mismo procedimiento de la *Spring edition de 2017*.

Igualmente, tenemos 37.792.849 referencias de *Scopus* del período 2003-2017 de todos los tipos documentales presentes en *Scopus*.

3.2. Preselección de parejas candidatas

Con los datos anteriores podemos deducir que hay $9x10^{14}$ posibles parejas formadas por una referencia NPL de *Patstat* y una referencia *Scopus*. Debido a la falta de normalización se hace necesaria una comparación directa que es imposible abordar manualmente con un número tan grande de parejas. Por esa razón, esta fase tiene como objetivo reducir el número de parejas a una cifra más manejable, pero que al mismo tiempo sea suficientemente amplia como para minimizar la posibilidad de que una pareja real se quede fuera.

Con ese fin, se utiliza una serie de reglas que se aplican en forma de sentencias SQL sobre los datos obtenidos de la fase anterior. Las reglas utilizadas son las correspondientes a las siguientes coincidencias:

- DOI
- Revista, volumen (NPL_VOLUME) y primera página (NPL_PAGE_FIRST)
- Revista, volumen (NPL_VOLUME) y número (NPL_ISSUE)
- Revista y apellido del primer autor
- Revista, volumen (incluido en NPL_BIBLIO) y primera página (incluida en NPL_BIBLIO)
- Revista, volumen (incluido en NPL_BIBLIO) y número (incluida en NPL_BIBLIO)
- Revista, volumen (incluido en NPL_BIBLIO) y apellido del primer autor
- Revista, año y primera página (incluida en NPL_BIBLIO)
- Revista, año y última página (incluida en NPL_BIBLIO)
- Revista, primera página (incluida en NPL_BIBLIO) y última página (incluida en NPL_BIBLIO)
- Primer autor y primera página (NPL_PAGE_FIRST)
- Primer autor y última página (NPL_PAGE_LAST)
- Primer autor, primera página (incluida en NPL_BIBLIO) y última página (incluida en NPL_BIBLIO)
- Revista, primera página (NPL_PAGE_FIRST) y última página (NPL_PAGE_LAST)
- Título del paper reducido a la raíz (NPL_Title1)
- Título del paper reducido a la raíz (primer entrecomillado de NPL_BIBLIO)
- Título del paper reducido a la raíz (segundo entrecomillado de NPL_BIBLIO)
- Título del paper reducido a la raíz (tercer entrecomillado de NPL_BIBLIO)
- Inclusión en NPL_BIBLIO del año y dos de las 4 raíces menos frecuentes del título de la referencia de *Scopus*
- Inclusión en NPL_BIBLIO del año y un término del título de la referencia de *Scopus* que aparezca en menos de 1.000 referencias NPL
- Apellido del primer autor (la palabra candidata a ser el apellido normalizado que se extrajo en la fase anterior) y dos de las 4 raíces menos frecuentes del título de la referencia de *Scopus*
- Apellido del primer autor (la palabra candidata a ser el apellido normalizado que se extrajo en la fase anterior) y un término del título de la referencia de *Scopus* que aparezca en menos de 1.000 referencias NPL

Como se ve, siempre que fue posible se utilizaron los campos específicos de la tabla TLS214_NPL_PUBLN, pero como no están rellenos en un gran porcentaje, también se buscaban los términos o números correspondientes en la referencia textual (NPL_BIBLIO).

Se han generado parejas candidatas para 14.758.096 referencias NPL de los 24 millones de los que partimos. Muchas referencias no enlazan a NPLs, mientras que otras apuntan a bibliografía no cubierta en *Scopus* o que fue publicada fuera

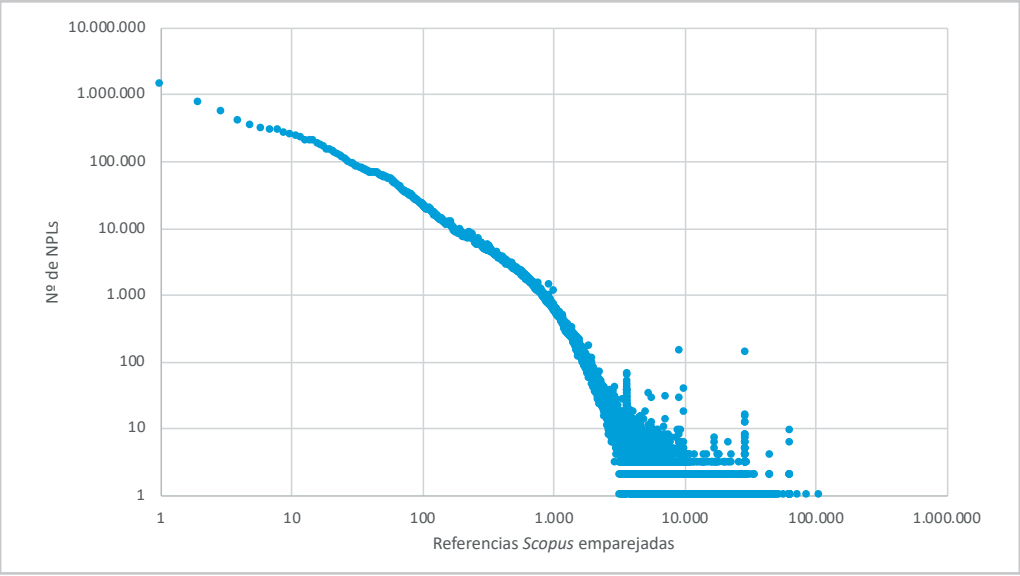


Figura 4. Gráfico de dispersión donde se representa el número de referencias *Scopus* emparejadas con cada referencia NPL

del período estudiado. El procedimiento de preselección nos genera 2.280.503.246 parejas candidatas, una vez eliminadas las correspondientes a las asignadas en la edición *Spring de 2017*. Esto quiere decir que en muchos casos una misma referencia NPL tiene muchas referencias *Scopus* candidatas. Hay 307.901 referencias NPL que tienen cada una más de 1.000 referencias *Scopus* candidatas, mientras que solamente hay 1.389.571 referencias NPL que tienen una única referencia *Scopus* candidata. La distribución sigue una ley de potencia (*power law*), como se puede observar en la figura 4.

3.3. Evaluación automática de las parejas candidatas

El objetivo de esta fase es asignar una puntuación que permita seleccionar para cada referencia NPL la referencia *Scopus* que más probabilidad tiene de referirse al mismo documento. Para ello se ha diseñado una serie de rutinas que buscan los elementos más importantes de la referencia de *Scopus* en el registro de la tabla TLS214_NPL_PUBLN. Los elementos buscados, para cada uno de los cuales se ha diseñado una rutina independiente, son los siguientes:

- Año de publicación
- Apellido del primer autor
- Título del trabajo
- Revista
- Volumen
- Número
- Páginas

En función de la calidad y de la importancia de la coincidencia cada rutina asigna una puntuación:

- En caso de que no contenga un elemento se le asigna un valor menor que uno, salvo en el caso del título o primer autor, que se le asigna uno (algunas referencias NPL no contienen el título o primer autor, pero están completamente especificadas).
- La puntuación está en función del tamaño del matching, aunque se multiplica por un factor en función de la calidad del encaje (este factor es mayor si el encaje es sin reducción a la raíz, o en los campos específicos de la tabla TLS214_NPL_PUBLN).
- La puntuación total del matching para cada pareja candidata se obtiene multiplicando el valor asignado por todas las rutinas.

3.4. Validación humana

Como se ha visto en los apartados anteriores, una referencia NPL puede no tener ninguna referencia *Scopus* candidata, puede tener una o puede tener varias. Lógicamente si alguna de las candidatas corresponde con la referencia NPL esta debería ser la que más puntuación obtiene, pero puede que ninguna de las asignadas sea válida. Por esta razón se hace necesaria una validación manual.

Con este fin se ha desarrollado una aplicación que permite la cooperación de muchas personas en la validación humana (figura 5).

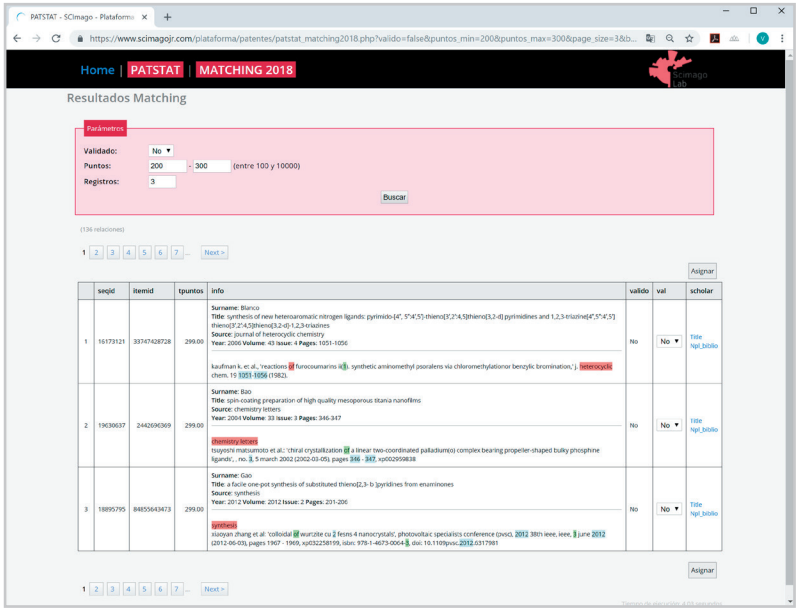


Figura 5. Captura de pantalla de la aplicación que permite validar manualmente los emparejamientos entre referencias NPL de *Patstat* y de *Scopus*.

4. Resultados

En la tabla 3 se pueden ver los resultados del proceso de emparejamiento tras la validación humana, con los porcentajes de error correspondientes en cada intervalo. Dichos porcentajes de error hasta 1.000 puntos son absolutos y en el resto de intervalos se ha hecho un muestreo de 100 parejas. Las referencias con más de 10.000 puntos se incorporan de manera automática sin necesidad de validación humana. Como se puede observar el acierto es del 100% para estos emparejamientos. A estas referencias de la edición *Spring 2018* se les deben sumar las 2.604.437 que obtuvieron 10.000 o más puntos con la versión *Spring 2017*.

Tabla 3. Distribución de las referencias NPL por intervalos de puntuación recibida con la referencia de *Scopus* que mejor encaja, y porcentaje de error

Mín. puntos	Máx. puntos	Referencias NPL	% error	Errores	Aciertos	% aciertos acumulados	% referencias procesadas
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		

En la tabla 3 podemos observar cuatro grupos de referencias procesadas con sus mínimos y máximos de puntuación. Los dos grupos inferiores (en azul y verde) han sido procesados por completo y para ellos el porcentaje de referencias NPL emparejadas con éxito es muy alto (un 99,6% de las referencias procesadas). El tamaño de estos dos grupos es de 3.206.150 referencias, lo que constituye alrededor del 18,5% de las referencias totales, pero estimamos que suponen alrededor del 83% de las referencias a las que se puede asignar un paper. El grupo de referencias con puntuaciones mínimas entre 1.000 y 10.000 (en verde) ha sido validado manualmente y constituye tan sólo el 1,4% de las referencias, dentro de las cuales se incluye el 5,9% de las referencias a las que se puede asignar un paper. En este grupo de referencias el porcentaje de acierto es muy alto (95,2%), lo que reduce mucho el esfuerzo humano necesario.

El grupo de referencias con entre 300 y 1.000 puntos muestra todavía unos porcentajes de error moderadamente bajos (en promedio inferiores al 25%) e incluye un 9,2% de las referencias correctamente emparejables en un volumen de datos equivalente al 2,7% del total. Aunque la relación entre referencias a procesar y referencias correctamente emparejables es buena, el esfuerzo neto a realizar es todavía alto, por lo que es en esta región donde entendemos que existe un mayor margen de mejora en los procedimientos automáticos. El último grupo de referencias ofrece sin embargo un balance muy poco alentador entre el esfuerzo que se debe realizar (procesar alrededor del 79% de los datos) y el beneficio esperable (un 7,9% de las referencias emparejables correctamente).

El número total de referencias incorporadas desde la nueva versión de la base de datos por este procedimiento es de 590.410, que se suman a las detectadas por el mismo procedimiento en la edición anterior, con un total agregado de unos 3,2 millones de referencias a trabajos indexados en *Scopus*. En un futuro próximo se estima que se enlazarán alrededor de 350.000 referencias más revisando manualmente las referencias con entre 300 y 1.000 puntos.

Como se puede ver en la figura 6, el número de solicitudes ha aumentado de forma exponencial a lo largo de los últimos años, pero el crecimiento del número de referencias NPL enlazadas con más de 1.000 puntos ha sido todavía más pronunciado. Parece razonable pensar que el fuerte incremento de referencias enlazadas im-

> La fase de evaluación automática puede emplearse para procesar emparejamientos que hasta ahora no se revisaban y sobre de los que sólo teníamos una estimación de su calidad

plica que los datos se están haciendo cada vez más robustos. Esta figura también permite describir el ritmo de incorporación de referencias NPL. Después de la primera fecha de publicación todavía se realizarán en muchos casos otras publicaciones que enriquecerán la bibliografía citada en las patentes, una bibliografía que no está disponible todavía en los datos de las patentes publicadas por primera vez en los últimos años.

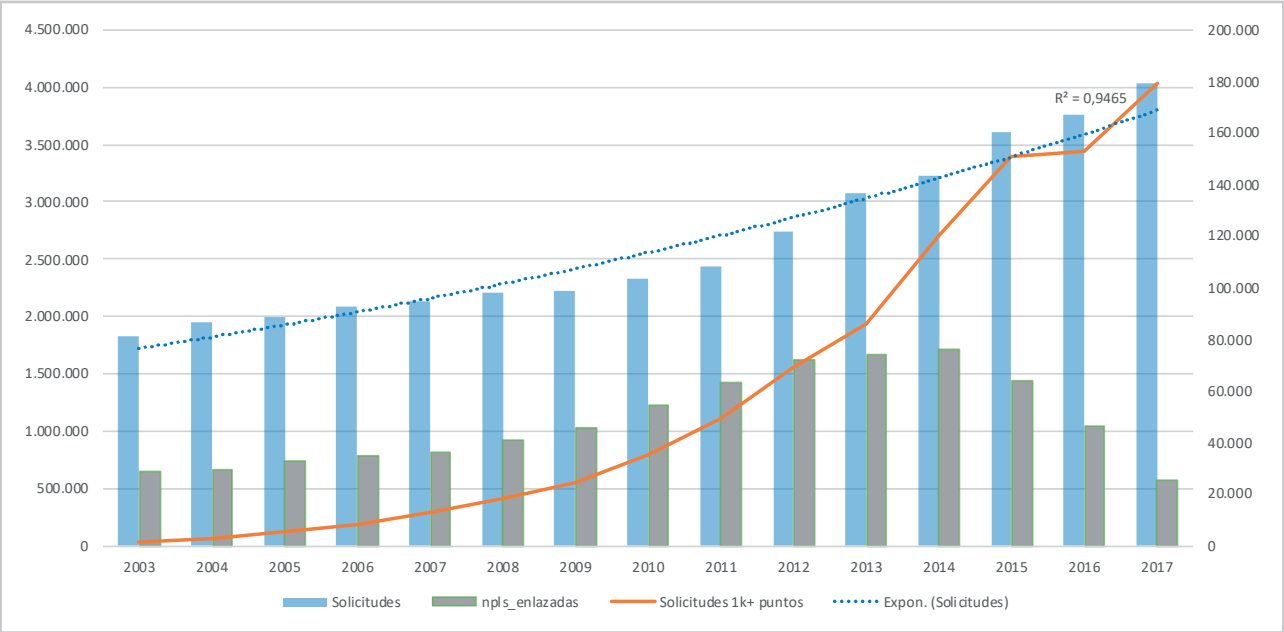


Figura 6. Evolución del número de solicitudes, NPL enlazadas y NPL enlazadas con más de 1.000 puntos (eje secundario de la derecha).

5. Conclusiones

El porcentaje de referencias correctamente emparejadas es todavía mejorable con respecto del total de referencias disponibles. Entendemos que esta situación va a evolucionar positivamente en próximas ediciones de *Patstat*, ya que los datos antes expuestos indican una mejora visible en las puntuaciones de los emparejamientos a lo largo de los últimos años. Es razonable pensar que esta tendencia va a continuar, por lo que el porcentaje de referencias correctamente emparejadas aumentará. Sin embargo, es necesario tener en cuenta que de los 24 millones de referencias procesadas hay 6,7 millones de referencias que no tienen nada que ver con ningún registro *Scopus*.

Por otra parte, existe margen todavía para la mejora en el método que hemos utilizado para llevar a cabo los emparejamientos, como hemos visto. La diferencia entre el número de referencias con 10.000 o más puntos y el número máximo que teóricamente podemos emparejar correctamente con este procedimiento es todavía importante. El trabajo continuado para mejorar la fase de evaluación automática debería redundar en un aumento efectivo del número de referencias bien emparejadas. Esto a su vez permitiría reducir el esfuerzo humano de validación necesario y emplearlo en procesar emparejamientos que hasta ahora no se revisaban y sobre de los que sólo teníamos una estimación de su calidad.

6. Nota

1. *Crios-PatStat* es una base de datos de patentes creada por un equipo de investigadores del *Centro di Ricerca su Innovazione, Organizzazione e Strategia* (*Crios*), de la *Università Bocconi*, en Milán.

En esta base de datos el usuario puede encontrar, para solicitudes de la *Oficina Europea de Patentes*, nombres de inventores y solicitantes desambiguados, así como otros datos que a menudo son difíciles de encontrar en otras bases de datos de patentes.

7. Referencias

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