Uso del Lenguaje R en Recuperación de Información aplicado a análisis bibliométrico

Lautaro Ramos[1], Soledad Retamar[1], Natalia Rapesta[1], Anabella De Battista[1],

Leandro Lepratte[2], Norma Herrera[3]

[1]Grupo de Investigación en Bases de Datos - FRCU - UTN - Entre Ríos – Argentina

[2] Grupo de Investigación de Desarrollo, Innovación y Competitividad - FRCU - UTN - Entre Ríos - Argentina

[3] Universidad Nacional de San Luis - San Luis - Argentina

{ramosl,retamars,rapestan,debattistaa,leprattel}@frcu.utn.edu.ar, nherrera@unsl.edu.ar

Resumen. *Los indicadores bibliométricos son instrumentos que permiten medir la producción científica y se utilizan para identificar, a partir del análisis de la literatura científica y tecnológica publicada, los outputs del sistema científico, en términos de performance y estructura del conocimiento. Para realizar análisis de dichos indicadores se emplean herramientas computacionales para la obtención, el tratamiento y el análisis de datos. Este trabajo presenta el uso de distintas librerías de R* (R Project n.d.) *y herramientas de visualización de información en el análisis de la producción de conocimiento, tomando como caso de estudio publicaciones científicas de Argentina y el resto de mundo que incluyen la palabra clave pesticidas.*

1. Introducción

Existe en un amplio consenso en los estudios sobre tecnología, ciencia y sociedad (CTS), al sostener que existe una relación simétrica entre estos campos del hacer humano. El rol protagónico de la ciencia y la cantidad de investigadores alrededor del mundo explorando gran variedad de disciplinas y áreas de investigación crece continuamente (Montoya et al. 2018). Del mismo modo se evidencia un crecimiento de las redes de colaboración entre investigadores y grupos de investigación de distintos países, como forma de aprovechamiento de las distintas líneas de financiamiento, que emplean criterios cada vez más complejos para determinar las líneas de investigación que se continuarán financiando. En este contexto, el análisis de redes científicas de colaboración es una herramienta que permite determinar las principales áreas de especialización de universidades y centros de investigación (Callon 1987).

Por otro lado, los indicadores bibliométricos son instrumentos que permiten medir la producción científica utilizados en el campo de la cienciometría. Se utilizan para identificar, a partir del análisis de la literatura científica y tecnológica publicada, los outputs del sistema científico, en términos de performance y estructura del conocimiento. Para realizar análisis de dichos indicadores se emplean herramientas computacionales para la obtención, el tratamiento y el análisis de datos. Este trabajo presenta el uso de distintas librerías de R (R Project n.d.) en el análisis de la producción de conocimiento, en cuanto a performance (autoría) y estructura (cognitiva y

colaboración) basado en publicaciones científicas de Argentina y el resto de mundo, utilizando como término clave de búsqueda "pesticidas". Las publicaciones científicas analizadas se obtuvieron mediante la API de Scopus (Scopus APIs - Elsevier Developer Portal n.d.). Los resultados permiten considerar diferentes indicadores de la producción de conocimiento conforme al Modo 1 (Gibbons 2015; Indicadores RICYT n.d.), tales como: cantidad de publicaciones que incluyen el término "pesticidas", clasificación de publicaciones por autor, tipo de publicación, qué palabras aparecen con más frecuencia relacionadas con el término de búsqueda, en qué áreas disciplinares están catalogadas las publicaciones, tanto para artículos producidos por autores con filiación Argentina como del resto del mundo.

2. Bibliometría

La bibliometría, como subdisciplina de la cienciometría, aplica métodos matemáticos y estadísticos en el análisis de publicaciones científicas y de los autores que la producen, a fin de de estudiar y analizar la actividad científica (Araujo Ruiz and Arencibia Jorge 2002). Parte de la necesidad de cuantificar ciertos aspectos de la ciencia para poder comparar, medir y objetivar la actividad científica. La primer definición de bibliometría menciona la aplicación de métodos estadísticos y matemáticos para definir los procesos de la comunicación escrita, la naturaleza y el desarrollo de las disciplinas científicas mediante técnicas de recuento y análisis de la comunicación (Pritchard and others 1969). En la actualidad, la definición hace referencia a la evaluación de la investigación científica a través de estudios de publicaciones que presentan resultados de investigación. Los análisis bibliométricos se basan en el supuesto de que la mayoría de los descubrimientos científicos y resultados de investigación se publican en revistas científicas internacionales a fin de permitir ser leídos y citados por otros investigadores (Rehn and Kronman 2008).

3. API de Scopus

Las API (Application Programming Interfaces) son interfaces web que permiten que distintas aplicaciones se comuniquen entre sí de forma programada. Los principales beneficios que proveen son: acceso directo a metadatos en tiempo real; arquitectura RESTful; soporte de estándares y especificaciones como W3C CORS, Dublin Core, PRISM; facilidad de integración con aplicaciones o sitios web, múltiples formatos de respuesta API compatibles; las respuestas API incluyen links a distintos recursos que simplifican la navegación y el acceso.

Scopus es la base de datos con mayor cantidad de abtracts y referencias de literatura científica revisada por pares: revistas científicas, libros y actas de congresos. Permite obtener una visión global de la producción de investigación en las áreas de la ciencia, la tecnología, la medicina, las ciencias sociales y las artes y las humanidades (Scopus APIs - Elsevier Developer Portal n.d.).

Scopus ofrece herramientas inteligentes para rastrear, analizar y visualizar la investigación, denominadas APIs, orientadas a obtener distintas metadatas de las publicaciones que almacena, como: Abstract Citation Count API, Citation Overview API, Serial Title API, Subject Classifications API, Abstract Retrieval API, Affiliation Retrieval API, Author Retrieval API, Affiliation Search API, Author Search API, Scopus Search API (Scopus APIs - Elsevier Developer Portal n.d.).

Mediante el uso de estas API se puede acceder a la literatura científica solicitada mediante distintos criterios de búsqueda para realizar análisis bibliométricos.

4. Utilización de R

Considerando diferentes indicadores bibliométricos para realizar el análisis propuesto en este trabajo, se generó en primer lugar una base de datos con la producción de artículos encontrados realizando búsquedas con la palabra clave "pesticidas", para el período 1971 - 2016, accediendo a la API de SCOPUS (Scopus APIs - Elsevier Developer Portal n.d.) mediante el uso del paquete *httr* (httr Package n.d.) de R. Se conformó un dataset registrando distintas variables de las publicaciones accedidas como: identificador, título, autores, nombre de la revista en la que fue publicada, ISSN, volumen, rango de páginas, fecha de publicación, tipo de publicación, filiación, país.

En general, las revisiones se clasifican en macro, meso o microestudios. Macro engloba el estudio de la producción científica de un país, ciudad o provincia; meso, a instituciones o grupos de investigación, y micro, a investigadores o revistas específicas.

Se obtuvieron 121.328 artículos en formato *JSON* (JSON n.d.), de los cuales 1.005 tienen filiación Argentina y el resto fueron redactados por autores con filiación de otros países. Se transformaron los artículos a un formato tabular utilizando los paquetes *jsonlite* (jsonlite package n.d.) y *plyr* (plyr Package n.d.), este último se empleó específicamente para tratar los datos de filiación de los artículos.

Empleando el paquete *ggplot2* (ggplot2 package n.d.) se confeccionaron gráficas de líneas y barras, para realizar un análisis con un enfoque cualitativo. Se procesaron los títulos, abstracts y palabras clave de los artículos para reconocer términos relevantes vinculados a la temática de estudio, y haciendo uso de la librería *igraph* (igraph package n.d.), se elaboraron redes con el fin de detectar la interrelación entre los tópicos identificados, así como también entre las áreas de conocimiento en las que se clasifican las publicaciones. Por último, para medir la relevancia de los términos relacionados a la palabra clave de búsqueda se confeccionó una nube de palabras que denota la frecuencia de aparición de cada tópico en las publicaciones haciendo uso del paquete *wordcloud* (wordcloud package n.d.).

La Figura 1 representa la cantidad de publicaciones que mencionan la palabra *pesticidas* ya sea en su título, resumen o palabras claves. De dicha figura se desprende que en la misma medida en que el tema fue tomando relevancia e incrementando la literatura científica relacionada a nivel mundial, lo hizo en Argentina.

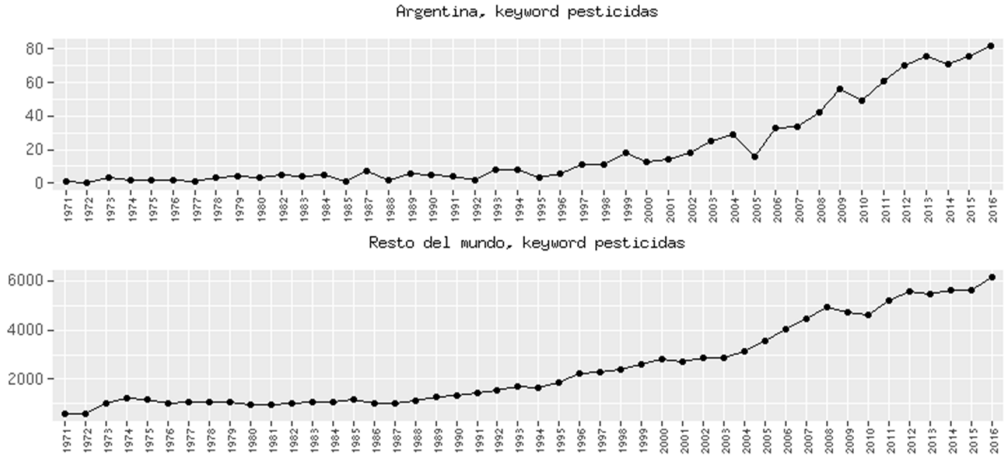

Figura 1. Cantidad de publicaciones por año para la Argentina y el resto del mundo

En la Tabla 1 se presenta el listado de áreas disponibles en Scopus para realizar la clasificación de publicaciones científicas.

Table 1. Áreas temáticas disponibles en Scopus para la clasificación de literatura científica

Listado de áreas
Environmental Science
Agricultural and Biological Sciences
Medicine
Chemistry
Biochemistry. Genetics and Molecular Biology
Pharmacology. Toxicology and Pharmaceutics
Engineering
Chemical Engneering
Immunology and Microbiology
Earth and Planetary Sciences
Materials Science
Social Sciences
Physics and Astronomy
Veterinary
Multidisciplinary
Neuroscience

Computer Science
Energy
Economics. Econometrics and Finance
Mathematics
Nursing
Business. Management and Accounting
Arts and Humanities
Health Professions
Psychology
Decision Sciences
Dentistry

Las Figuras 2 y 3 muestran la distribución de trabajos científicos publicados según áreas disciplinares, para Argentina y el resto del mundo, respectivamente. Se puede apreciar que en ambos casos la mayoría de las publicaciones pertenece al área disciplinar *Environmental Science*, siguiendo *Agricultural and Biological Science* y en tercer lugar *Medicine*. Esto significa que tanto en el país como internacionalmente las investigaciones sobre pesticidas tienen una distribución similar en esas tres áreas del conocimiento, dando lugar a la interacción y posible conformación de un campo científico y tecnológico de controversias (Latour 2011), en particular entre enfoques relacionados con la I+D y producción de este tipo de sustancias y el de la salud como se evidencia en estos datos. Lo que llevaría en una futura investigación a explorar un mapeo al interior de cada disciplina y entre las disciplinas sobre estas posibles controversias (Whatmore 2009).

En este trabajo se abordó también la detección de redes de colaboración, en particular se presenta en la Figura 7 la red de países con los que Argentina ha producido literatura científica en la que aparece la palabra clave *pesticidas*. Para el período considerado (1971-2016) los tres países con los que se han elaborado más publicaciones conjuntas son: Estados Unidos con 88, España con 79 y Brasil con 43 publicaciones en conjunto. Para esta red de colaboración se ha calculado la densidad de la red, en la que aparecen 52 países colaboradores (incluido Argentina) y 187 conexiones establecidas, por lo que la densidad de la misma es del 14%.

En las Figuras 4 y 5, se representan los tópicos principales relaciones al tema de estudio, basados en las palabras clave definidas para los artículos con filiación Argentina.

En la Figura 6 se representan los países colaboradores con puntos cuyo tamaño varía según la cantidad de publicaciones conjuntas con Argentina.

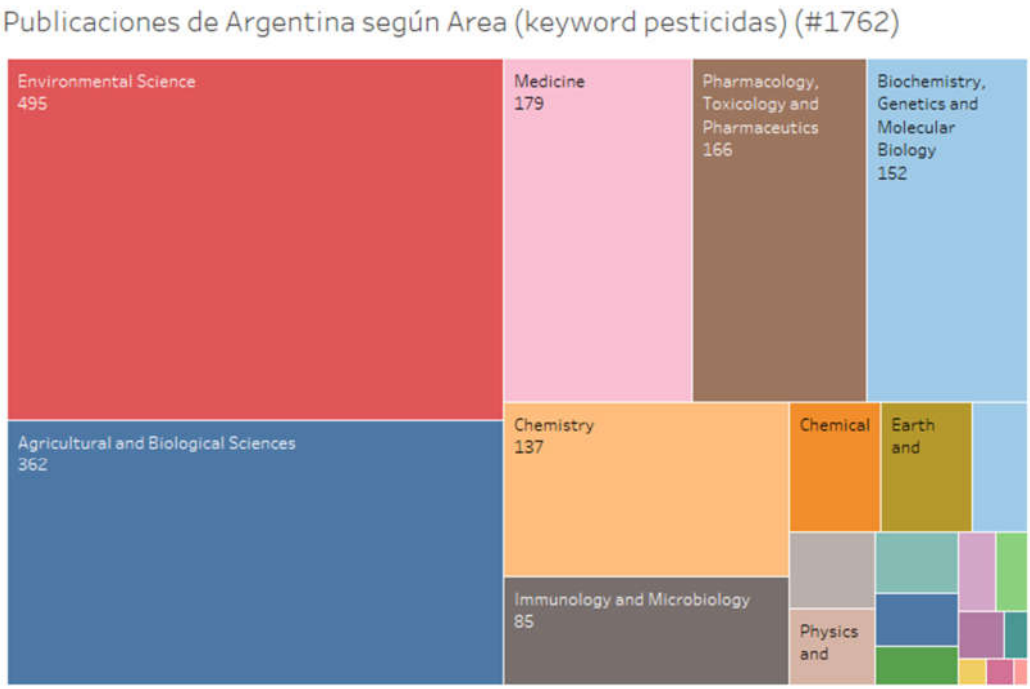

Figura 2. Cantidad de publicaciones por área disciplinar de la Argentina

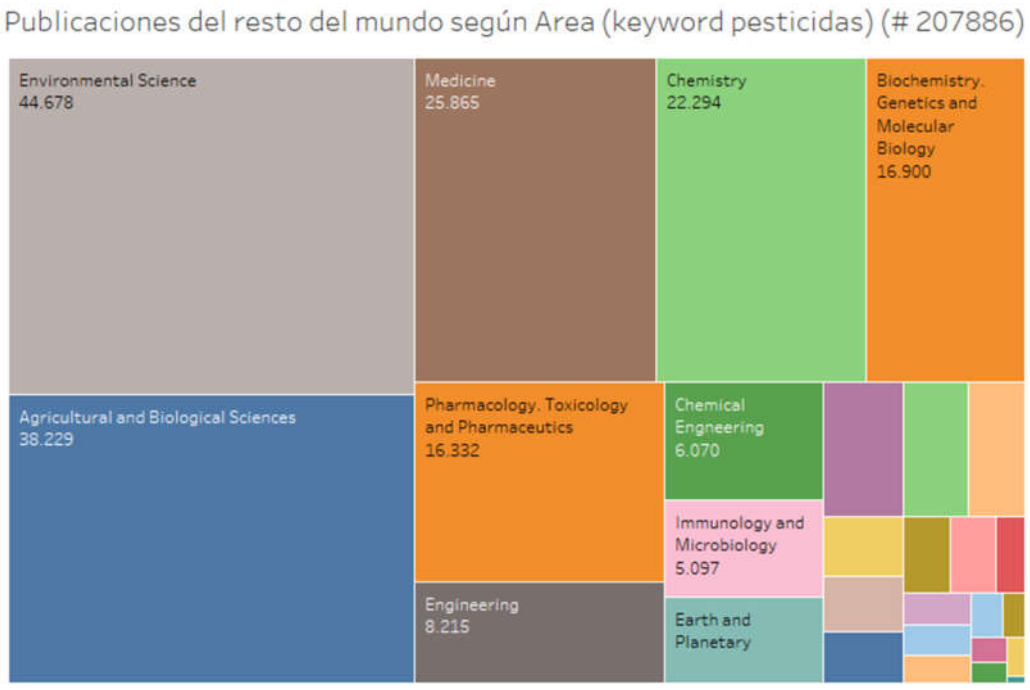

Figura 3. Cantidad de publicaciones por área disciplinar del resto del mundo

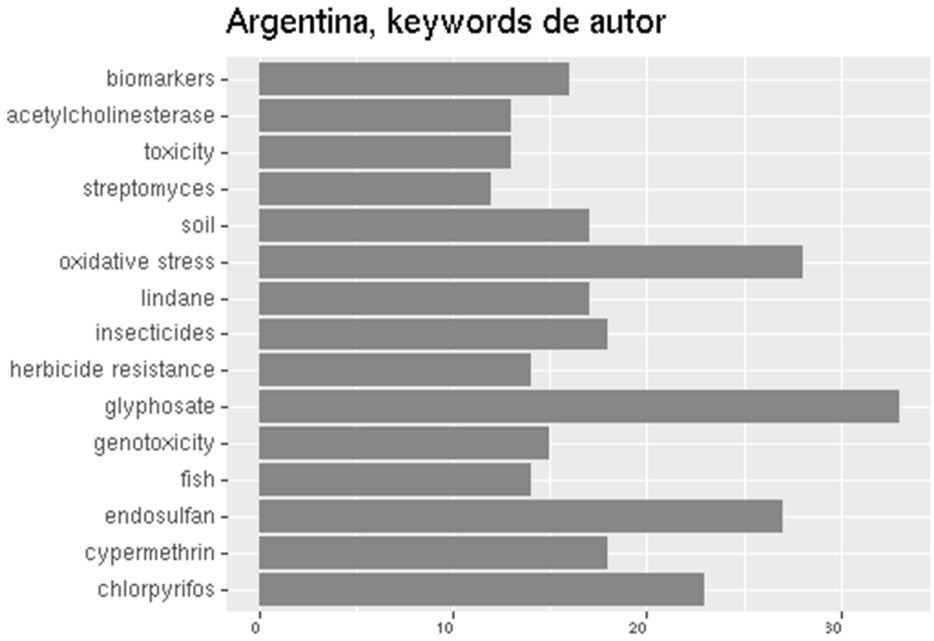

Figura 4. Tópicos principales en base a las palabras clave definidas por el autor

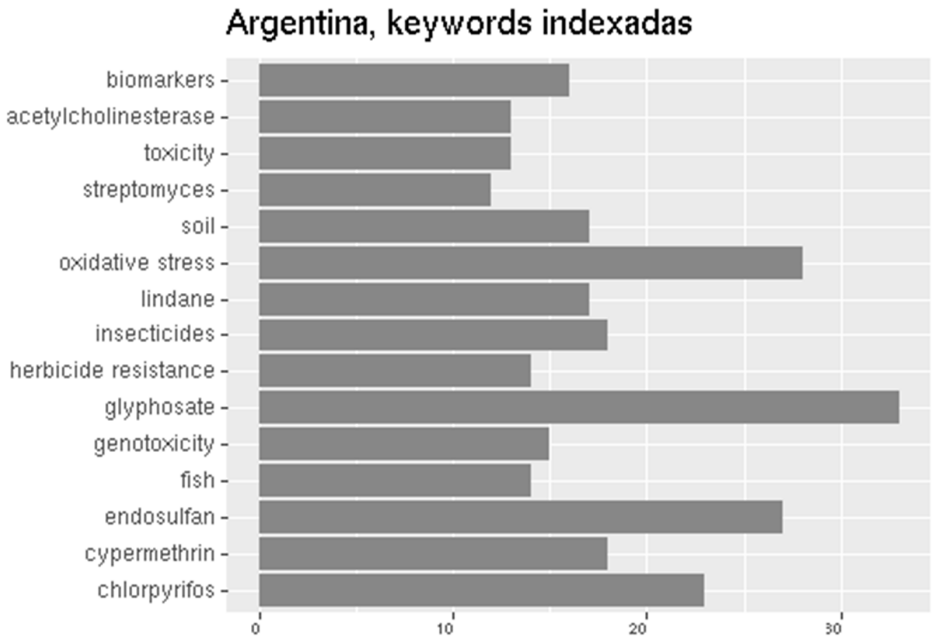

Figura 5. Tópicos principales en base a las palabras clave indexadas

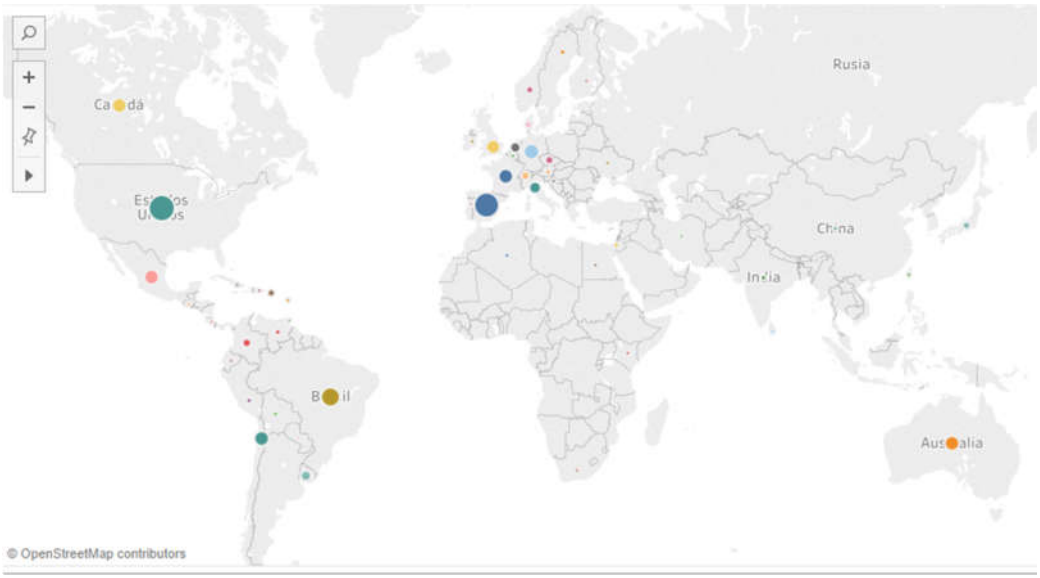

Figura 6. Países que colaboran en publicaciones con Argentina

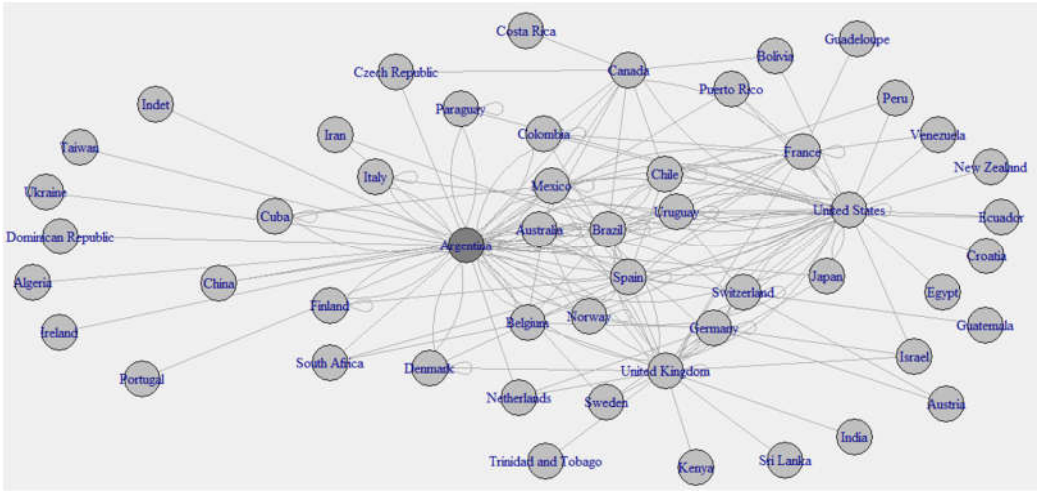

Figura 7. Red de países que colaboran en publicaciones con Argentina

5. Conclusiones

En este trabajo se presenta la primera etapa de un análisis de literatura científica especializada sobre la temática pesticidas. Se presenta el estudio de artículos publicados tanto por autores argentinos como del resto del mundo. En esta primer etapa se elaboraron una serie de conclusiones, a saber: el autor con más publicaciones para Argentina para el período considerado 1971-2016 pertenece al CONICET, con un total de 25 publicaciones a su nombre; respecto al tipo de publicaciones, la mayoría aparecen en revistas, tanto para Argentina (95%) como para el resto del mundo (91%); en el análisis de términos que aparecen con mayor frecuencia vinculados al término pesticidas se encontraron: para la Argentina *efecto/s* (135 apariciones), *ambiental* (42 apariciones) y *glifosato* (41 apariciones), y para el resto de mundo, se repite el término

efecto/s (476 apariciones) y aparecen en los primeros lugares los términos *residuos* (327 apariciones) y *agua* (280 apariciones). Para el caso argentino, la presencia del concepto *glifosato* evidencia otro posible campo de controversia sociocientífica a analizar en su especificidad para nuestro país.

Se ha propuesto como trabajo futuro analizar las redes de colaboración entre instituciones a nivel país e internacionales, analizar cuál es la estructura latente de la red de colaboración y la centralidad de los países en la investigación latinoamericana sobre pesticidas, las fuentes de financiamiento, estudiar la correlación entre la cantidad de publicaciones por áreas disciplinares y la cantidad de registros de patentes, entre otros factores de interés para estudiar el avance de la ciencia en un área determinada. El análisis de redes de colaboración entre investigadores es clave en la búsqueda de soluciones para problemas científicos y tecnológicos. Para realizar este análisis de forma sistemática se trabajará con un proceso automatizado para la extracción de grandes volúmenes de información utilizando scripts (ResNetBOT) y la API de Scopus y luego del procesamiento de los datos, herramientas de visualización, en particular de redes o grafos que permitan representar las redes de colaboración.

Algunas gráficas adicionales pueden visualizarse accediendo en la Sección Cienciometría del blog del proyecto FRCU DATA LAB disponible en (Blog del FRCU DATA LAB n.d.).

References

Araujo Ruiz, Juan A, and Ricardo Arencibia Jorge. 2002. "Informetría, Bibliometría y Cienciometría: Aspectos Teórico-Prácticos." *ACIMED* 10: 5–6. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352002000400004&nrm=iso.

"Blog Del FRCU DATA LAB."

Callon, Michel. 1987. *Society in the Making: The Study of Technology as a Tool for Sociological Analysis.*

"Ggplot2 Package." goo.gl/z1FkBw%0A (April 30, 2018).

Gibbons, Michael. 2015. "La Nueva Producción Del Conocimiento La Dinámica de La Ciencia y La Investigación En Las Sociedades Contemporáneas." *Tecnología y Construcción* 28(2). https://goo.gl/M6wQA4.

"Httr Package." https://goo.gl/VsjBcp (April 30, 2018).

"Igraph Package." https://goo.gl/VNXg3a (April 30, 2018).

"Indicadores RICYT." https://goo.gl/fhPAMW (April 30, 2018).

"JSON."

"Jsonlite Package." https://goo.gl/SDG3qX (April 30, 2018).

Latour, Bruno. 2011. "Network Theory| Networks, Societies, Spheres: Reflections of an Actor-Network Theorist." *International Journal of Communication* 5(0). https://goo.gl/zgqN13.

Montoya, Francisco G, Alfredo Alcayde, Raúl Baños, and Francisco Manzano-Agugliaro. 2018. "A Fast Method for Identifying Worldwide Scientific

Collaborations Using the Scopus Database." *Telematics and Informatics* 35(1): 168–85. http://www.sciencedirect.com/science/article/pii/S0736585317305415.

"Plyr Package." https://goo.gl/rKdHMx (April 30, 2018).

Pritchard, Alan, and others. 1969. "Statistical Bibliography or Bibliometrics." *Journal of documentation* 25(4): 348–49.

"R Project." https://www.r-project.org.

Rehn, Catharina, and Ulf Kronman. 2008. "Bibliometric Handbook for Karolinska Institutet."

"Scopus APIs - Elsevier Developer Portal." https://goo.gl/Qu1XHD (April 30, 2018).

Whatmore, Sarah J. 2009. "Mapping Knowledge Controversies: Science, Democracy and the Redistribution of Expertise." *Progress in Human Geography* 33(5): 587–98. https://doi.org/10.1177/0309132509339841.

"Wordcloud Package." https://goo.gl/gU1tc8 (April 30, 2018).