El devenir de las bases de datos académicas y sus diferentes paradigmas

The evolution of academic databases and their different paradigms

José-Luis Ortega

Ortega, José-Luis (2024). "El devenir de las bases de datos académicas y sus diferentes paradigmas". *Anuario ThinkEPI*, v. 18, e18e03.

https://doi.org/10.3145/thinkepi.2024.e18a03

Publicado en *IweTel* el 20 de febrero de 2024

José-Luis Ortega
https://orcid.org/0000-0001-9857-1511
https://www.directorioexit.info/ficha426
Instituto de Estudios Sociales Avanzados (IESA-CSIC)
Plaza Campo Santo de los Mártires
714004 Córdoba, España
jortega@iesa.csic.es

Resumen: Esta entrada pretende hacer un recorrido por los diferentes tipos de bases de datos académicas, y cómo han sido entendidas a lo largo del tiempo según su contexto histórico y tecnológico. Repasaremos el origen de los primeros índices de citas, la aparición de los buscadores académicos con la llegada de la Web y concluiremos con las nuevas bases de datos basadas en terceros. De esta forma se pretende mostrar la situación actual de estos distintos enfoques y cómo pueden responder ante los retos actuales de digitalización y ciencia abierta.

Palabras clave: Bases de datos científicas; Buscadores académicos; Índices de citas; Bases de datos basadas en terceros.

Abstract: The aim of this entry is to take a tour through the different types of scholarly databases, and how they have been understood throughout the time according to their historical and technological contexts. We will review the origins of the first citation indexes, the appearance of academic search engines with the advent of the Web and conclude with the new databases based on third parties. In this way, we intend to show the current status of these different approaches and how they can respond to the current challenges of digitization and open science.

Keywords: Scientific databases; Scholarly search engines; Citation indexes; Third-party based databases.

El origen de las bases de datos científicas

La evolución de las bases de datos científicas ha estado definida por diferentes paradigmas que han enmarcado la forma en que percibimos y apreciamos estas herramientas. Los distintos cambios tecnológicos en el procesamiento y almacenaje de datos han configurado diferentes formas de crear y entender las bases de datos académicas. Esta nota pretende describir esta evolución,

caracterizando los diferentes paradigmas y discutiendo la necesidad de plantear una perspectiva equilibrada que tenga en cuenta tanto el valor de la cobertura como la calidad y riqueza de los datos indexados.

Los primeros índices de citas nacen bajo un paradigma selectivo. Sólo incluyen lo más importante, el núcleo de una disciplina científica, ya que por motivos técnicos, económicos y de tiempo les era imposible cubrir toda la literatura existente

El fin de la Segunda Guerra Mundial consolidó un nuevo modelo científico caracterizado por la incorporación de una gran masa de científicos profesionales, el desarrollo de grandes programas de investigación (salud pública, energía, carrera espacial), y la participación fundamental del estado, tanto en su financiación como en su organización (**Bernal**, 1939). Este nuevo modelo llevó consigo la conocida explosión documental, donde el número de publicaciones científicas se iba doblando cada 5 años (**De-Solla-Price**, 1963). Esta ingente producción documental requirió de nuevos sistemas de información que pudieran facilitar el tratamiento, acceso y búsqueda de esta documentación. Los existentes servicios de resúmenes (*Chemical Abstracts*, *Physical Abstract*) no eran herramientas suficientes y ágiles para seleccionar los resultados más importantes en cada área (**Rice**; **Bernier**; **Baker**, 1960). Las nacientes bases de datos como *Excerpta Medica* (1947) o *Medline* (1966) sí ofrecían una mayor utilidad al ser más exhaustivas y requerir menos costes de producción.

Índices de citas: el paradigma selectivo

Los primeros índices de citas son creados por el *Institute of Scientific Information* (*ISI*) en 1961 (**Weinatoek**, 1971). Se trataba de bases de datos que incorporaban un novedoso índice: el índice de citas. Basado en los índices de citas jurídicos *Shepard's Citations*, estos índices permitían relacionar registros de una base de datos bibliográfica usando las citas incluidas en cada documento. Estas conexiones se revelaban altamente informativas ya que establecían un vínculo entre publicaciones que los índices permutados, de materias o palabras clave no podían capturar (**Garfield**, 1979). Además, estos índices podían ampliar la búsqueda de forma considerable, ya que superaban las limitaciones de los índices temáticos en las búsquedas interdisciplinares. Sin embargo, el principal problema de estos índices es que las citas podían apuntar a documentos fuera de la base de datos, poniendo de manifiesto limitaciones de cobertura. Estos índices, como la mayoría de las bases de datos del momento, basaban su cobertura en las fuentes primarias de publicación, las revistas. Por lo tanto, era fundamental justificar las revistas usadas, ya que su selección podría producir sesgos. Por ese motivo, en 1971 se publicó el *Journal Citation Report*. Un informe que ordenaba las revistas según el número de citas que recibían sus artículos (*Journal Impact Factor*), lo que permitía evidenciar la razón de su inclusión en la base de datos. De esta forma, los primeros índices de citas nacen bajo un paradigma selectivo. Esto es, las bases de datos sólo tienen que indizar lo más importante, el núcleo de una disciplina científica, ya que por motivos económicos y de tiempo era imposible cubrir toda la literatura existente. Sin embargo, este criterio puede producir importantes sesgos al no definir correctamente el núcleo. De hecho, los índices creados por el *ISI* (hoy *Web of Science*) presentan aún hoy importantes sesgos de cobertura, puesto que el núcleo está dominado por revistas de lengua inglesa, obviando el peso de otras lenguas en áreas más locales como las Humanidades, Ciencias Sociales y Recursos Naturales (**Van-Leeuwen** *et al.*, 2001; **Liang**; **Rousseau**; **Zhong**, 2013).

Buscadores académicos: el paradigma inclusivo

La aparición de la Web supuso una revolución de todo el sistema de publicación científica y las bases de datos académicas no quedaron al margen de esta transformación. La publicación de revistas electrónicas y la formación de repositorios permitió que cada vez más se pudiera acceder a re-

sultados científicos en la Web. A medida que las publicaciones científicas crecían en el medio digital, se vio necesaria la creación de sistemas de información que recopilasen y recuperasen estos documentos. *CiteSeer* (1997) puede ser considerado el pionero de los buscadores académicos, al ser el primer sitio que usaba un robot para recopilar documentos científicos, extraer metadatos y citas de ellos, y un motor de búsqueda para recuperarlos (**Ortega**, 2014). La principal ventaja de estos servicios es que no requieren fuentes para seleccionar las publicaciones. No importa si están publicados en revistas, o no son estrictamente artículos de investigación, sólo es necesario que sean accesibles a través de un URL. Esta libertad posibilita que el tamaño de los buscadores dependa únicamente de la capacidad del *bot* de rastrear la Web, y no tanto de cómo y dónde se ha publicado. Sin embargo, el talón de Aquiles de estos servicios reside en la indización. La extracción de datos directamente de los documentos (*parsing*) genera muchos errores y pérdidas de información, debido fundamentalmente a que muchos de ellos no incorporan suficientes y normalizados metadatos. Así, mientras los buscadores se destacan en la localización de publicaciones, su principal inconveniente está en la pobre calidad de los datos que ofrecen. Un ejemplo paradigmático es *Google Scholar*. Creado en 2004, este buscador académico incorpora el potencial de *Google* localizando documentos, pero obvia la descripción detallada de ellos. En un principio recibió críticas por sus errores de indización y el pobre tratamiento de los documentos, pero su gran alcance pronto hizo olvidar estas limitaciones, convirtiéndose en el mejor competidor de los índices de citas tradicionales (**Jacsó**, 2005; 2008; **Orduña-Malea**; **Martín-Martín**; **Delgado-Lopez-Cozar**, 2017). Este poderío tecnológico hizo bascular el paradigma de las bases de datos académicas hacia un paradigma inclusivo. Ahora primaba el tamaño de la base de datos frente a la calidad de la información que contenía. A medida que el sistema de publicación se volcaba al mundo digital, los buscadores como *Google Scholar* y *Microsoft Academic* iban creciendo en tamaño, convirtiéndose en las principales fuentes en la búsqueda de literatura científica (**Martín-Martín** *et al.*, 2021).

El poderío tecnológico de los buscadores hizo bascular el paradigma de las bases de datos académicas hacia un paradigma inclusivo Ahora primaba el tamaño de la base de datos frente a la calidad de la información

Bases de datos basadas en terceros: hacia un equilibrio de paradigmas

La segunda década del siglo XXI vio surgir una nueva generación de productos. *Scilit* (2014), *Dimensions* (2018) y la sección de publicaciones científicas de *The Lens* (2018) son algunos ejemplos de bases de datos académicas basadas en terceros (*thrid-party databases*). Estas plataformas se caracterizan por ofrecer búsquedas en abierto, sin coste de suscripción, y construidas a partir de datos externos suministrados por fuentes abiertas como *Crossref*, *PubMed* o *Microsoft Academic Graph*. Dentro del anterior paradigma inclusivo, la principal apuesta de estos productos es la cobertura. Con filtros muy laxos a la hora de incorporar contenidos de fuentes y formatos diferentes (patentes, financiación, libros, datos, etc.). Sin embargo, este afán por el tamaño en sí está llevando a muchas bases de datos a desatender la propia calidad informativa de los documentos que indizan. En un reciente estudio (**Delgado-Quirós** *et al.*, 2024) se ha podido comprobar cómo el uso acrítico de fuentes secundarias como *Crossref* está inflando la cobertura de bases de datos como *Scilit* o *The Lens* y, en menor grado, *Dimensions* con materiales editoriales como portadas, índices, anuncios, comentarios, etc. Por el mero hecho de tener un DOI asignado, se incluyen documentos con dudosa calidad o valor informativo. Aunque el peso de estos materiales no es muy elevado (se estima un 5% en *Crossref*), puede provocar ruido en la recuperación al no poder ser filtrados adecuadamente. Pero quizás más preocupante es la calidad de los metadatos de estas bases de datos. A excepción de *Dimensions*, muchas de estas nuevas bases de datos presentan problemas en la indización de

resúmenes, asignación de fechas, descripciones bibliográficas, tipologías documentales, etc., evidenciando en algunos casos un pobre procesamiento de los datos (**Delgado-Quirós**; **Ortega**, 2024).

Los buscadores académicos deben ser entendidos como grandes localizadores, pero pobres descriptores de literatura científica

Los buscadores académicos no están exentos de estos problemas. Es más, este tipo de documentación poco informativa es más visible en este tipo de productos. La obtención de datos directamente del rastreo de la Web puede ocasionar que se incorporen documentos no estrictamente científicos. Un ejemplo lo tenemos en *Google Scholar*. Sus criterios de inclusión (*Inclusion Guidelines for Webmasters*) nos dicen que

> "Contenidos como noticias, artículos de magacines, reseñas de libros y editoriales no son apropiados para *Google Scholar*" (*Google Scholar*, 2024).

Nótese que dice no apropiado, no que sean excluidos. Lo que viene a decir que *Google Scholar* no revisará si los documentos que localiza se ajustan a estos criterios. Sólo tenemos que hacer una búsqueda por "*book review*" o "editorial" y ver los resultados.

Pero quizás el mayor problema de los buscadores son las dificultades de acceso a los metadatos o a las propias publicaciones. Nuestro estudio detectó que entorno a un 30% de las publicaciones (**Delgado-Quirós** *et al.*, 2024) que un buscador no localiza es debido a directrices de exclusión de robots. Redireccionamientos, páginas no activas y otros problemas técnicos impiden el acceso y la posterior indización de estas publicaciones. A esto también tenemos que sumar que muchas editoriales y repositorios incluyen pocos o ningún metadato, o utilizan programación, lo que dificulta la extracción de datos sobre las publicaciones. Desde sus comienzos, estos productos han arrastrado este déficit en la calidad de sus metadatos, por lo que su incapacidad de solucionarlo deja patente que los buscadores académicos deben ser entendidos como grandes localizadores, pero pobres descriptores de literatura científica.

En definitiva, los tradicionales índices de citas, consciente de su incapacidad de competir con los buscadores en cobertura, apuestan por un modelo exclusivo. Se centran en cubrir el núcleo central de la ciencia y se afanan en competir por el tratamiento y enriquecimiento de datos. Mientras, los buscadores triunfan en cobertura, pero con pobre valor añadido. En medio, las nuevas bases de datos basadas en terceros tienen el reto de equilibrar ambos modelos. Tienen la oportunidad de contar de forma libre y fácil con diferentes y grandes fuentes de datos, pero por otro se enfrentan a la obligación de competir en procesamiento y tratamiento de datos a fin de crear herramientas fiables y de gran valor informativo. Por ejemplo, a día de hoy, todas las nuevas bases de datos basadas en fuentes externas, tiene como principal proveedor a *Crossref*. Esto nos dice que en cuestión de cobertura no hay mucho en donde competir, la diferencia estará en el valor añadido, en cómo esos datos son limpiados y corregidos; son enriquecidos con identificadores e información complementaria de autores, organizaciones, revistas o disciplinas; y nuevos y consistentes indicadores son incorporados. En un escenario altamente competitivo, sólo unos pocos tendrán éxito y podrán hacer valor su modelo de negocio. En cualquier caso, el usuario es el máximo beneficiado, ya que podrá elegir y exigir.

Las bases de datos basadas en terceros deberán competir más en el valor añadido que aporten a sus datos (metadatos, identificadores, limpieza, etc.), que en la coberturaa

Referencias

Bernal, John D. (1939). *The social function of science*. London: George Routledge & Sons.

De-Solla-Price, Derek J. (1963). *Little science, big science*. New York: Columbia University Press. ISBN: 978 0231918442

Delgado-Quirós, Lorena; **Aguillo, Isidro F.**; **Martín-Martín, Alberto**; **Delgado López-Cózar, Emilio**; **Orduña-Malea, Enrique**; **Ortega, José-Luis** (2024). "Why are these publications missing? Uncovering the reasons behind the exclusion of documents in free-access scholarly databases". *Journal of the Association for Information Science and Technology*, v. 75, n. 1, pp. 43-58.
https://doi.org/10.1002/asi.24839

Delgado-Quirós, Lorena; **Ortega, José-Luis** (2024). "Completeness degree of publication metadata in eight free-access scholarly databases". *Quantitative Science Studies*.
https://doi.org/10.1162/qss_a_00286

Garfield, Eugene (1979). *Citation indexing its theory and application in science, technology and humanities*. New York: Wiley. ISBN: 978 0894950254

Google Scholar (2024) *Inclusion Guidelines for Webmasters*.
https://scholar.google.com/intl/es/scholar/inclusion.html#content

Jacsó, Peter (2005). "Google Scholar: the pros and the cons". *Online information review*, v. 29, n. 2, pp. 208-214.
https://doi.org/10.1108/14684520510598066

Jacsó, Peter (2008). Google scholar revisited. *Online information review*, v. 32, n. 1, pp. 102-114.
https://doi.org/10.1108/14684520810866010

Liang, Liming; **Rousseau, Ronald**; **Zhong, Zhen** (2013). "Non-English journals and papers in physics and chemistry: Bias in citations?". *Scientometrics*, n. 95, 333-350.
https://doi.org/10.1007/s11192-012-0828-0

Martín-Martín, Alberto; **Thelwall, Mike**; **Orduña-Malea, Enrique**; **Delgado-López-Cózar, Emilio** (2021). "*Google Scholar*, *Microsoft Academic*, *Scopus*, *Dimensions*, *Web of Science*, and *OpenCitations' COCI*: a multidisciplinary comparison of coverage via citations". *Scientometrics*, v. 126, n. 1, pp. 871-906.
https://doi.org/10.1007/s11192-020-03690-4

Orduña-Malea, Enrique; **Martín-Martín, Alberto**; **Delgado-Lopez-Cozar, Emilio** (2017). "Google Scholar as a source for scholarly evaluation: A bibliographic review of database errors". *Revista española de documentación científica*. v. 40, n. 4.
https://doi.org/10.3989/redc.2017.4.1500

Ortega, José-Luis (2014). *Academic search engines: A quantitative outlook*. Cambridge, UK: Chandos Publishing (Elsevier). ISBN: 978 1843347910

Rice, Randall G.; **Bernier, Charles L.**; **Baker, Dale B.** (1960). "Scientific abstracting and indexing services". *STWP Review*, v. 7, n. 3, pp. 11–15.
http://www.jstor.org/stable/43091061

Van-Leeuwen, Thed N.; **Moed, Henk F.**; **Tijssen, Robert J.**; **Visser, Martijn S.**; **Van-Raan, Anthony F.** (2001). "Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance". *Scientometrics*, n. 51, pp. 335-346.
https://doi.org/10.1023/A:1010549719484

Weinatoek, Melvin (1971). "Citation indexes". *Encyclopaedia library and information science*, 5, 16-40.
http://www.garfield.library.upenn.edu/essays/V1p188y1962-73.pdf