

# Cobertura lingüística en las nuevas bases de datos

RODRIGO SÁNCHEZ-JIMÉNEZ

Departamento de Biblioteconomía y Documentación.  
Universidad Complutense de Madrid

Este trabajo analiza la cobertura lingüística de cinco bases de datos bibliográficas, comparando los modelos tradicionales de suscripción (Web of Science y Scopus) con nuevas infraestructuras abiertas y agregadores como OpenAlex, OpenAIRE y SciLit. El estudio parte de la problemática de los sesgos geográficos y lingüísticos inherentes a las fuentes clásicas para evaluar si las nuevas plataformas ofrecen una representación más diversa de la ciencia global. Para ello, se recuperó y armonizó la totalidad de la producción indexada hasta el año 2025 en cada fuente, examinando la distribución de los veinte idiomas principales con el fin de cuantificar las diferencias de visibilidad que ofrecen los distintos modelos de indexación.

Los resultados revelan disparidades abismales en el volumen de documentos no anglófonos, con plataformas como OpenAlex que afloran millones de registros en lenguas asiáticas (japonés, indonesio, coreano) y de Oriente Medio que permanecían prácticamente invisibles en las bases comerciales. No obstante, el análisis advierte sobre el compromiso entre cantidad y calidad del metadato: frente a la selección editorial de WoS o Scopus y la recolección "notarial" de OpenAIRE (con menor completitud), OpenAlex logra una cobertura masiva mediante inferencias algorítmicas que introducen cierto margen de error. Se concluye así que el profesional y el investigador se enfrentan a dos ecosistemas complementarios, debiendo elegir entre el control de la élite científica o la panorámica inclusiva —pero más ruidosa— de la ciencia global.

Bases de datos bibliográficas, sesgos indexación, calidad metadatos, cobertura lingüística, visibilidad investigación, acceso abierto

This study examines the linguistic coverage of five bibliographic databases, comparing traditional subscription-based models (Web of Science and Scopus) with emerging open infrastructures and aggregators such as OpenAlex, OpenAIRE, and SciLit. The analysis begins with the well-documented issue of geographic and linguistic biases inherent in classical sources, in order to assess whether these new platforms provide a more diverse representation of global science. To this end, the complete body of indexed output available in each source up to 2025 was retrieved and harmonized, and the distribution of the twenty most prevalent languages was analyzed to quantify differences in visibility across indexing models.

The results reveal stark disparities in the volume of non-English documents, with platforms like OpenAlex surfacing millions of records in Asian languages (Japanese, Indonesian, Korean) and Middle Eastern languages that remain largely invisible in commercial databases. However, the analysis also highlights the trade-off between quantity and metadata quality: whereas WoS and Scopus rely on editorial selection and OpenAIRE on a more "notarial" harvesting approach (with lower completeness), OpenAlex achieves massive coverage through algorithmic inference, introducing a certain margin of error. The study concludes that professionals and researchers now navigate two

*complementary ecosystems, needing to choose between the curated selectivity of the scientific elite and the more inclusive—though noisier—panorama of global science.*

Bibliographic databases, indexing biases, metadata quality, linguistic coverage, research visibility, open access

## Introducción

La evaluación de la producción científica internacional ha dependido históricamente de grandes bases de datos como Web of Science (WoS) y Scopus. Aunque estas infraestructuras han permitido estandarizar la medición de la actividad científica, numerosos estudios han señalado que presentan sesgos significativos en términos de cobertura geográfica y lingüística (Vera-Baceta, 2019; Asubiaro, 2024). En particular, muestran una mayor presencia de publicaciones anglófonas y de disciplinas centrales dentro de la ciencia internacional, lo que tiene implicaciones para la visibilidad de comunidades científicas que publican en otras lenguas y tradiciones académicas locales (Singh et al., 2021; Visser et al., 2021). Estas diferencias en cobertura, ampliamente documentadas en la literatura, se reflejan tanto en la selección de revistas indexadas como en las dinámicas de citación y evaluación.

En los últimos años, la aparición de nuevas infraestructuras como Dimensions, The Lens o incluso las “novísimas bases de datos”, como OpenAlex, ha ampliado el panorama de la información científica. Estas plataformas se caracterizan por políticas de indexación diferentes, por la posibilidad de acceso abierto o semilibre a grandes volúmenes de datos y por una mayor transparencia en la estructura de los metadatos. Debido a ello, diversos autores han planteado que estas bases podrían ofrecer una representación diferente de los sistemas científicos en diversas lenguas y culturas científicas.

Es importante tener en cuenta que el número creciente de bases de datos que han ido surgiendo a lo largo de los últimos años es posible tanto por la demanda de este tipo de productos como por la amplia disponibilidad de metadatos básicamente gratuitos. La demanda viene de la mano de una visión más diversa de la actividad científica, y también de las sucesivas iniciativas que pretenden reformar la forma en que se evalúa la actividad científica.

Desde el punto de vista de la disponibilidad, a día de hoy es posible acceder de forma básicamente gratuita a grandes conjuntos de datos, incluyendo PubMed y Crossref, pero también los registros de Microsoft Academic Graph, el difunto experimento cuyo cuerpo fue donado a la ciencia (y a la comunidad científica) y OpenCitations. Confluyen por tanto factores que no estaban presentes hace unos años y que facilitan emprender proyectos de envergadura suficiente como para convertirse en elementos relevantes del ecosistema de los sistemas de información científica.

¿Ahora bien, tienen realmente estas nuevas bases de datos una cobertura significativamente distinta de la ciencia desde el punto de vista geográfico y lingüístico? En el caso de Dimensions, una de las primeras “nuevas bases de datos”, parece que existe una ligera mejora en la cobertura de la ciencia escrita en idiomas diferentes del inglés. Visser et al. (2021) cifran esta diferencia en un 14%, frente al 10% de Scopus. Delgado-Quirós y Ortega (2024), en su análisis de las nuevas bases de datos abiertas encuentran una creciente presencia de trabajos escritos en idiomas diferentes al inglés. Sin embargo, también apuntan a un problema importante para aquellos que estudian el problema: la información sobre el idioma de la

publicación no siempre está disponible, y cuando lo está, se ha generado de formas distintas según la base de datos, por lo que hay que tratar cualquier estudio sobre la cobertura lingüística con un poco de reserva.

Los primeros estudios exhaustivos sobre la cobertura lingüística en OpenAlex exploran esta cuestión. Céspedes et al. (2025), en un análisis comparativo con Web of Science y mediante la validación manual de miles de documentos, muestran que OpenAlex dispone de una cobertura lingüística más amplia que WoS. Sin embargo, también subrayan que la calidad del metadato "idioma" presenta inconsistencias: los algoritmos de identificación pueden sobrerrepresentar algunas lenguas e infradetectar otras, lo que introduce incertidumbres en los análisis que utilizan directamente estas etiquetas. De forma complementaria, Holmberg Sainte-Marie et al. (2025) evalúan el rendimiento de algoritmos de detección de lengua aplicados a OpenAlex y confirman que persisten errores sistemáticos, especialmente en lenguas con escasa presencia en los corpus de entrenamiento.

De esta forma, si queremos analizar patrones lingüísticos o incluso geográficos, debemos tener cuidado y considerar las 'peculiaridades' de cada base de datos, ya sean inconsistencias en los metadatos o sesgos en los algoritmos. Con esa cautela necesaria, este artículo pretende ofrecer una visión general y actualizada del panorama, asumiendo que los datos disponibles no son perfectos. Es cierto que las nuevas bases de datos ofrecen una visión más inclusiva de la ciencia, pero la 'torre de Babel' científica sigue presentando desafíos técnicos importantes. Entender estas limitaciones no es solo cosa de expertos en bibliometría; es importante para tener una imagen más completa, global e inclusiva de la actividad científica actual. En las siguientes páginas veremos qué podemos esperar realmente de estas nuevas herramientas.

## Las bases de datos

Hemos seleccionado cinco bases de datos representativas de diferentes modelos. Scopus y Web of Science representan el modelo comercial tradicional por suscripción. Son bases de datos con coberturas amplias de la ciencia internacional, que tienen procesos de selección y comités editoriales. Esto implica al menos una preocupación por la calidad de las fuentes que indexan, que son además revisadas y están sujetas a la desindexación cuando dejan de cumplir los criterios exigidos. Por otra parte, han invertido tradicionalmente en la calidad y completitud de sus datos, por lo que son fuentes de prestigio, aunque tienen los problemas y sesgos ya mencionados.

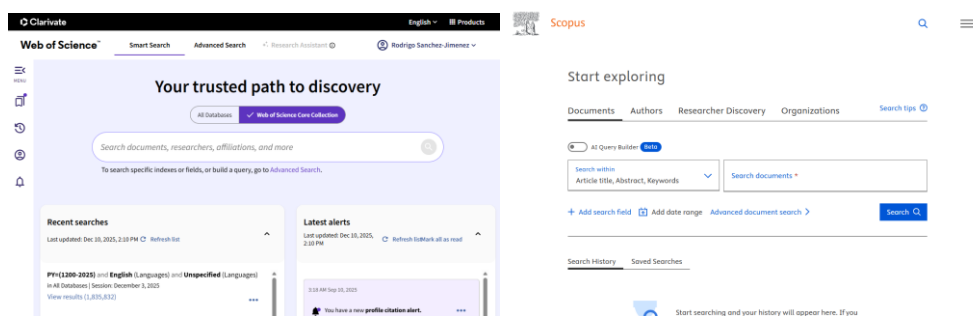


Figura 1: capturas de las páginas de inicio de Web of Science y Scopus

OpenAlex y OpenAIRE son lo que podríamos denominar bases de datos libres, que no sólo permiten la realización de búsquedas de forma gratuita, sino que permiten además utilizar el API con restricciones muy generosas, e incluso descargar sin coste los datos al completo y utilizarlos bajo licencias de acceso abierto CCo (uso libre sin restricciones) y CC-BY (uso libre con atribución), respectivamente. Ninguna de estas bases de datos tiene criterios editoriales per se, e indexan básicamente todo lo que encuentran en repositorios o en otras fuentes de datos bien establecidas, como PubMed, Crossref y Microsoft Academic Graph. Sí que se encargan (en general con bastante éxito) de identificar y consolidar duplicados, lo que no es una tarea sencilla. Por otra parte, aunque ambos proyectos han dedicado recursos crecientes a la mejora de sus datos, todo parece indicar que en general no están a la altura de sus competidoras comerciales. Hay otras bases de datos abiertas muy relevantes, como The Lens, pero no la hemos incluido al no ofrecer filtros o análisis por idioma en su interfaz público.

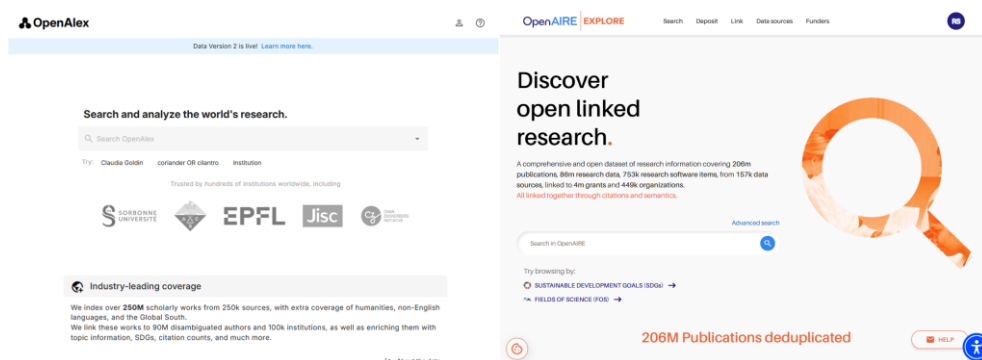


Figura 2: capturas de las páginas de inicio de OpenAlex y Scopus

Existe un tercer modelo a caballo entre las bases de datos comerciales y las bases de datos libres, el de las bases de datos gratuitas (pero no libres), o parcialmente gratuitas (freemium). Este es el caso de Dimensions, propiedad de Holtzbrinck, también propietaria de Springer Nature. Es probablemente el ejemplo más conocido, pero no hemos podido incluirlo precisamente porque entre los filtros o análisis disponibles en la versión gratuita no se incluye el idioma. En su lugar hemos utilizado SciLit, que ofrece acceso gratuito con pocas limitaciones al interfaz de búsqueda y varias opciones de análisis interesantes, aunque no es una base de datos libre. SciLit es editado por MDPI y aunque tiene un consejo asesor, no tiene una política de indexación específica. Funciona como un agregador masivo alimentado principalmente por el flujo de Crossref y datos de editoriales, fundamentalmente al margen del grafo de Microsoft Academic. A pesar de que agrega contenidos fundamentalmente libres se trata de un producto propietario, lo que limita muy significativamente tanto la recogida de datos como su uso.

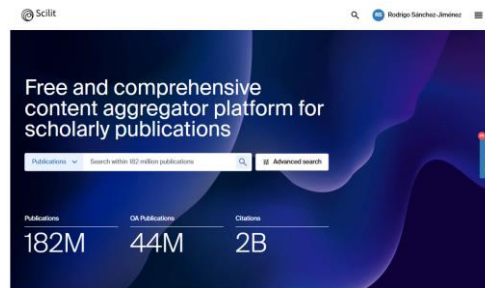


Figura 3: captura de la página de inicio de ScilLit

La selección de bases de datos pretende reflejar una diversidad de modelos que actualmente parecen gozar de cierto vigor de manera simultánea, lo que es notable teniendo en cuenta el hecho de que son competidores mutuos. Las limitaciones de nuestra metodología hacían que quedara fuera un cuarto modelo, el de los buscadores académicos, como Google Scholar o Semantic Scholar, que no ofrecen filtros o herramientas de análisis que permitan determinar su cobertura lingüística.

## Metodología

Para realizar este análisis, recuperamos la totalidad de la producción indexada en cada una de las bases de datos seleccionadas, abarcando desde el inicio de sus registros hasta el año 2025 (inclusive). La extracción de los datos se llevó a cabo el 3 de diciembre de 2025, utilizando las herramientas de búsqueda avanzada y los filtros de análisis propios de cada plataforma para obtener la distribución por idiomas. Una vez obtenidos los datos brutos, realizamos una labor de limpieza y armonización: unificamos las denominaciones de las lenguas —que variaban ligeramente entre fuentes— y descartamos aquellos registros que no tenían idioma asignado o figuraban como 'desconocido' (unknown o unspecified). El resultado final se presenta en una tabla con los 20 idiomas principales, ordenados según su frecuencia en OpenAlex, criterio que hemos elegido por ser esta la base de datos con mayor volumen total de registros.

Hemos tenido que llegar a una definición operativa de la cobertura lingüística que es necesario explicar. Dado que es posible asignar múltiples idiomas a un mismo registro, resulta difícil aislar con precisión cuántos documentos únicos tienen disponible este dato. Por ello, hemos optado por una aproximación: calculamos el volumen de datos disponibles sumando todas las etiquetas de idioma encontradas, aunque esto implique contar varias veces los trabajos multilingües. Somos conscientes de que este método tiende a ofrecer una visión optimista —o un 'techo máximo'— de la cobertura, pero resulta útil para comparar la riqueza de metadatos entre las distintas plataformas.

## Resultados

### Complejidad del campo "idioma"

Un primer elemento importante que analizar es la disponibilidad efectiva de información sobre el idioma, que no es igual entre plataformas. Scopus muestra un 99,7% estimado de completitud

del campo, mientras que Web of Science alcanza un 93,6%. Es importante considerar que nuestra consulta en WoS incluía todas sus bases de datos, incluyendo "Research Commons", que incorpora registros de fuentes externas como OpenAlex. SciLit se sitúa en un nivel intermedio con un 97%, mientras que OpenAlex y OpenAIRE muestran resultados muy diferentes: OpenAlex alcanza un máximo teórico del 99% y OpenAIRE tan solo un 67,3%.

La explicación más probable a esta diferencia es el diferente enfoque de ambas bases de datos frente a los datos. Mientras que OpenAlex tiene un enfoque proactivo, en el que intenta mejorar los datos utilizando aproximaciones algorítmicas, OpenAIRE actúa más como "fedatario" de los datos. Su objetivo es por tanto enlazar e integrar los datos de terceros en un inmenso grafo, más que intervenir de manera sustancial en los registros proporcionados por los proveedores de datos originales. Esto implica que la calidad de los datos será tan buena como lo era originalmente (signifique lo que signifique esto), pero también implica una cierta renuncia a refinarlos. Por otra parte, los datos de OpenAlex se refieren más a la exhaustividad que a la precisión, y ya hemos visto que tienen margen de mejora.

<b>Idioma</b>	<b>Web of Science</b>	<b>Scopus</b>	<b>SciLit</b>	<b>OpenAlex</b>	<b>OpenAIRE</b>
<b>English</b>	140,822,401	91,495,628	153,400,000	200,880,947	107,920,697
<b>German</b>	3,379,142	2,731,741	7,740,000	10,462,392	6,021,551
<b>Japanese</b>	600,942	769,418	629,600	10,191,314	3,109,462
<b>Spanish</b>	2,915,968	1,047,277	2,770,000	9,455,980	2,764,853
<b>French</b>	3,054,833	1,552,877	3,530,000	8,208,523	4,222,014
<b>Chinese</b>	1,328,103	2,743,427	20,800	4,828,441	382,000
<b>Portuguese</b>	1,408,325	384,029	2,340,000	4,587,827	1,101,307
<b>Korean</b>	1,661,793	86,371	211,000	3,212,915	76,000
<b>Indonesian</b>	634,525	6,392	1,160,000	3,116,435	345,000
<b>Russian</b>	1,752,718	1,259,677	882,800	3,003,096	3,524,023
<b>Italian</b>	869,036	561,927	652,300	1,765,942	2,091,441
<b>Polish</b>	395,327	308,932	291,900	1,417,403	694,021
<b>Dutch</b>	255,996	158,542	325,100	1,128,829	693,536
<b>Turkish</b>	358,476	94,352	342,100	939,228	1,424,968
<b>Arabic</b>	22,226	9,400	537,500	760,477	164,000
<b>Swedish</b>	206,569	75,586	73,000	603,778	228,000
<b>Ukrainian</b>	46,617	21,000	335,800	563,000	331,000
<b>Persian</b>	5,784	12,000	184,300	525,000	33,000
<b>Czech</b>	228,580	160,164	95,000	415,000	503,322
<b>Norwegian</b>	108,395	19,000	93,600	346,000	243,000
<b>Totales</b>	171,964,966	104,391,065	181,832,574	272,400,000	205,941,178

Tabla 1: Top 20 idiomas en las diferentes bases de datos seleccionadas



Los datos disponibles en esta primera tabla tienen una representatividad muy elevada, con porcentajes de entre el 98 y el 99,5% del total de todas las etiquetas de idioma en las respectivas bases de datos.

### Rasgos generales de la cobertura lingüística

El inglés representa la mayor proporción de documentos en todas las bases de datos analizadas, aunque con variaciones significativas. En Scopus, el inglés constituye el 87,9% del total de asignaciones de idioma, mientras que en Web of Science se sitúa en el 87,5%, ligeramente por encima de SciLit, con en el 87%. Las bases de datos libres tienen porcentajes visiblemente inferiores, un 77.9% para OpenAIRE y un 74.5% para OpenAlex, que se sitúa como la base de datos con mayor porcentaje de literatura escrita en idiomas diferentes al inglés.

A parte de este rasgo notable de la cobertura lingüística de las diferentes bases de datos hay algunos otros hallazgos interesantes que se pueden ver mejor agrupando los idiomas por zonas lingüísticas, como se puede ver en la tabla 2.

Zona lingüística	Web of Sci.	Scopus	SciLit	OpenAlex	OpenAIRE
<b>Inglés</b>	87.51%	87.91%	86.95%	74.47%	77.91%
<b>Idiomas asiáticos</b>	2.63%	3.46%	1.15%	7.91%	2.82%
<b>Idiomas europeos</b>	6.62%	6.67%	8.14%	10.70%	14.42%
<b>Iberoamérica (*)</b>	2.69%	1.38%	2.90%	5.21%	2.79%
<b>Medio Oriente y N. de África</b>	0.02%	0.02%	0.41%	0.48%	0.14%

Tabla 2: distribución de los idiomas por zonas lingüísticas. (\*) Incluye Español y Portugués.

Todas las bases de datos tienen como segundo foco de atención los idiomas europeos (excluyendo el inglés). La diferencia entre este grupo y el siguiente en relevancia (los idiomas asiáticos) es grande para Scopus y Web of Science, pero enorme para SciLit y OpenAIRE. El tercer grupo, formado por el Español y el Brasileño (idiomas de Iberoamérica) es un poco conflictivo, porque en realidad está a caballo entre Europa y América. Tiene una presencia muy destacada en OpenAlex, y porcentajes mucho más modestos en el resto de las bases de datos. Es destacable el escaso papel que juega en Scopus en relación con las demás. El resto de los idiomas principales recogidos en esta tabla son plenamente marginales, con porcentajes entre el 0,02 y el 0,485.

### Lenguas europeas occidentales: variaciones significativas

Las lenguas asiáticas protagonizan las brechas de cobertura más extremas entre plataformas. El contraste es abismal en casos como el japonés o el coreano, donde OpenAlex multiplica exponencialmente los registros de las bases tradicionales: el japonés salta de los 600.000 documentos en WoS/Scopus a más de 10 millones, y el coreano pasa de 86.000 (Scopus) a 3,2 millones. El chino sigue una tendencia similar, oscilando entre el millón de WoS y los casi cinco millones de OpenAlex, aunque aquí destaca la anomalía de SciLit (apenas 20.800 registros), probablemente debida a fallos técnicos en la ingesta de caracteres no latinos.

El caso del indonesio merece una mención aparte por la magnitud de la disparidad. Mientras Scopus apenas recoge 6.000 documentos y WoS unos 635.000, OpenAlex aflora 3,1

millones de registros. Estas diferencias, que en ocasiones superan varios órdenes de magnitud, no son aleatorias: ilustran dos modelos de "visibilidad" opuestos. Mientras las bases tradicionales priorizan revistas con circuitos de revisión internacionalizados, las nuevas infraestructuras capturan la inmensa producción de los sistemas universitarios locales —muy activos en Indonesia—, ofreciendo una imagen mucho más fiel de la actividad científica real de estas regiones.

### Lenguas asiáticas: perfiles de cobertura diversos

Las lenguas asiáticas protagonizan las brechas de cobertura más extremas entre plataformas. El contraste es abismal en casos como el japonés o el coreano, donde OpenAlex multiplica exponencialmente los registros de las bases tradicionales: el japonés salta de los 600.000 documentos en WoS/Scopus a más de 10 millones, y el coreano pasa de 86.000 (Scopus) a 3,2 millones. El chino sigue una tendencia similar, oscilando entre el millón de WoS y los casi cinco millones de OpenAlex, aunque aquí destaca la anomalía de SciLit (apenas 20.800 registros), probablemente debida a fallos técnicos en la ingesta de caracteres no latinos.

El caso del indonesio merece una mención aparte por la magnitud de la disparidad. Mientras Scopus apenas recoge 6.000 documentos y WoS unos 635.000, OpenAlex aflora 3,1 millones de registros. Estas diferencias, que en ocasiones superan varios órdenes de magnitud, no son aleatorias: ilustran dos modelos de "visibilidad" opuestos. Mientras las bases tradicionales priorizan revistas con circuitos de revisión internacionalizados, las nuevas infraestructuras capturan la inmensa producción de los sistemas universitarios locales —muy activos en Indonesia—, ofreciendo una imagen mucho más fiel de la actividad científica real de estas regiones.

### Otras lenguas: patrones de variabilidad

Este patrón de variabilidad se replica, con matices, en otras regiones lingüísticas. En Oriente Medio, lenguas como el árabe o el persa, que en WoS y Scopus tienen una presencia testimonial (a menudo por debajo de los 10.000 documentos), alcanzan en OpenAlex volúmenes de cientos de miles (760.000 y 525.000 respectivamente), revelando un ecosistema científico regional que permanecía oculto.

En Europa del Este, las variaciones son significativas, aunque menos extremas. El ucraniano, por ejemplo, se multiplica por diez al pasar de WoS a OpenAlex. El ruso, por su parte, encuentra su "techo" en OpenAIRE (3,5 millones frente al 1,3 de Scopus), una cifra impulsada seguramente por los acuerdos de cosecha con repositorios de la región. Finalmente, incluso en lenguas de países con sistemas científicos consolidados y pequeños, como el neerlandés, el sueco o el noruego, se observan "inflaciones" considerables en las nuevas bases: el neerlandés, por ejemplo, escala de 158.000 registros en Scopus a más de un millón en OpenAlex, demostrando que la infrarrepresentación en las bases clásicas no es exclusiva del Sur Global.



## Conclusiones

El análisis comparativo de la cobertura lingüística en cinco bases de datos bibliográficas revela diferencias sustanciales en la representación de diversas lenguas. Estas diferencias no son meramente técnicas, sino que reflejan distintas filosofías de indexación, diferentes modelos de negocio, y prioridades específicas en cuanto a tipos de publicaciones y regiones geográficas cubiertas.

Las bases de datos tradicionales por suscripción tienden a concentrarse en publicaciones con mayor proyección internacional, lo que en la práctica se traduce en una mayor proporción de contenido en inglés y en menor representación de publicaciones de circulación principalmente local o regional. Las bases de acceso abierto, por su parte, suelen incorporar volúmenes mayores de contenido de diversas fuentes, incluyendo repositorios institucionales y publicaciones locales, lo que resulta en mayor diversidad lingüística pero también en mayor heterogeneidad en la calidad y completitud de los metadatos.

Ninguna aproximación es intrínsecamente superior: cada una responde a necesidades y contextos diferentes. Para análisis centrados en la ciencia de mayor impacto internacional, las bases tradicionales pueden ofrecer colecciones más cuidadas, con niveles de calidad y completitud exigibles para algunos de los análisis bibliométricos. Para estudios que buscan capturar la diversidad de la producción científica global, incluyendo publicaciones de circulación local, las bases abiertas ofrecen coberturas más amplias. En este caso destaca claramente OpenAlex. Para análisis centrados en producción europea depositada en repositorios, OpenAIRE presenta ventajas específicas.

Los profesionales de la información científica deben seleccionar las fuentes de datos considerando los objetivos específicos de sus análisis, las limitaciones metodológicas de cada plataforma, y la necesidad de transparencia sobre las decisiones tomadas. La disponibilidad de múltiples fuentes con características diferentes no simplifica el trabajo bibliométrico, pero sí ofrece la posibilidad de aproximaciones más matizadas y conscientes de las limitaciones inherentes a cualquier sistema de medición de la actividad científica.

## Referencias

- Asubiaro, T., Onaolapo, S., & Mills, D. (2024). Regional disparities in Web of Science and Scopus journal coverage. *Scientometrics*, 129(3), 1469-1491.  
<https://doi.org/10.1007/s11192-024-04948-x>
- Céspedes, L., Kozłowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimi, S., Ayeni, P., Filali, S., Li, B., & Larivière, V. (2025). Evaluating the linguistic coverage of OpenAlex: An assessment of metadata accuracy and completeness. *Journal of the Association for Information Science and Technology*, 76(6), 884–895.  
<https://doi.org/10.1002/asi.24979>
- Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, 5(1), 31-49.  
[https://doi.org/10.1162/qss\\_a\\_00286](https://doi.org/10.1162/qss_a_00286)
- Holmberg Sainte-Marie, M., Kozłowski, D., Céspedes, L., & Larivière, V. (2025). Sorting the Babble in Babel: Assessing the performance of language-detection algorithms on the OpenAlex database. arXiv. <https://doi.org/10.48550/arXiv.2502.03627>

- Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, 126(6), 5113–5142. <https://doi.org/10.1007/s11192-021-03948-5>
- Vera-Baceta, M. A., Thelwall, M., & Kousha, K. (2019). Web of Science and Scopus language coverage. *Scientometrics*, 121(3), 1803-1813. <https://doi.org/10.1007/s11192-019-03264-z>
- Visser, M., Van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative science studies*, 2(1), 20-41. [https://doi.org/10.1162/qss\\_a\\_00175](https://doi.org/10.1162/qss_a_00175)

### Sobre el autor



#### RODRIGO SÁNCHEZ-JIMÉNEZ

ID ORCID: <https://orcid.org/0000-0002-3685-7060>

Profesor Titular en la Universidad Complutense de Madrid, donde ejerce la docencia desde hace más de veinte años. Doctor en Documentación por la UCM en 2006, se ha especializado en cienciometría. Su actividad investigadora se centra en el análisis cuantitativo de la actividad científica, abordando líneas como los costes de publicación (APCs), las brechas de género y la evolución de las tesis doctorales. Asimismo, cuenta con una amplia experiencia en gestión editorial y colabora activamente con el grupo SCImago en el desarrollo de nuevos indicadores y el estudio de fuentes de datos abiertas.