Abordando el problema de la cobertura en las bases de datos bibliográficas: propuesta de una herramienta para la comparación del contenido de bases de datos bibliográficas

Robles-Belmont E. y López Bonifacio, J.G.

Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. MMSS, Apdo. Postal 20-126, 04510, Coyoacán, Cd. de México, México. eduardo.robles@iimas.unam.mx y jose.lopezb@iimas.unam.mx

Introducción

En las actividades de evaluación de la ciencia y los estudios sobre la producción científica las bases de datos bibliográficas ocupan un lugar importante por ser las fuentes de información. Cuando estamos frente a la elección de cuál base de datos emplear, varios son los problemas que enfrentamos: la falta de homogeneización de las bases de datos (Vuotto et al., 2020), las categorías ambiguas de las clasificaciones (Wang & Waltman, 2016), la limitada cobertura para algunos campos y disciplinas (Mongeon & Paul-Hus, 2016), así como la subrepresentación geográfica (Rafols et al., 2016).

En este trabajo nos interesamos en explorar y analizar la cobertura de la base de datos de OpenAlex, con respecto a otras bases de datos comerciales (Web of Science y Scopus). Los datos que usamos en este ejercicio conciernen a dos objetos de estudio. El primero es el campo emergente de Climate Change Eutrophication Lakes (CCEL) y el segundo concierne a un grupo de investigadoras e investigadores de un departamento de investigación en ciencias sociales y humanidades de la UNAM. Estos dos ejemplos los hemos seleccionado por tratarse de ejemplos ilustrativos de dos usos comunes de datos bibliográficos: estudios de ciencias emergentes y evaluación académica. Como se trata de un ejercicio que busca comparar diferentes bases de datos, el abordaje se basa en la teoría de conjuntos tradicional, para la identificación de coincidencias y divergencias entre las bases de datos comparadas.

Método

En la Imagen 1, se muestra la estrategia metodológica ha consistido en a. descargar los datos de las tres fuentes, b. tratar los datos para eliminar duplicados (a partir de los títulos de los documentos), c. normalizar el texto del campo a comparar, d. comparar las bases de datos (Cuadro 1 y Cuadro 2) y visualizar resultados. El tratamiento de los datos y la comparación lo hemos realizado con un código en Python que hemos escrito y será liberado en el momento de la publicación del estudio.

Imagen 1. Estrategia metodológica.

- Intersección SC ∩ WS ∩ OA: Documentos compartidos por las tres bases.
- Intersección OA ∩ SC - WS: Documentos en OA y SC pero no en WS.
- Intersección OA ∩ WS - SC: Documentos en OA y WS pero no en SC.
- Intersección SC ∩ WS - OA: Documentos en SC y WS pero no en OA.
- Complemento OA - (SC ∪ WS): Documentos solo en OpenAlex.
- Complemento SC - (OA ∪ WS): Documentos solo en Scopus.
- Complemento WS - (OA ∪ SC): Documentos solo en Web of Science.

Cuadro 1. Operaciones para calcular la intersección de las 3 bases de datos.

- Intersección de los tres conjuntos (SC ∩ WS ∩ OA).
- Intersecciones parciales (OA ∩ SC - WS, OA ∩ WS - SC, SC ∩ WS - OA).
- Complementos de cada base (OA - (SC ∪ WS), SC - (OA ∪ WS), WS - (OA ∪ SC)).

Cuadro 2. Operaciones para calcular las intersecciones y diferencias entre las 3 bases de datos.

Resultados

En este ejercicio se han identificado documentos duplicados para las bases de datos OA y SC para ambos casos. Siendo mayores los duplicados en OA confirma que esta base de datos debe mejorar su sistema de catalogación de documentos.

En los resultados mostramos en la Imagen 2, la visualización de la comparación para CCEL muestra que SC presenta la cobertura más amplia seguido de WS y OA. Sin embargo, en términos de registros únicos, SC tiene la mayor cobertura en estos registros seguido de OA y WS. Esto indica que para este campo emergente las bases de datos privadas cuenta con una mayor cobertura, y para los registros únicos lo que incluye OA no es negligente. Esto puede deberse a las dinámicas propias de las disciplinas que aportan a este campo CCEL y sus prácticas en la publicación. Por otro lado, hemos observado los histogramas de los documentos en cada una de las bases de datos y su comportamiento es similar, lo que refuerza la idea anterior sobre las prácticas de publicación (por cuestión del espacio en el cartel no hemos incluido las gráficas, pero pueden solicitar a los autores).

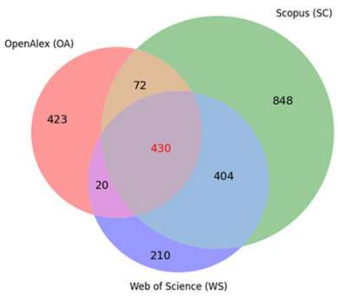

Imagen 2. Comparación de OpenAlex, Wos y Scopus para CCEL

Base de datos	Referencias	% total	Referencias únicas	% únicos
OA	945	39.26%	423	17.57%
SC	1754	72.87%	848	35.23%
WS	1064	44.20%	210	8.72%
Total	2407			

Tabla 1. Distribuciones de cobertura en las bases de datos

Por otro lado, sobre el caso del grupo del departamento de investigación de la UNAM, los resultados se muestran en la Imagen 3, la visualización de la comparación muestra que OA presenta la cobertura más amplia seguido de WS y SC es la de menor cobertura. En los registros únicos, OA tiene la mayor cobertura en estos registros seguido de WS y SC. Esto indica que este grupo de las ciencias sociales y humanidades esta subrepresentado en las bases de datos privadas, y para los registros únicos es OA con mayor cobertura. Al observar los histogramas de las publicaciones, es la de OA la que presenta un crecimiento estable.

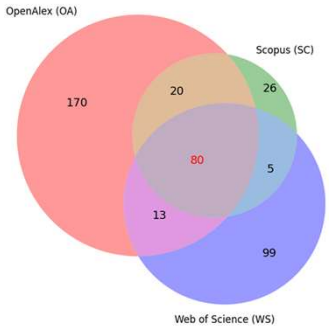

Imagen 3. Comparación de OpenAlex, Wos y Scopus para el departamento de investigación del IIMAS

Base de datos	Referencias	% total	Referencias únicas	% únicos
OA	283	46.32%	170	27.82%
SC	131	21.44%	26	4.26%
WS	197	32.24%	99	16.20%
Total	611			

Tabla 2. Tipo de documento para el departamento de investigación del IIMAS en OpenAlex (OA)

Conclusiones

En conjunto, los resultados sugieren que la base de datos OpenAlex ofrece una alternativa viable frente a las bases comerciales, aunque aún presenta inconsistencias en la identificación de duplicados y en la cobertura disciplinaria. Por otro lado, los resultados confirman que la cobertura de los datos comerciales se inclina hacia las ciencias naturales sobre el detrimento de las ciencias sociales y humanidades. Para estas últimas, OpenAlex presenta mayor cobertura lo que permite avanzar hacia mapeos y evaluaciones pertinentes.

Referencias

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics, 106(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Rafols, I., Chavarro, D., & Ciarli, T. (2016). Under-representation of research in the global south Biases in mainstream journal indexing systems.

Vuotto, A., Di Césare, V., & Pallotta, N. (2020). Fortalezas y debilidades de las principales bases de datos de información científica desde una perspectiva bibliométrica. Palabra Clave, 10(1). https://doi.org/10.24215/18539912e101

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. Journal of Informetrics, 10(2), 347–364. https://doi.org/10.1016/j.joi.2016.02.003

Agradecimientos

Este estudio ha sido desarrollado en el marco del proyecto PAPIIT IN302623 “Indicadores de la ciencia y la tecnología en el contexto de la Ciencia Abierta”.