Datos bibliométricos para las Ciencias Sociales y las Humanidades: un método para el acopio, validación y análisis con herramientas de acceso gratuito

Silvia Eunice Gutiérrez de la Torre

Biblioteca «Daniel Cosío Villegas», El Colegio de México

segutierrez@colmex.mx

José Valentín Ortiz Reyes

Biblioteca «Daniel Cosío Villegas», El Colegio de México

jvortiz@colmex.mx

Jonathan Israel Escobar Farfán

Biblioteca «Daniel Cosío Villegas», El Colegio de México

jescobar@colmex.mx

Tomás Bocanegra Esqueda

Biblioteca «Daniel Cosío Villegas», El Colegio de México

tbocanegra@colmex.mx

Víctor Cid Carmona

Biblioteca «Daniel Cosío Villegas», El Colegio de México

vjcid@colmex.mx

Claudia Escobar Vallarta

Biblioteca «Daniel Cosío Villegas», El Colegio de México

cescobar@colmex.mx

María Lourdes Quiroa Herrera

Biblioteca «Daniel Cosío Villegas», El Colegio de México

lourdes_verd@hotmail.com

CAMELIA ROMERO MILLÁN

Biblioteca «Daniel Cosío Villegas», El Colegio de México

cromero@colmex.mx

RESUMEN

La baja cobertura en los índices globales más importantes, como Web of Science (WOS) y Scopus, de la investigación en español de las Ciencias Sociales y las Humanidades (CSyH) obligan, a quienes analizan las citas que se hacen a la producción académica de sus comunidades, a explorar alternativas para recopilar datos de mayor alcance, confiables y validados. En este documento se describe un método que emplea Google Académico (GA) como principal fuente de información para el análisis de citación de los artículos publicados en revistas arbitradas por la planta académica de El Colegio de México entre los años 2012-2020. A partir de una base de datos compilada y validada desde 2016, por personal académico de la Biblioteca «Daniel Cosío Villegas»[1], se ofrece evidencia sobre el alcance de GA para el análisis de citación, destacando sus ventajas y limitaciones. Se describe el proceso de acopio, análisis e interpretación de los datos, el cual se realiza exclusivamente con software gratuito y de libre acceso (Google Drive, Zotero, R y GA). Con esto nos proponemos compartir un método que pueda ser replicado en contextos similares.

PALABRAS CLAVE

Análisis de citas; servicios bibliométricos; herramientas de acceso gratuito. Citation Analysis, Bibliometric Services, Free Access Tools.

[1] Tomás Bocanegra, Víctor Cid, Claudia Escobar, Israel Escobar, Silvia Gutiérrez, Valentín Ortiz, Carolina Palacios, Lourdes Quiroa, Camelia Romero, Arón Sánchez y en ediciones anteriores: Mariana Córdoba, Máximo Domínguez, Lourdes Guerrero, José Manuel Mórales.

Introducción

En los últimos años el uso de métricas para valorar la producción académica ha tomado mayor relevancia. Por esta razón, la reflexión sobre las fuentes y métodos empleados para la obtención de datos se ha centrado en la búsqueda de parámetros más justos y apegados a las prácticas de publicación de cada disciplina. Para el caso del trabajo que se realiza en la Biblioteca «Daniel Cosío Villegas» (BDCV) de El Colegio de México (COLMEX), esta reflexión se ha centrado, por la naturaleza de los programas de formación y líneas de investigación que se desarrollan, en las Ciencias Sociales y las Humanidades (CSyH), tarea que derivó inicialmente del servicio de análisis de citas que se ofrece a la planta académica.

El COLMEX es una institución pública dedicada a la formación e investigación en CSyH, a nivel licenciatura y, principalmente, a nivel de posgrado. Cuenta con ocho centros de estudios: Centro de Estudios de Asia y África (CEAA), que cuenta con 26 personas en su planta académica; Centro de Estudios Demográficos, Urbanos y Ambientales (CEDUA), 30 personas; Centro de Estudios Económicos (CEE), 16; Centro de Estudios Históricos (CEH), 25; Centro de Estudios Internacionales (CEI), 24; Centro de Estudios Lingüísticos y Literarios (CELL), 30; Centro de Estudios Sociológicos (CES), 22 y el Programa de Estudios Interdisciplinarios (PEI), 50, que se integra por el personal académico jubilado.

Durante 2016 se analizaron los indicadores que ofrecen distintas fuentes de información sobre la citación de artículos publicados por la planta académica. A partir de aquel ejercicio se identificó que ninguna base de datos ofrece información confiable y actualizada para analizar las citas a nivel institucional. Al realizar el ejercicio de búsqueda con el filtro de filiación institucional en Scopus, se ubicaron 353 documentos publicados entre 2012 y 2016. Sin embargo, al revisar puntualmente los documentos, se ubicaron artículos cuyas temáticas estaban asociadas con ciencias duras. El error que identificamos fue que había autores que pertenecían a El Colegio Nacional, pero que para Scopus tenían filiación con el COLMEX. Además de esto, identificamos que los

números o indicadores que disponen las bases de datos están limitados al universo de documentos que albergan, es decir, no consideran otras fuentes de información para generar datos y los que ofrecen son limitados por su propio alcance.

Como adelantamos, los ejercicios de análisis de citas evidenciaron problemas de cobertura, pues el número de citas que se ubicaban en WOS o Scopus son mínimas, en comparación con las identificadas por medio de GA. La anotación al margen en este caso es que hablamos de textos relativos al estudio de las CSyH escritos en español, condiciones que restringen su representatividad en ambas bases de datos de pago (ver PLAZA *et al.*, 2018). En el análisis que realizamos en 2019, con la base de datos Scopus, sobre la distribución por idioma de artículos publicados, entre 2014 y 2018, en el área de Ciencias Sociales, encontramos que el 89 % fueron escritos en inglés y un 2,54 % en español. En un escenario similar, ubicamos en WOS, que el 87,12 % de los artículos publicados en el área de Ciencias Sociales fueron escritos en inglés y tan sólo un 3,11 % en español (*Las revistas científicas y el español*, 2019). Con todo, a pesar de ser la herramienta disponible más adecuada, está lejos de ser perfecta. En un análisis reciente se ha probado que la visibilidad de los artículos en español está significativamente afectada por los algoritmos de relevancia de GA, lo cual afecta la posibilidad de cosechar citas, pues a menor facilidad de encuentro, menor posibilidad de ser leído o citado (ROVIRA, CODINA, & LOPEZOSA, 2021).

Estudios como los de HARZING y ALAKANGAS (2016) han discutido la necesidad de realizar estudios longitudinales para tener mayor evidencia sobre la cobertura o alcance de Scopus, WOS y GA, pues el uso *per se* de estos recursos puede determinar las métricas o resultados obtenidos y, con ello, modificar las conclusiones que se puedan hacer sobre los datos. Parte de su argumentación apunta a que GA ofrece una cobertura más amplia, y, por lo tanto, indicadores más altos que otras fuentes de información. Sin embargo, requiere de un trabajo adicional de control de calidad, pues rastrea cualquier información que esté disponible en la red. Otro de los inconvenientes que apuntan estas autoras es el gran número de documentos duplicados, que

varían en detalles nimios de sus respectivos registros. El problema con el contenido que ofrece GA, de acuerdo con Gingras (2016) es que puede incluir documentos provenientes de publicaciones académicas, pero también toma documentos de páginas personales, documentos de trabajo, borradores, pruebas de imprenta. A esto se suma el riesgo, al que volveremos más delante, de que los documentos citantes pueden aparecer o desaparecer, modificando así la validación de los indicadores, calculados por el propio recurso.

Para Martín Martín, Orduna Malea, Thelwall y Delgado López Cózar (2018), GA encuentra más citas que WOS Core Collection y Scopus en todas las áreas temáticas. A partir de un conjunto de datos de distintas disciplinas descubrieron que casi todas las citas encontradas por WOS (95 %) y Scopus (92 %) fueron también encontradas por GA. A esto se le suma que esta última base encontró una cantidad sustancial de citas únicas que no fueron ubicadas por las otras bases de datos. En las áreas de CSyH las citas únicas de GA, según este estudio, superan el 50 % de todas las citas del área.

Una de las herramientas que se valoró en 2016 para el acopio de información fue Publish or Perish, un programa de acceso gratuito desarrollado por Anne-Wil Harzing, que permite búsquedas avanzadas para el acopio y exportación de reportes numéricos de las citas provenientes de fuentes como GA y Microsoft Academic Research. Esta herramienta ha sido útil para otros reportes en los que no se hace la revisión manual de las citas pues, si bien la última versión permite recuperar los documentos citantes de cada ítem, este proceso requiere correr la búsqueda por cada artículo y hacer después el cotejo de la veracidad de la cita, tratamiento que describiremos más adelante.

Frente al requerimiento institucional de ofrecer evidencias sobre las citas a la producción académica del COLMEX, se delineó una estrategia para obtener información confiable sobre las citas a artículos publicados por la planta académica en revistas propias y ajenas, desde 2012 hasta el 2016. Estos son algunos de los elementos que se consideraron para dar forma a la base de datos que el personal académico, con el apoyo de personal auxiliar, ha actualizado y mantenido en los últimos cinco años y la cual ha servido como

principal insumo para la elaboración de reportes anuales[2]. A continuación, se describe el método empleado para el acopio de datos, así como los criterios que se aplican en su validación. Se muestra en detalle el procedimiento de sistematización de los datos con el uso de herramientas como Google Drive, Zotero, R en RStudio y OpenRefine. Al final se exponen algunas consideraciones a tomar en cuenta en la aplicación de este método.

Descripción del método

Anualmente se analizan las citas reportadas por GA para los artículos publicados por profesores-investigadores del COLMEX en revistas académicas. En este análisis se buscan citas únicamente a los artículos que son reportados, por la propia comunidad académica, en el Currículum Electrónico COLMEX de cada profesor-investigador, y no se consideran las citas a libros, capítulos de libros y otros documentos. Estos registros, que se realizan de forma personal, no están exentos de errores de captura, pero para el estudio se normalizan los campos de año y lugar de publicación.

Se validan únicamente citas asentadas en publicaciones del periodo comprendido entre 2012 y el año inmediato anterior al del reporte anual. Por ejemplo, para el reporte del 2021, se analizaron sólo artículos que citan la producción COLMEX entre 2012 y 2020. Se excluyen las que se consideran citas «no válidas», sobre lo cual se ahondará más adelante. Para el proceso de validación se aplican los siguientes criterios de inclusión para seleccionar los trabajos citantes:

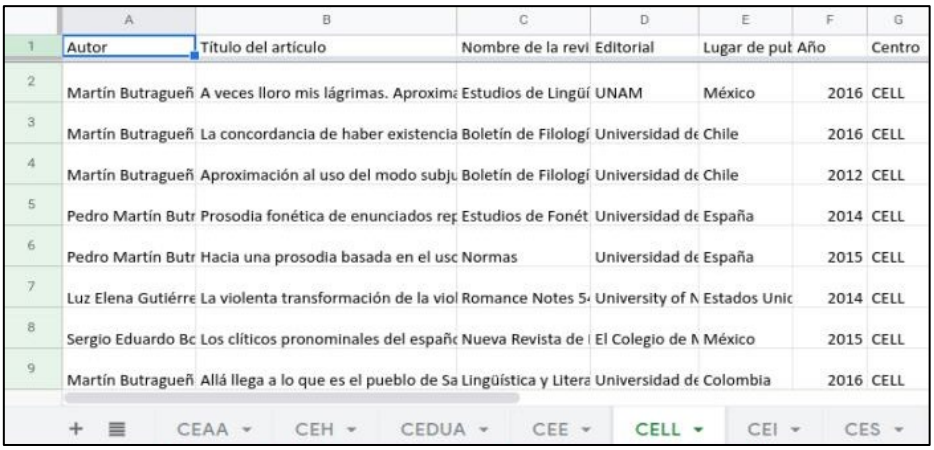
- Se validan sólo aquellos a los que se tiene acceso en texto completo y en cuya bibliografía se haya cotejado la cita al documento publicado por investigadores COLMEX.
- Se validan aquellos que no hayan sido escritos por la persona citada, es decir, que sean «autocitas».

[2] Estos informes están disponibles en: https://www.colmex.mx/es/documentacion-institucional

- Se consideran sólo publicaciones arbitradas. Así, se excluyen como citas válidas las hechas en: documentos de trabajo, ponencias, presentaciones y memorias de actos académicos publicadas sin ISBN o ISSN, o, en su defecto, que se encuentre en trámite. Si bien estos identificadores no aseguran que se trate de una publicación arbitrada, requieren un nivel de trámite superior a la simple subida de un archivo digital.
- Sólo se validan citas en tesis en las que la persona citada no haya tenido ninguna participación como tutor o lector.
- Se excluyen los documentos citantes publicados en el año en el que se realiza el análisis.

Procedimiento de acopio con Google Académico (GA) y Google Drive. En esta sección se describe la hoja de cálculo colaborativa, las búsquedas que realizan los compañeros auxiliares, y se detalla el uso que se da a las columnas de control.

La Bibliotecaria de Humanidades Digitales (BHD) crea en Google Drive una hoja de cálculo colaborativa para el año de análisis correspondiente, partiendo de una copia de la hoja de cálculo usada en el año anterior. Así se mantiene un registro histórico del análisis, y se aprovechan avances anteriores. En esta nueva hoja de cálculo se crea una pestaña para cada centro de estudios, y en cada pestaña se actualiza el listado de artículos. Cada artículo corresponde a un renglón de la hoja de cálculo y es descrito con las siguientes columnas: *Autor, Título del artículo, Nombre de la revista en la que fue publicado, Editorial, Lugar de publicación, Año, y Centro* (ver FIGURA 1).



	A	B	C	D	E	F	G
1	Autor	Título del artículo	Nombre de la revi	Editorial	Lugar de pub	Año	Centro
2	Martín Butragueñ	A veces lloro mis lágrimas. Aproxima	Estudios de Lingüí	UNAM	México	2016	CELL
3	Martín Butragueñ	La concordancia de haber existencia	Boletín de Filologí	Universidad de	Chile	2016	CELL
4	Martín Butragueñ	Aproximación al uso del modo subju	Boletín de Filologí	Universidad de	Chile	2012	CELL
5	Pedro Martín Butr	Prosodia fonética de enunciados rep	Estudios de Fonét	Universidad de	España	2014	CELL
6	Pedro Martín Butr	Hacia una prosodia basada en el usc	Normas	Universidad de	España	2015	CELL
7	Luz Elena Gutiérre	La violenta transformación de la viol	Romance Notes 5	University of N	Estados Unic	2014	CELL
8	Sergio Eduardo Bc	Los clíticos pronominales del españ	Nueva Revista de	El Colegio de M	México	2015	CELL
9	Martín Butragueñ	Allá llega a lo que es el pueblo de Sa	Lingüística y Litera	Universidad de	Colombia	2016	CELL

FIGURA 1. COLUMNAS PARA DESCRIBIR ARTÍCULOS EN GOOGLE DRIVE

Además de las columnas anteriores, se incluye una columna *Liga GA*, en la que se registra la URL de resultados para la búsqueda de cada artículo en GA. Cuando GA no reporta resultados para la búsqueda de un artículo de la lista, se consigna la nota: *No reconocido.* Colaboradores auxiliares son responsables de correr las búsquedas siguiendo la URL correspondiente, y actualizarla, si es el caso, para artículos no reconocidos en años anteriores, o para aquellos que recién se integran a la lista.

En cada búsqueda, se coteja el número de citas reportadas para el artículo, y se consigna en una columna *Citas GA XXXX*, donde XXXX corresponde al año analizado. En otra columna llamada *Diferencias* se calcula la diferencia entre las citas reportadas por GA el año anterior y las del año analizado. Por ejemplo, en el estudio de 2020 se comparan las citas reportadas por GA ese año, con las reportadas en el 2019. En principio, solamente aquellos artículos que hayan presentado una diferencia en citas reportadas son los que tendrían nuevas citas para validar.

A continuación, la BHD actualiza las columnas de control (ver FIGURA 2) donde XXXX corresponde al año de análisis, es decir, aquel en el que se realiza el ejercicio. Estas columnas sirven para distinguir el número de citas válidas, desglosando el número de citas «no válidas» reportadas por GA. Las *Citas Zotero* son las citas validadas de acuerdo con los criterios mencionados anteriormente. La columna *Sin Acceso* registra la cantidad de citas que no fue posible verificar porque no se tuvo acceso al texto completo del documento

citante. En *Autocita*, se cuentan aquellas publicaciones citantes cuyo autor es el mismo del artículo citado. *No arbitrado* sirve para contabilizar las citas hechas en documentos sin revisión, como se explicó previamente. En la columna *Otro* se cuentan documentos que no pertenecen a ninguna de las categorías anteriores, por ejemplo, duplicados, errores, tesis dirigidas por el profesor investigador citado, o publicaciones del año en el que se realiza el análisis.

	Lugar de pub	Año	Centro	Citas GA 2019	Citas GA 2020	Diferencias	Citas Zotero 2021	Sin acceso 2021	Autocita 2021	No arbitrado 2021	Otro 2021	Comprobacion	Liga GA	Notas
2	México	2016	CELL	8	11	3	6	0	3	0	3	-1	https://scholar.google.com/scholar?cites=15556044187529	1 AC en Ha JIEF 2021:
3	Chile	2016	CELL	11	12	1	6	1	2	0	3	0	https://scholar.google.com/scholar?hl=es&as_sdt=2005&sc	JIEF 2020: JIEF 2021:
4	Chile	2012	CELL	21	25	4	19	0	3	1	2	0	https://scholar.google.com/scholar?cluster=452858605257	1 autocita JIEF 2021:
5	España	2014	CELL	15	22	7	7	2	9	0	4	0	https://scholar.google.com/scholar?cluster=2509717891332	1 AC en A JIEF 2021:
6	España	2015	CELL	7	10	3	8	0	2	0	1	-1	https://scholar.google.com/scholar?hl=es&as_sdt=0%2C5&	GS ya no lis JIEF 2021:
7	Estados Unic	2014	CELL	4	7	3	4	1	0	0	2	0	https://scholar.google.com/scholar?hl=es&as_sdt=0%2C5&	1 REP de La JIEF 2021:
8	México	2015	CELL	5	11	6	5	0	2	0	4	0	https://scholar.google.com/scholar?cluster=161426336412	1 AC en Ora JIEF 2021:
9	Colombia	2016	CELL	7	11	4	8	0	3	0	1	-1	https://scholar.google.com/scholar?cites=13083579098513	GS ya no in JIEF 2021:

FIGURA 2. COLUMNAS DE CONTROL EN GOOGLE DRIVE

Finalmente, la BHD integra una columna para *Comprobación* en la cual se calcula la diferencia entre la suma de las columnas de control y el número de posibles citas reportadas por GA para el año analizado. En principio, una diferencia de cero indicaría que se ha dado cuenta de todas las posibles citas reportadas por GA, y el número total de éstas se ha distribuido entre las columnas de control. Una columna para notas es utilizada para hacer las anotaciones que puedan ayudar, en ejercicios subsecuentes, a la mejor identificación de citas ya validadas o ya descartadas. Una vez preparada la hoja de cálculo colaborativa en Google Drive, se procede a la validación de citas y al registro de artículos citantes en Zotero.

Validación e integración de las citas en Zotero. En esta sección se describe la biblioteca compartida de Zotero, y se hacen precisiones sobre las citas a más de un artículo, y las coautorías. Los bibliógrafos asignados a cada centro de estudios son los encargados de validar las posibles citas reportadas por GA, y de compilar la información bibliográfica de los artículos citantes en una biblioteca compartida de Zotero

Como requisitos previos, cada bibliógrafo debe instalar un cliente de Zotero en su computadora de trabajo (STILLMAN, 2021). También debe instalar el complemento de Zotero para su navegador de internet, que le permite importar las referencias bibliográficas desde la lista de resultados de GA, o desde la base de datos donde se aloja el documento citante.

Para revisar las posibles citas a cada artículo, se pueden ordenar los datos en cada pestaña de centro por la columna *Diferencias*, para analizar aquellas que en principio tuvieron aumento de citas. Por cada artículo, se sigue la liga de GA con la lista de posibles documentos citantes, y de esta lista se validan las citas que cumplen los criterios establecidos, y se contabiliza cada cita «no válida» usando, en cada caso, las columnas de control descritas en la sección anterior.

Los datos bibliográficos de un documento citante que incluye una cita válida se integran a Zotero. Dicha biblioteca tiene la siguiente estructura (ver FIGURA 3). Esta estructura jerárquica permite concentrar la información por centro, autor y sus artículos citados. Las carpetas de cada artículo se nombran con una combinación del año de publicación del artículo y su título.

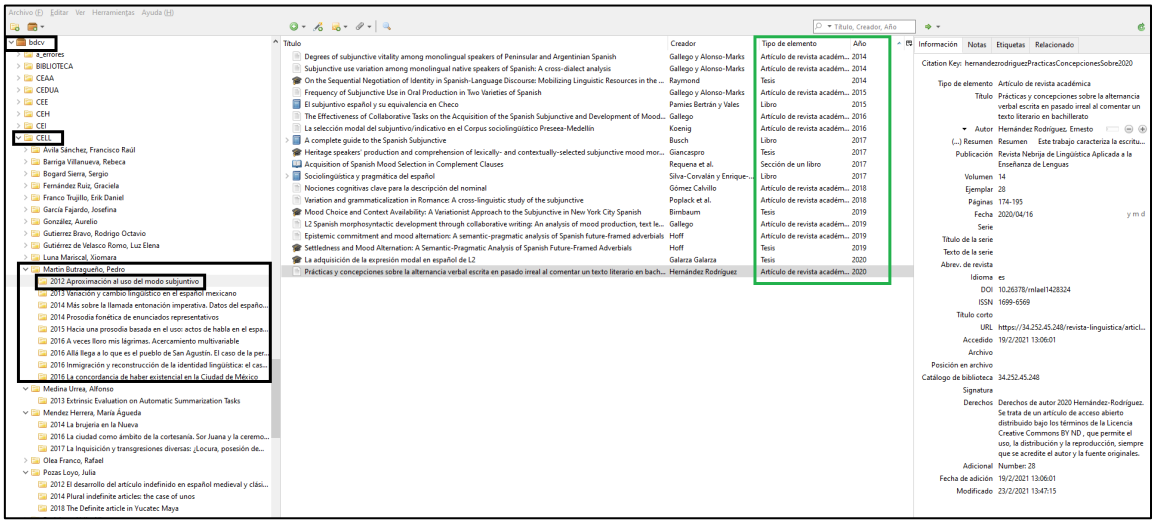


FIGURA 3. ESTRUCTURA JERÁRQUICA DE LA INFORMACIÓN CONCENTRADA EN ZOTERO

Para guardar las referencias de los artículos citantes, los bibliógrafos pueden, en el caso ideal, valerse del complemento de Zotero para su navegador

de internet, pero, a veces, es necesario ingresar la referencia manualmente. En cualquier caso, siempre es indispensable verificar que los datos de la referencia agregada estén completos y sean correctos. Se pone especial atención en el tipo de documento citante, el orden correcto del nombre y apellidos del autor, el título, año de publicación y editorial.

Conforme se avanza en la lista de los documentos citantes, el bibliógrafo actualiza la cuenta de citas validadas en la hoja colaborativa o, en su caso, en las otras columnas de control («autocita», «sin acceso», etc.).

Respecto al proceso de validación de documentos citantes a más de un artículo analizado, y de citas hechas a artículos en coautoría, cabe hacer las siguientes precisiones:

Si un documento cita a dos o más artículos de la lista, este documento debe ser duplicado e integrado a la carpeta de Zotero de cada autor/artículo citado. Es importante recalcar que en Zotero debe existir un registro del documento citante en cada carpeta, y no un solo registro asociado a dos o más carpetas. Ejemplo: el artículo Y de Suárez Méndez cita a Martín Butragueño (2012) y a Barriga Villanueva (2010). El artículo Y de Suárez Méndez se integra tanto a la carpeta del artículo de Martín Butragueño (2012) como a la de Barriga Villanueva (2010).

Si un documento cita un artículo de la lista en co-autoría, este documento debe ser integrado a la carpeta de Zotero de cada autor. Ejemplo: Martín Butragueño y Barriga Villanueva escriben un artículo X en 2007. Un artículo Y cita dicho texto. El artículo Y se integrará tanto a la carpeta del artículo X en la colección de Martín Butragueño, como en la de Barriga Villanueva. En la hoja de cálculo colaborativa, el artículo X se encuentra en dos renglones: uno corresponde a la autoría de Martín Butragueño y otro a la de Barriga Villanueva, y en ambas filas se contabiliza la cita hecha por el artículo Y.

Si un documento Y del autor A cita al artículo X de ese mismo autor, cuenta como autocita para el autor A, pero como cita validada para sus coautores, de ser el caso. Ejemplo: Martín Butragueño y Barriga Villanueva escriben un artículo X en 2007. En 2008 Martín Butragueño, en el artículo Y, cita al artículo

X; así entonces, la cita en el artículo Y cuenta como cita para Barriga Villanueva, pero como autocita para Martín Butragueño.

Integración de la información y análisis. En esta sección se describen las herramientas utilizadas para integrar los insumos y hacer el análisis sistemático de los datos. Una vez que se ha terminado la revisión de las citas reportadas por GA, la BHD crea una carpeta para el año de análisis y, en ésta, crea un proyecto de RStudio (RSTUDIO TEAM, s/f). Además, genera una búsqueda de todas las citas por cada centro en Zotero. Los bibliógrafos verifican que ese número coincida con la suma de la columna *Citas Zotero XXXX* de la pestaña correspondiente a su centro en la hoja de cálculo colaborativa. Éste es uno de los procesos más demandantes, pues requiere de una revisión exhaustiva de cada carpeta en caso de que el número de citas guardadas en Zotero no coincida con la suma de citas en la hoja de cálculo. La BHD también revisa que todos los artículos citantes tengan: año de publicación (éste no puede ser mayor al año de análisis) y tipo de documento (no puede contener tipos de documentos no válidos). A continuación, exporta la carpeta Zotero de artículos citantes de cada centro en formato CSV a la carpeta donde se guarda el proyecto de RStudio. Usando un script de R, conjunta los CSV en uno solo, al que se añade la columna *Centro*, para saber a qué centro pertenece el artículo citado.

Con la hoja de cálculo colaborativa, la BHD agrupa a todos los centros en una misma pestaña, recuperando las siguientes columnas: *Autor, Título del artículo, Nombre de la revista, Editorial, Lugar de publicación, Año, Centro, Citas GA XXXX, Diferencias, Citas Zotero XXXX, Sin acceso XXXX, Autocita XXXX, No arbitrado XXXX, Otro XXXX y Liga GA*. Este archivo es analizado en OpenRefine (HUYNH & MAZZOCCHI, 2021) para asegurar que haya consistencia en los datos. Se uniforman las variables *Año* y *Lugar de publicación*, utilizando los distintos métodos de *clustering* disponibles en esta herramienta y se editan manualmente las inconsistencias no identificadas con estos métodos. La hoja de cálculo así revisada, que lista todas los artículos publicados y el número de citas validadas, se guarda en la carpeta del proyecto de RStudio.

Con los archivos de datos producidos a partir de la biblioteca colaborativa en Zotero y la hoja de cálculo en Google Drive, y usando scripts de R, la BHD lleva a cabo los análisis necesarios para generar los datos que se reportan a la Presidencia del COLMEX.

Consideraciones respecto al método

El uso de GA para identificar citas implica un proceso de minuciosa revisión y validación de datos, que se afina año con año. En nuestra experiencia con este método, hemos detectado algunas complicaciones generales que requieren la debida consideración. Se presentan a continuación siete advertencias que el equipo de la BDCV ha distinguido a la luz del análisis de 1.236 artículos publicados entre 2012 y 2020, las 6.561 citas que fueron identificadas por GA, y el proceso que se requiere para afinar los resultados derivados de esta información.

Inconsistencias y verificación. La primera advertencia es que no debe subestimarse la cantidad de tiempo que se debe invertir en verificar y rectificar las inconsistencias. Una de ellas está asociada a los nombres de autores hispanoamericanos. El algoritmo de GA parece lo suficientemente sofisticado para agrupar las variaciones de los nombres de un determinado autor y recuperar citas a pesar de las variantes con que se consignan los nombres en los documentos citantes. En contraste, el complemento del navegador para importar citas a Zotero frecuentemente importa los nombres hispanoamericanos alterando el orden de los elementos, y consigna el apellido materno como apellido paterno, y éste como un segundo o tercer nombre de pila. El uso de guiones para unir los apellidos maternos y paternos en algunas publicaciones mitiga este problema, pero es una práctica muy poco difundida aún. A menudo se requiere un trabajo de verificación y edición de los nombres de autor en Zotero.

Otra inconsistencia tiene que ver con los títulos repetidos o paralelos. Un autor citado puede presentar dos trabajos con un mismo título, o con títulos casi idénticos, que a menudo corresponden a distintas etapas de una misma

investigación, o retoman algún trabajo previo, como una tesis doctoral. En estos casos, los bibliógrafos pueden invertir tiempo considerable para identificar el trabajo citado. También ocurre que GA reporta como citante a un mismo artículo tantas veces como títulos paralelos encuentre. El ejemplo típico es el de los artículos indizados en SciELO, que a menudo tienen un título en dos o tres idiomas. En consecuencia, GA reporta un resultado por cada título encontrado en SciELO, pero únicamente se debe contar una cita, y consignar los otros resultados como repeticiones en la columna de control *Otro XXXX*. Este tipo de artículos demandan especial cuidado, y su monitoreo puede ser de los más extenuantes, y de los más proclives a causar errores de validación.

Falta de acceso. La segunda advertencia está relacionada con la posible falta de acceso al documento citante. Esto podría impedir la validación de muchas citas reportadas por GA. Es posible verificar las citas en documentos de acceso abierto y, en nuestro caso, muchos recursos pueden ser verificados gracias a las suscripciones de la BDCV, o a que forman parte de su acervo impreso. Sin embargo, es importante considerar que, aunque pocas (ver *Citas «no válidas»* abajo), hay publicaciones a las que no tenemos acceso, y cuyas citas no pueden validarse. Por tanto, la cantidad de recursos a los que una biblioteca tenga acceso puede limitar su posibilidad de validar citas reportadas por GA.

Variaciones por año de publicación y por disciplina. La tercera advertencia tiene que ver con la distribución desigual de los datos por año. Como se puede apreciar en la FIGURA 4, la cantidad de artículos publicados por año/centro fluctúa de manera irregular. Si bien podría decirse que se mantiene un promedio de 17 artículos por año, el 2015 es el año con el promedio más bajo (13,8 artículos por centro) y el 2013 el más alto (20,3).

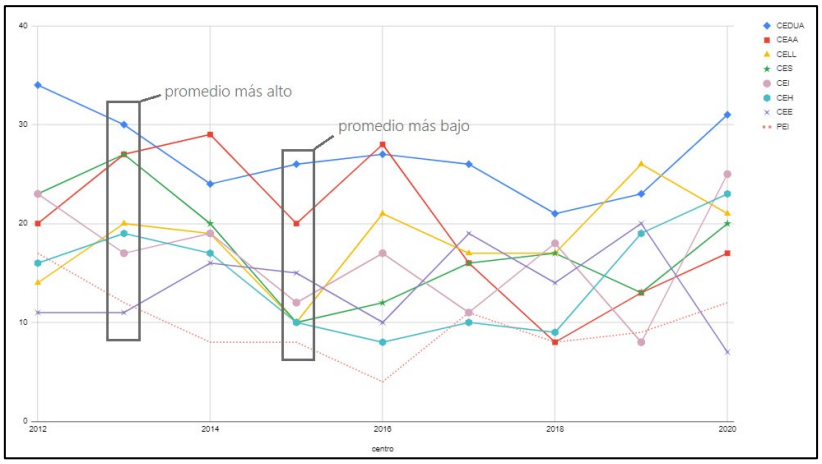


FIGURA 5. NÚMERO DE ARTÍCULOS POR CENTRO SEGÚN AÑO DE PUBLICACIÓN

Así, si un análisis como éste fuera a ser replicado en otra institución, recomendaríamos tomar una muestra de por lo menos cinco años para obtener las diferencias de comportamiento por año y por departamento. Con este último objetivo en mente, recomendamos visualizaciones como la FIGURA 5, cuyo diagrama de caja permite leer las siguientes características:

- La media por centro. Como puede observarse, ésta oscila entre 26 (CEDUA) y 9 (PEI) artículos por año.
- Las medidas más bajas y más altas por centro. Por ejemplo, para el CES, esta medida sería de aproximadamente 10 artículos por año, pues en 2015, 2016 y 2019 publicó entre 10 y 13 artículos por año; en contraste, en el caso del CELL, podemos ver un punto, que representa un valor atípico, pues sólo un año (en 2015) se publicaron 10 artículos.
- La dispersión de los datos. Vemos que la variación más alta es en el CEAA, cuya producción anual, en su mayoría puede ser descrita entre los 17 y 27 artículos anuales; mientras que el CELL tiene una producción más constante que oscila entre los 17 y 21 artículos por año.

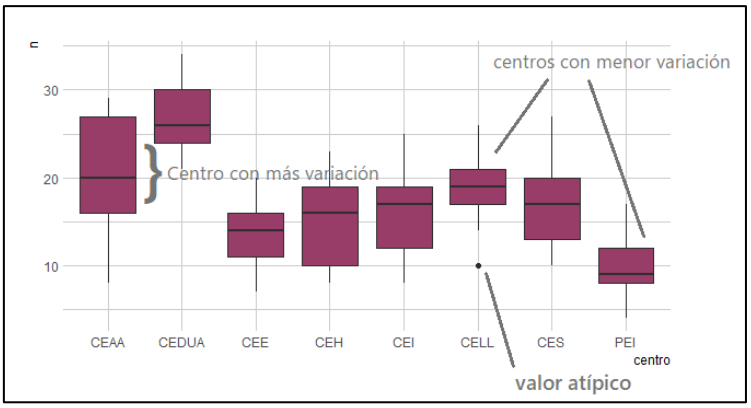


FIGURA 6. DIAGRAMAS DE CAJA DE NÚMEROS DE ARTÍCULOS POR CENTRO

Con datos como los de FIGURA 4, se pueden crear políticas institucionales que analizan las condiciones de los años más productivos, pero también que promuevan el establecimiento de una ventana de tiempo adecuada para el análisis de la producción. Ésta deberá establecerse después de verificar la variabilidad de los años con producción baja (para el caso de este estudio, el 2015) y alta (como el 2013). Por otro lado, la información que obtenemos con la FIGURA 5, nos permite tener un panorama de la variación por centro. Por ejemplo, mientras que los datos para un centro como el CELL o el PEI podrían ser significativos casi en cualquier rango de años (pues su variación no es tan extrema), no lo serían para centros como el CEAA o el CEH que tienen una variación más grande por año.

Citas «válidas». La cuarta advertencia corresponde a las complicaciones que existen para discernir e interpretar las citas válidas de las que detecta GA, en su dimensión temporal. En 2021, GA identificó 6.561 posibles citas a los artículos publicados entre 2012 y 2020. Después de la verificación manual 4.348 citas, es decir, un 66,27 % del total fueron consideradas válidas. Esta proporción varía desde un 53 % en 2020 hasta un 71 % en 2012. Es decir, nuestros datos sugieren que entre más tiempo haya pasado entre la publicación del artículo y el análisis de citas, mayor probabilidad existe de que la cita recuperada por GA sea una cita válida, pues el margen de error es menor a mayor cantidad de datos. Así, mientras el 47 % de citas no válidas en el 2020 equivale a unas cuantas citas (18 de un total de 39), el 71 % de citas válidas del 2012 equivale a 460 citas no válidas de un total de 1.562 (FIGURA 6).

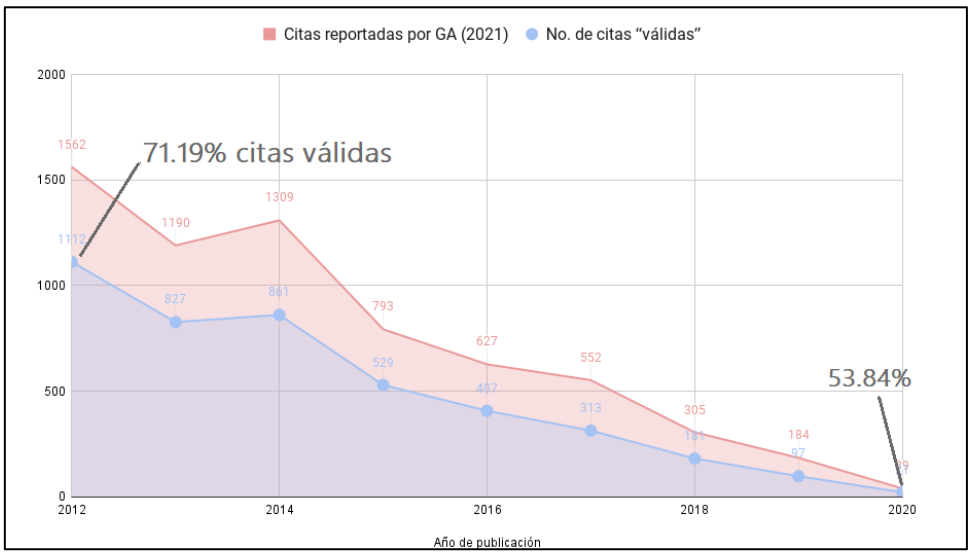


FIGURA 7. CITAS SEGÚN GA POR AÑO DE PUBLICACIÓN DE ARTÍCULOS VS CITAS «VÁLIDAS»

En cuanto a la interpretación es importante considerar el correlato, es decir que la mayoría de las citas válidas corresponderá al extremo más antiguo de los datos. Por ejemplo, en nuestro análisis el 23 % de las citas identificadas por GA (es decir 1.562 citas) corresponden a las hechas a los artículos publicados en 2012, el primer año del periodo analizado; en contraste con el 2 % (39 citas) que es para los artículos de 2020 (ver FIGURA 6). Es decir, los artículos de 2012 han tenido ocho años para acumular citas; mientras que los de 2020, sólo tuvieron algunos meses o incluso días para ser descubiertos y citados por la comunidad académica.

Citas «no válidas». Como se ha mencionado, en el ejercicio de la BDCV además de filtrar los documentos no arbitrados, se han establecido otras categorías de citas «no válidas» cuyo desglose numérico es el siguiente: 1.094 duplicados, errores y otros (16 % del total de citas identificadas por GA y 46,8 % del total de no válidas); 609 se encontraban en documentos no arbitrados (9 % del total de citas identificadas de GA); 408 eran autocitas (6 % del total) y 226 citas provenían de documentos a los que no se tuvo acceso y no pudieron ser corroboradas (3 %).

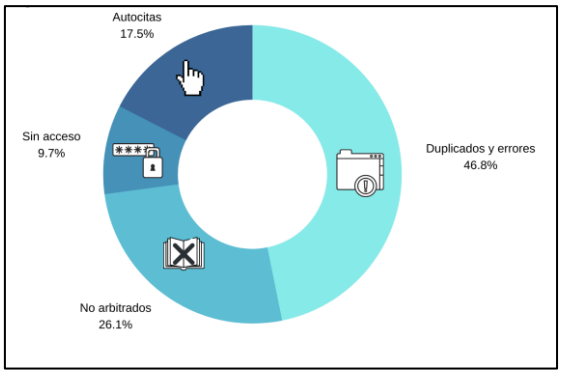


FIGURA 8. CITAS NO VÁLIDAS, PORCENTAJES POR TIPO

Esta es la primera vez, hasta donde sabemos, que se cuenta con una comprobación manual de los datos recuperados por el algoritmo de GA para artículos de CSyH y esperamos que contribuya a hacer lecturas más cautas sobre la información que este buscador maquinalmente ofrece. Es decir, quien tome datos de GA para un análisis de citas a artículos de estas áreas mayoritariamente en español, puede calcular que habrá alrededor de un 16 % que corresponden a duplicados, errores y otros; cerca de un 9 % podría tratarse de documentos no arbitrados; y 6 % podrían ser autocitas. En consecuencia, si se aplican estos tres criterios, hasta un 32 % de las citas identificadas por GA serían citas no válidas. Más aún, como se puede leer en la FIGURA 6, esta proporción podría aumentar dependiendo del tiempo que han tenido para acumular citas. Así, la proporción de citas válidas en años más lejanos (ver 2012) podría ser mucho más alta que la de años más recientes en los que ha pasado poco tiempo para la acumulación de citas (ver 2020).

Documentos des-indexados. Si a las citas identificadas por GA que son 6.561, le restamos las citas válidas (4.348) obtenemos un total de 2.213 registros «no válidos». Sin embargo, si sumamos el total de «citas no válidas» de acuerdo con nuestro control, obtenemos 2.337 citas de este tipo. Es decir, existe una diferencia de 124 documentos «no válidos» que parecerían sobrar. Esto nos lleva a la sexta advertencia: hay documentos citantes que dejan de estar indexados en GA año tras año, lo cual se explica porque el índice de GA funciona automáticamente y si la cita no está escrita conforme a los

lineamientos que Google usa para su indización ese año, éstos no son incluidos (Google Scholar Help, s/f).

Por ejemplo, puede suceder que, en determinado año, el algoritmo de GA identifique que un artículo X tiene dos citas y que, al siguiente año, reporte cero, porque en el ajuste de su algoritmo dos artículos ya no son identificados como citantes. O bien, que en otro año se reporten dos citas y al año siguiente, cuatro, pero que estas cuatro sean todas nuevas, pues las del primer año ya no son identificados por el algoritmo como válidas. En el ejercicio que se hace en la BDCV, esto no es la norma, sino la excepción, pero calculamos que alrededor de 124 documentos citantes «no válidos» estaban presentes en búsquedas anteriores, pero han desaparecido en los resultados de GA.

Esto también se expresa en números negativos en la columna *Diferencias* y *Comprobación* de la hoja de control. Es decir, a veces al calcular la diferencia entre citas reconocidas por GA el año anterior al estudio para obtener las citas nuevas identificadas salen números negativos por lo explicado arriba. La práctica, en este caso, permite identificar que en GA hay registros que fueron validados en años anteriores y posteriormente ya no figuran entre los resultados. Esta volatilidad de los resultados de las citas que indiza GA, hace más compleja esta labor y justifica el uso de notas detalladas para el seguimiento de estos casos año tras año. Como si se tratará de una cadena de evidencia.

Diferencias por área. La séptima y última advertencia tiene que ver con los comportamientos de citación distintos de las Humanidades y las Ciencias Sociales. Si bien, de los 1.236 artículos publicados entre 2012 y 2020, 597 (el 48 %) fueron citados al menos una vez, este porcentaje varía entre los diferentes Centros.

Por ejemplo, los cuatro centros que pudieran considerarse más cercanos a lo que se denominan «Ciencias Sociales», son los que concentran la proporción más alta de citas a sus artículos publicados. El 73 % de los 123 artículos publicados por el profesorado del CEE fueron citados al menos una vez; está seguido por el CES, donde el 59 % de sus 158 artículos tuvieron por lo menos

una cita; en tercer lugar, está el CEDUA, que obtuvo una proporción ligeramente menor (57 % de sus 242 artículos) y por último, el CEI (46 % de sus 150 artículos). En contraste, los centros que pudieran considerarse más «humanistas», se encuentran en el extremo más bajo de la proporción: el CEH (45 % de 131), CELL (33 % de 165) y CEAA (28 % de 178).

Ahora, la advertencia va dirigida no sólo a reconocer esta diferencia sino a comprender que: primero, como se ha expresado en los reportes anuales que entrega la Biblioteca, la bibliografía apunta a que las Humanidades tienden a reportar pocas citas a artículos en revistas académicas, pues su circulación científica tiende a moverse más en materiales monográficos (LARIVIÈRE, GINGRAS, & ARCHAMBAULT, 2009; *Las revistas científicas y el español*, 2019), los cuales no han sido incluidos en este estudio. Segundo, la tendencia a publicar en línea también favorece la recuperación de citas hechas en artículos de revistas científicas; en contraste, un número (difícil de determinar) de citas hechas en libros son invisibles a una herramienta como GA, a menos que los motores de búsqueda de esta última tengan acceso a una versión legible por computadora del libro donde la posible cita aparece y esto afecta principalmente a las humanidades. Tercero, como se ha mencionado antes, un periodo tan corto (ocho años en nuestro estudio) afecta especialmente a las disciplinas de las humanidades, las cuales no suelen tratar temas de coyuntura y tardan en acumular citas.

Aún con todo, las proporciones obtenidas en nuestro estudio son alentadoras si se consideran las alarmantes cifras que DEREK CURTIS BOK recuperó en su libro sobre educación superior en Estados Unidos, en el que afirma que un 98 % de las publicaciones en humanidades y un 75 % de las de ciencias sociales, nunca obtienen una cita (BOK, 2013, p. 330). Este dato fue tomado de un artículo escrito 22 años antes (TAINER *et al*., 1991) al que PENDLEBURRY, en una carta abierta de ese mismo año, hizo una corrección numérica que no es menos preocupante: 93 % de los artículos de humanidades permanecen sin citas los primeros cinco años después de su publicación y entre 45 % y 49 % en el caso de las ciencias sociales (PENDLEBURY, 1991).

Sin embargo, considerando la misma ventana de tiempo (es decir, artículos que han tenido por lo menos cinco años para acumular citas), esa cifra jamás estuvo por arriba del 75 % de artículos sin cita para el caso de las Humanidades y el promedio de artículos con cita para ese periodo (2012-2015) fue de 34 % para el CEAA, 54 % para el CELL y 68 % para el CEH; es decir, siempre muy lejos de tener más del 93% de sus publicaciones sin cita como refiere Pendleburry (ver FIGURA 8).

Por otro lado, el de las Ciencias Sociales, aunque en 2014 el CEI tuvo su punto más bajo (47 % de artículos con citas) su promedio se mantuvo en 53 % de artículos con cita para ese periodo, CEDUA con 70 % de artículos con cita, y CEE con 83 %.

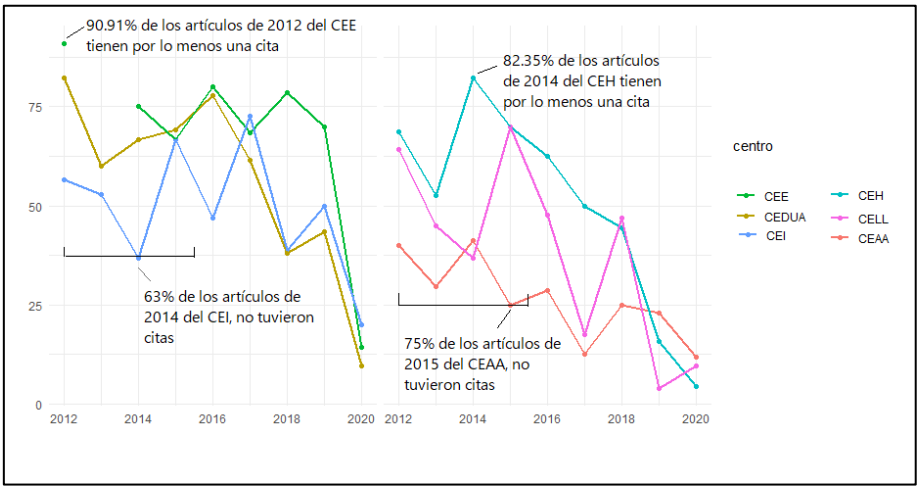


FIGURA 9. PORCENTAJE DE ARTÍCULOS CON POR LO MENOS UNA CITA

Conclusiones

En este artículo se ha descrito un método replicable para el análisis de citas utilizando GA como la mejor fuente abierta disponible para la identificación a artículos de CSyH en español. Hemos realizado una comparación ilustrativa de la dinámica de citación de la producción institucional de un tipo de documentos, los artículos. La dinámica en cuanto a los canales de difusión de cada disciplina es variable y obedece a lógicas de su propio campo de

producción. De tal modo que la medida justa para evaluar su desempeño debería tomar como referencia su propio campo.

Sin embargo, esto no quiere decir que esta sea la única vía y hemos mostrado las limitaciones de este método. Por un lado, quedado fuera del alcance de este estudio tratar iniciativas como **OpenCitations - Home** y **I4OC: Initiative for Open Citations** para aquellas instituciones que no tienen acceso a bases de datos de texto completo de paga para cotejar sus citas. Por otro lado, las limitaciones quedarán aún mejor definidas en cuanto se hagan estudios longitudinales para identificar cuantitativamente la diferencia de GA respecto de otras bases de datos, en relación con la cobertura, traslape y citas únicas para nuestro universo de artículos y en un futuro, la necesaria integración de otro tipo de publicaciones.

Cabe mencionar que este método es perfectible. Desde la BDCV ya existen esfuerzos para resarcir carencias, por ejemplo, crear una base de datos relacional en la que se guarden automáticamente los resultados de GA cada año para así disminuir, hasta cierto punto, el trabajo manual de comprobación, por lo menos, en el nivel de la comparación de documentos citantes con años anteriores y el registro de la cantidad de citas identificadas por los algoritmos de GA.

Por último, quisiéramos destacar que, a pesar de ser un procedimiento arduo, los análisis que se pueden derivar de este tipo de datos son sumamente ricos. Por ejemplo, en los informes anuales entregamos vistas de pájaro sobre la cobertura geográfica de los lugares de publicación de los diferentes centros, así como de las revistas que concentran la mayor cantidad de citas. Este tipo de información puede ser de utilidad para las instituciones académicas que desean tener una impresión de la cobertura geográfica de las citas hechas a los trabajos académicos, pero también para quienes deseen hacer análisis más finos de la circulación del conocimiento por disciplinas y encontrar áreas de oportunidad. En informes futuros, esperamos poder integrar análisis más robustos sobre la inmediatez de las citas, es decir, el tiempo que pasa para que un artículo reciba su primera cita.

Considerando lo anterior, advertimos que es necesario el uso crítico de fuentes para el análisis bibliométrico. Como hemos visto, el uso de una fuente limita o condiciona los resultados. Si bien no existe aún alguna fuente absolutamente confiable y exhaustiva para el análisis de citas a literatura en español y de CSyH, si seguimos utilizando las mismas fuentes como un criterio estático, sin analizar sus limitaciones, pero también sus posibilidades, mantendremos una visión aún más sesgada del rico universo de la comunicación científica que se da desde nuestras regiones y en nuestro idioma.

Bibliografía

Bok, D. C. (2013). *Higher education in America*. Princeton University Press.

Gingras, Y. (2016). *Bibliometrics and research evaluation: Uses and abuses*.

Google Scholar Help. (s/f). https://scholar.google.com/intl/en/scholar/inclusion.html#indexing

Harzing, A.-W., y Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787-804. https://doi.org/10.1007/s11192-015-1798-9

Huynh, D., y Mazzocchi, S. (2021). OpenRefine (Versión v3.4.1) [Java]. https://github.com/OpenRefine

Larivière, V., Gingras, Y., y Archambault, É. (2009). The decline in the concentration of citations, 1900–2007. *Journal of the Association for Information Science and Technology*, *60*(4), 858-862.

Las revistas científicas y el español. (2019). El Colegio Nacional. [video] https://youtu.be/rd8-Rgr5bZE

Martín Martín, A., Orduna Malea, E., Thelwall, M., y Delgado López Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160-1177. https://doi.org/10.1016/j.joi.2018.09.002

Pendlebury, D. A. (1991). Science, citation, and funding. *Science*, *251*(5000), 1410-1411.

PLAZA, L. M., GRANADINO, B., GARCÍA CARPINTERO, E., ALBORNOZ, M., BARRERE, R., y MATAS, L. (2018). El valor del idioma español en ciencia y tecnología. *Rilce. Revista de Filología Hispánica*, 716-745. https://doi.org/10.15581/008.34.2.716-45

ROVIRA, C., CODINA, L., y LOPEZOSA, C. (2021). Language Bias in the Google Scholar Ranking Algorithm. *Future Internet*, *13*(2), 31. https://doi.org/10.3390/fi13020031

RSTUDIO TEAM. (s/f). RStudio: Integrated Development Environment for R (Versión 1.4.1106). Boston, MA: RStudio, PBC. http://www.rstudio.com/

STILLMAN, D. (2021). Zotero (Versión 5.0.96.2). Corporation for Digital Scholarship. https://www.zotero.org/

TAINER, J. A., ABT, H. A., HARGENS, L. L., BOTT, D. M., LANCASTER, F. W., PANNELL, J. H., … PENDLEBURY, D. A. (1991). Science, citation, and funding. *Science*, *251*(5000), 1408-1411.