ARBOR Ciencia, Pensamiento y Cultura
Vol. 197-799, enero-marzo 2021, a592 | ISSN-L: 0210-1963
https://doi.org/10.3989/arbor.2021.799007

EL MOVIMIENTO *OPEN CITATIONS* Y SUS IMPLICACIONES EN LA TRANSFORMACIÓN DE LA EVALUACIÓN CIENTÍFICA

OPEN CITATIONS MOVEMENT AND ITS ROLE IN THE TRANSFORMATION OF RESEARCH EVALUATION

José Luis Ortega
Instituto de Estudios Sociales Avanzados (IESA)
Consejo Superior de Investigaciones Científicas
0000-0001-9857-1511
jortega@iesa.csic.es

Cómo citar este artículo/Citation: Ortega, José Luis (2021). El movimiento Open Citations y sus implicaciones en la transformación de la evaluación científica. *Arbor,* 197(799): a592. https://doi.org/10.3989/arbor.2021.799007

Recibido: 8 septiembre 2020. Aceptado: 2 febrero 2021.
Publicado: 9 abril 2021

RESUMEN: El presente trabajo pretende hacer una revisión del naciente movimiento *Open Citations*, el cual aboga por la libre disposición de las citas bibliográficas incluidas en cada contribución científica. Este movimiento, enmarcado dentro de corrientes más generales como *Open Data* y *Open Access*, busca de esta forma que las citas bibliográficas sean un bien común para la comunidad científica, reforzando el desarrollo de la investigación bibliométrica y la construcción de sistemas de información científica autóctonos. Este cambio está suponiendo una revolución en el mercado de la documentación científica, al surgir nuevos productos y plataformas que permiten valorar la producción e impacto de investigadores e instituciones a partir de fuentes abiertas y alternativas. Esta transformación implica una oportunidad para el desarrollo de portales regionales o institucionales que, alimentados de estas fuentes abiertas, permitan una evaluación propia e independiente. En primer lugar, se hará un análisis del origen y contexto de este movimiento; se analizarán las fuentes de citas abiertas que están apareciendo (*Crossref*, *Microsoft Academic Knowledge Graph*, *Open Citation Corpus*) y algunos productos alternativos (*Lens*, *Dimensions*, *SemanticScholar*); por último, se analizará las implicaciones que todo este movimiento puede tener en la evaluación científica, haciendo hincapié en la posibilidad de desarrollar *Current Research Information Systems* (CRIS) locales destinados a la evaluación científica.

PALABRAS CLAVE: citas abiertas, ciencia abierta, datos enlazados, índices de citas, bibliometría.

ABSTRACT: The aim of this paper is to review the emerging Open Citations movement, which advocates for the free dissemination of bibliographic citations included in research publications. This movement, framed within the broader mainstream currents of Open Data and Open Access, thus attempts for bibliographic citations to be a common good for the scholarly community, reinforcing bibliometric research and the building of internal scientific information systems. This change is causing a revolution in the scientific information market, as new products and platforms are arising that make it possible to evaluate the production and impact of researchers and organizations using open and alternative sources. This transformation is an opportunity to develop institutional or national portals that, fed by these open sources, permit their own and independent evaluation. The paper begins with an introduction about the origin and context of this movement; then, the extraction and process of citations are explained; next, several sources of open citation data are described (Crossref, Microsoft Academic Knowledge Graph, OpenCitation Corpus) and some alternative products (Lens, Dimensions, SemanticScholar); finally, the implications this movement can have on scientific evaluation is analysed, highlighting the possibility of developing local Current Research Information Systems (CRIS) intended for scientific evaluation.

KEYWORDS: Open citations, Open Science, linked data, citation indexes, bibliometrics.

1. INTRODUCCIÓN

La aparición de la web en 1989 planteó la posibilidad de transformar las bases en que se fundamentaba el sistema de publicación científica. El sistema tradicional, centrado en la revista científica como principal vehículo de comunicación, estaba siendo cuestionado por su lentitud, elevado coste y dificultad de acceso (Byrd, 1990). La Web ofrecía una vía alternativa donde primaba la inmediatez y el bajo coste. Desde ese momento, el sistema de comunicación científica se ha ido transformando hacia un modelo abierto y colaborativo, gracias a la presión de iniciativas concretas (*Open Data*, *Open Peer-review*) que hoy en día agrupamos entorno al concepto de Ciencia Abierta (*Open Science*) y que tiene su origen en el movimiento inicial *Open Access* en 2006. Dentro de esta corriente de la Ciencia Abierta se ha venido a sumar en los últimos años, la iniciativa *Open Citations*. Un movimiento que aboga por la libre disposición de las referencias bibliográficas que permita el diseño independiente de índices de citas y, de esta forma, promover un sistema de evaluación basado en fuentes específicas y más transparentes en sus datos e indicadores.

En concreto, se pretende que las editoriales hagan público, además de los metadatos que describen una publicación (autor, título, revista, etc.), información sobre las referencias que contiene cada documento. Estas referencias forman parte de la propiedad intelectual de la revista, ya que estas poseen un formato específico y propio. Además, estas pasan por un proceso de edición, donde se comprueba su efectiva citación, la identificación inequívoca del documento o la eliminación de erratas.

2. ÍNDICES DE CITAS, BUSCADORES Y *BIG DATA*

Para entender mejor de qué trata este movimiento es necesario saber que las citas o referencias bibliográficas que se incluyen en un trabajo académico (artículo científico, libro, informe) forman parte esencial de la comunicación científica. Con ellas se puede conocer el contexto en que se inscribe una investigación, observando cómo otros trabajos influyen en la creación de nuevo conocimiento. Historiadores de la ciencia ven en la cita un instrumento clave para el estudio de la evolución del pensamiento científico (Price, 1970) mientras que los sociólogos la interpretan como un mecanismo de regulación del sistema de recompensas en la ciencia (Merton, 1973). Sin embargo, la cita adquiere también un importante valor económico, ya que su almacenamiento en bases de datos posibilita identificar publicaciones de gran valor y realizar análisis cuantitativos orientados a la evaluación científica.

Estas bases de datos de referencias enlazadas se las conoce como índices de citas, siendo la *Web of Science* (WoS) y *Scopus* las únicas con una cobertura global. Al estar la evaluación científica y el diseño de políticas cada vez más fundamentadas en indicadores de citas (factor de impacto, índice h, etc.), estos índices se están convirtiendo en piezas clave en el sistema de comunicación científica. Sin embargo, estos índices están desarrollados por empresas privadas (Clarivate para WoS y Elsevier para *Scopus*) que procesan las citas incluidas en las revistas que indizan. Los criterios para seleccionar estas revistas son en muchos casos opacos (Clarivate, 2021) o son científicamente discutibles (Elsevier, 2021), con claros sesgos temáticos y geográficos (Van Leeuwen *et al*., 2001; Mongeon y Paul-Hus, 2016). Otro problema es que los costes de acceso a estos índices son altos, ya que se trata de productos complejos y altamente especializados.

Una alternativa surgida a este modelo son los buscadores académicos. Se trata de índices autónomos de citas que de forma automática rastrean la Web en busca de publicaciones científicas extrayendo las citas insertadas en ellas (Ortega, 2014). Desde el pionero *CiteSeer* (1998) hasta *Google Scholar* (2004), *Microsoft Academic* (2011) y *SemanticScholar* (2015), estos productos se están convirtiendo en serios competidores de los índices de citas, ya que contabilizan las citas de forma más rápida (Thelwall y Kousha, 2017; Thelwall, 2018), tienen una cobertura muy superior (Martín-Martín *et al*., 2018) y su acceso es libre. Sus principales desventajas son que, al ser un proceso automático, son más fáciles de manipular (Delgado López-Cózar *et al*., 2014) y producen más errores e inconsistencias en la identificación y cómputo de la cita (Jacsó, 2010).

Sin embargo, otro fenómeno ajeno a la información científica, y de carácter técnico, vendría a explicar el creciente interés en la disponibilidad de las citas académicas, el *big data*. Este concepto alude a todo lo relacionado con la gestión y explotación de grandes volúmenes de datos, los cuales permitirían un análisis más detallado de la realidad y abrir nuevas vías de conocimiento (Joyanes Aguilar, 2016). Justificado por la reducción de los costes de procesamiento y almacenaje de información, unido a la masiva producción de datos que se vierten en la Web (ciudades inteligentes, internet de las cosas, redes sociales, etc.), el *big data* ofrece mecanismos más sencillos y transparentes de

extraer referencias de documentos científicos (*Web Scraping*, APIs) gestionar grandes volúmenes de información bibliográfica (*JavaScript Object Notation*, JSON y NoSQL) y posibilitar así mayor potencial a la bibliometría y la evaluación científica.

3. LOS ORÍGENES DEL MOVIMIENTO

El origen de este movimiento lo encontramos en un proyecto de investigación, *OpenCitations*, liderado por David Shotton en 2010. Este proyecto tenía un ámbito global y pretendía transformar el acceso a la literatura científica, diseñando un sistema de publicación en abierto de las referencias bibliográficas a través de un sistema de enlaces (*Linked data*) parecido a la Web.

Pero no fue hasta 2017 cuando este movimiento se institucionalizó en la *Initiative for Open Citations* (I4OC). Una fundación que se define como punto de encuentro entre editores científicos, investigadores y cualquier otro agente interesado en promover la disponibilidad en abierto de las referencias bibliográficas. Los patronos más relevantes de esta fundación son organizaciones dedicadas a promover datos abiertos (*Open Citations*, *Wikimedia Foundation*, *DataCite*), editoriales científicas (PLOS, eLife) y universidades (Universidad de Curtin). A finales de 2019, más de 1.200 editoriales científicas, entre las que se encuentran la importantes Springer Nature, Taylor & Francis, Sage y Wiley se habían adherido al movimiento y habían puesto en abierto sus referencias bibliográficas. Como resultado, el porcentaje de artículos con referencias en abierto en ese momento alcanzó el 59%, con más de 500 millones de referencias en *Crossref* (I4OC, 2020).

Posterior a esta institucionalización, han continuado realizándose diversas acciones para concienciar y potenciar medidas en apoyo a las citas en abierto. En diciembre de 2017, la principal sociedad en bibliometría, la *International Society for Scientometrics and Informetrics* (ISSI), lanzó una carta abierta a las principales editoriales para que entendieran la importancia que tiene para el avance de la ciencia el disponer de citas en abierto. Pero quizás la acción más llamativa y de mayor impacto en la comunidad científica fue la renuncia en bloque del comité editorial de la revista *Journal of Informetrics* en 2019, propiedad de Elsevier, en demanda de un acuerdo que permitiera la publicación en abierto de las citas bibliográficas. La respuesta de Elsevier fue que las referencias que se publicaban en sus revistas estaban sujetas a un proceso de revisión y elaboración que no les permitía ponerlas a libre disposición. Sin embargo, se baraja otra razón más poderosa, ya que Elsevier está en el mercado de las citas bibliográficas a través de su índice de citas *Scopus* y podría pensar que la libre circulación de referencias permitiría el desarrollo de productos competidores, como ya está sucediendo (Regier, 2019). En enero de 2020, Open Access Scholarly Publishing Association (OASPA), la asociación de editores académicos en acceso abierto, requirió a sus miembros que depositaran las referencias en abierto en *Crossref*.

Ante la creciente presión, en diciembre de 2020, Elsevier decide firmar la *Declaración de San Francisco sobre la evaluación científica* (DORA), y hacer accesible a través de *Crossref* todas sus referencias bibliográficas (Plume, 2020). Una posible explicación para este cambio de postura puede estar en un giro en la orientación de su negocio hacia el desarrollo de funcionalidades y servicios añadidos que a la explotación de datos en sí (Waltman, 2020). Esta decisión permite incrementar de forma muy considerable el volumen de citas en abierto, ya que Elsevier es el principal editor científico. Como resultado, el porcentaje de artículos con referencias en abierto ha pasado del 59% al 83% (I4OC, 2021). Esta reciente incorporación de Elsevier augura que muchos de los editores aún reticentes como IEEE o la *American Chemical Society* acepten ofrecer sus referencias en abierto, y conseguir en un futuro próximo el 100%.

4. *CROSSREF*, LA FUENTE ORIGINAL

Gran parte de las editoriales científicas de todo el mundo se agrupan en torno a *Crossref*, una fundación sin ánimo de lucro que tiene por objetivo recopilar la producción de estas editoriales y asignar un identificador único (*Digital Object Identifier*, doi) a sus publicaciones. Entre los servicios que ofrece a sus socios está *Cited-by*, un sistema que permite ver qué artículos han sido citados por otros trabajos. Pero para acceder a este servicio deben remitir previamente las referencias de sus artículos, además de sus metadatos. Sin embargo, estas citas son sólo visibles para los socios de *Crossref*. De esta forma, es necesario que cada editor haga explícito si quieren que sus referencias sean públicamente visibles (Tay, 2018). *Crossref* hace pública la lista de editoriales que ponen en abierto sus referencias. El punto de acceso a esta información es la API REST de *Crossref* (https://api.crossref.org/works).

Sin embargo, este depósito en abierto de citas en *Crossref* ha dejado en evidencia algunos problemas graves. Existen muchos artículos que no incluyen re-

ferencias en sus metadatos, incluso cuando la editorial asegura que las está depositando. Springer-Nature e Informa (Taylor & Francis) son las editoriales que tienen mayor porcentaje de citas perdidas (Van Eck, *et al*., 2018). En otros casos, la calidad de la cita es tan pobre que es imposible establecer un vínculo con el documento citado. Todos estos problemas implican que los datos en bruto de *Crossref* deben ser procesados previamente para su utilización en índices de citas.

5. *LINKED DATA* Y LOS MODELOS DISTRIBUIDOS

La principal limitación en la construcción de un índice de citas es que se necesita un gran volumen de documentos (corpus) que contengan citas entre sí. Así que cuantos más documentos contenga, más probable será que entre ellos haya una cita. No es de extrañar que el número de citas promedio que puede tener un artículo en la WoS o *Scopus* sea mucho menor que en *Google Scholar*, ya que el tamaño de *Google Scholar* es varias veces mayor que las anteriores (Moed, Bar-Ilan y Halevi, 2016; Martín-Martín *et al*., 2021). Pero esto tiene el problema de crear compartimentos estancos, en el que el impacto de un documento debe siempre ser observado en referencia a una determinada base de datos.

Para solventar esta limitación, en el que las citas deben estar insertas en un mismo corpus, el movimiento *Open Citations* propone un modelo distribuido en el que la cita a un documento pueda estar alojada en diferentes repositorios, permitiendo conocer el impacto de un documento según múltiples fuentes. Este modelo distribuido es idóneo para un sistema descentralizado en el que diferentes corpus contribuyen para mostrar un impacto global.

La tecnología para desarrollar este modelo la encontramos en la Web semántica y en los datos enlazados (*Linked data*) (Berners-Lee, 2006). El objetivo que se propone la Web semántica es construir una web paralela en el que los objetos estén enlazados entre sí de acuerdo a un significado. De esta forma, un artículo científico se relaciona con otro si este es citado, tiene el mismo autor o tiene las mismas palabras clave. Al igual que la web, no importa donde esté alojado el objeto ya que las relaciones se expresan con enlaces URI (Uniform Resource Identification) que contienen qué tipo de relación existe y donde está situado. Para construir esta red de relaciones conceptuales existe un método de publicación de datos que se denomina datos enlazados (*Linked data*).

Los datos enlazados se expresan en una especificación denominada RDF (*Resource Description Framework*). Esta especificación se basa en declarar los datos como una sintaxis sujeto-predicado-objeto, también llamadas tripletas. De esta forma, una cita podría definirse como artículo-cita-artículo. Como este esquema se fundamenta en URIs o direcciones web, se podría expresar como <https://doi.org/10.1007/s11192-010-0190-z><http://purl.org/spar/cito><https://doi.org/10.1002/asi.22708>. Lo interesante de este enunciado es que cada enlace es también un identificador (doi) que localiza un artículo independientemente de qué base de datos albergue el documento, y la ontología *Citation Typing Ontology* (CiTO) define de forma normalizada la relación de la cita (Peroni y Shotton, 2012). Este modelo de datos tiene la ventaja de que no necesita un corpus cerrado para computar la cita, sino que podemos extraer estas relaciones de múltiples repositorios e integrarlos en una única aplicación. Por ejemplo, el primer doi del ejemplo se puede referir a un documento indexado en *Scopus* y el segundo en la WoS, descubriendo una cita que de otro modo no sería computada.

Generalmente, el acceso a estos datos se realiza con un lenguaje específico para la consulta de datos enlazados denominado SPARQL (*Protocol and RDF Query Language*). Siguiendo una lógica parecida al SQL (*Structured Query Language*), se puede interrogar bases de datos de datos enlazados para recuperar la información necesaria. Sin embargo, muchos repositorios ofrecen más opciones como REST APIs, que permiten descargar registros según determinados parámetros, o ficheros en bruto (*dump files*) que contienen todos los registros del servicio.

6. PROVEEDORES DE CITAS ABIERTAS

Como hemos visto, *Crossref* es el lugar desde donde las editoriales hacen públicas sus citas, convirtiéndose en el principal proveedor de citas abiertas por volumen. Sin embargo, esta no es la única plataforma que recopila citas y las pone a disposición pública. Podemos distinguir dos grupos de plataformas, los servicios comerciales y los proyectos científicos.

Comerciales

En este grupo se incluyen los servicios que capturan citas bibliográficas de publicaciones científicas que, aunque pongan a disposición pública una parte significativa de sus citas, no son organizaciones sin ánimo de lucro.

Microsoft Academic Graph (MAKG)

Microsoft Academic es un buscador académico creado en 2011. Como otros buscadores, usa *crawlers* para rastrear la web académica en busca de publicaciones científicas. En febrero de 2021 contaba con 250 millones de documentos. A diferencia de otros buscadores, *Microsoft Academic* construye todo un grafo de entidades asociadas a la publicación (autores, organizaciones, revistas, temas, etc.) que se relacionan entre sí para convertirse en una auténtica herramienta de descubrimiento, donde no sólo es posible buscar publicaciones, sino también ver informes detallados sobre disciplinas u organizaciones. Gran parte de esta información es publicada en abierto a través del *Microsoft Academic Knowledge Graph* (MAKG), una representación en datos enlazados de las relaciones de todos los elementos de *Microsoft Academic*. En su última actualización (noviembre 2018), dispone de 146 millones de citas en tripletas RDF que pueden ser consultadas a través de un punto de acceso SPARQL y en ficheros en bruto.

Semantic Scholar

Semantic Scholar es un buscador surgido en 2015 que viene a competir en el mercado de los buscadores académicos aportando soluciones basadas en inteligencia artificial. Desarrollado por *Allen Institute for Artificial Inteligent*, este buscador, además de rastrear la web académica, actúa también como repositorio, ya que descarga y almacena las versiones en abierto de un resultado científico. Actualmente cuenta con más de 184 millones de documentos. *Semantic Scholar* ofrece de forma abierta una API (https://api.semanticscholar.org/v1) para acceder a toda su información. Además, a través del *Open Research Corpus*, pone a disposición los ficheros en bruto de datos bibliográficos para descargar (http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/download/). Desgraciadamente, *Semantic Scholar* sólo ofrece citas para aquellos documentos que tengan una versión en abierto y cuyas referencias puedan ser extraídas automáticamente.

SN SciGraph

Es una iniciativa aún en desarrollo entre el grupo editorial Springer Nature y la empresa especializada en información científica *Digital Science*. Basándose en tecnologías de la web semántica, tiene por objetivo crear un sistema de conocimiento que otorgue valor a las publicaciones generadas por el grupo editorial. Así, cuenta actualmente con cerca de diez millones de documentos y 160 millones de citas, que pueden ser directamente accesibles a través del grafo (https://scigraph.springernature.com) o descargando sus ficheros en *figshare* (https://sn-scigraph.figshare.com/). Por ahora, estas citas sólo enlazan a los contenidos de Springer Nature, pero su modelo de datos enlazados permitirá integrarlo con otros corpus.

Proyectos científicos

A continuación, se detallan algunas iniciativas desde el ámbito científico orientadas a la distribución y generación de citas bibliográficas enlazadas. En muchos casos se tratan de proyectos y prototipos de corto alcance pero que ilustran el potencial de este movimiento y su desarrollo en diferentes entornos.

National Institutes of Health, NIH Open Citation Collection

Como parte de *iCite*, un índice de citas especializado en biomedicina, el *NIH Open Citation Collection* en PubMed (Hutchins *et al*., 2019). El núcleo d contiene en torno a 420 millones de citas entre artículos indizados el servicio lo forma citas de *Crossref* (≈60%), *National Library of Medicine* (≈35%) y una pequeña parte de un sistema propio de extracción de citas, *Machine Learning pipeline* (≈5%). Los datos son accesibles a través del interfaz web de *iCite* https://icite.od.nih.gov), la API de *iCite* (https://icite.od.nih.gov/api), o a través de ficheros (http://doi.org/10.35092/yhjc.c.4586573).

Scholar Index

Se define como índice de citas colaborativo especializado en Arte y Humanidades. Su origen está en el proyecto *Linked Books*, desarrollado por la Escuela Politécnica Federal de Lausana, y con el objetivo de recopilar literatura sobre la historia de Venecia. Utilizando técnicas de *machine learning*, consigue extraer citas bibliográficas de documentos antiguos escaneados, enlazando fuentes primarias y secundarias (Romanello y Colavizza, 2018). Por ahora sólo tiene disponible un corpus, *Venice Scholar*, con 3,8 millones de referencias extraídas y 80 mil citas. Su propósito es ampliar este prototipo atrayendo la participación de bibliotecas y museos que desean escanear sus documentos y extraer las referencias. De esta forma, a partir de la unión de colecciones particulares de corpus de citas, pretende construir un gran índice federado de citas de documentación histórica. Emplea el

ARBOR Vol. 197-799, enero-marzo 2021, a592. ISSN-L: 0210-1963 https://doi.org/10.3989/arbor.2021.799007

mismo modelo de datos de *Open Citations*, por lo que permite integrar las citas de *Scholar Index* en el *Open Citations Corpus*. El acceso a sus datos se puede realizar a través de su interfaz web (https://venicescholar.dhlab.epfl.ch) o a través de la API (https://api-venicescholar.dhlab.epfl.ch/v1/swagger.json).

Linked Open Citation Database (LOC-DB)

Es un proyecto local de la biblioteca de la Universidad de Mannheim. Su objetivo es digitalizar, extraer y publicar las citas de todos los documentos en ciencias sociales adquiridos en 2011, principalmente libros y artículos de revista. Con esta muestra se pretende obtener medio millón de referencias. Siguiendo el mismo modelo de datos de *Open Citations*, espera ser un referente local para la extracción de citas, y que estas puedan incorporarse al corpus general de *Open Citations* (Lauscher *et al*., 2018). El proyecto está en fase de desarrollo y cuenta con una página web (https://locdb.bib.uni-mannheim.de/blog/en/).

Wikicite

Se trata de una iniciativa para crear una base de datos bibliográfica a partir de toda la información recogida en la *Wikimedia*, el nombre que agrupa a todas las iniciativas en torno a la *Wikipedia*. Se inició en 2016 y su objetivo es mejorar y normalizar los procesos de cita dentro de la *Wikimedia*, a la vez que pretende crear un repositorio de publicaciones abierto y enlazado. Se organiza en conferencias anuales, donde los colaboradores aportan resultados y propuestas. En septiembre de 2019 contaba con 22 millones de publicaciones y 155 millones de citas de artículos. Estos datos son accesibles a través de un interfaz de búsqueda (http://wikicite.org/access.html), un punto de acceso SPARQL (https://query.wikidata.org/) y en grandes ficheros (https://www.wikidata.org/wiki/Wikidata:Database_download).

Excite

Otro proyecto de origen alemán es Excite (Universidad de Koblenz), aunque está centrado más en desarrollar tecnología para la extracción de citas de documentos en PDF, que en la de proveer de citas abiertas. Está especializado en literatura en ciencias sociales en alemán. http://excite.west.uni-koblenz.de/website/

Open Citations

Como se ha comentado anteriormente, *Open Citations* es el proyecto germen del movimiento *Open Citations*. El resultado más importante fue la creación de un repositorio de referencias y citas bibliográficas entrelazadas entre sí de acuerdo a una ontología (SPAR). A este repositorio se le llamó *Open Citations* Corpus (OCC). En 2015, Silvio Peroni se unió al proyecto, mejorando los metadatos y la alimentación automática de registros. Actualmente, OCC está instalado en la Universidad de Bolonia (Peroni y Shotton, 2020). Los datos pueden consultarse de varias formas: a través de una REST API, un punto de consulta SPARQL y un interfaz de búsqueda. También ofrece la posibilidad de descargar los ficheros completos de la plataforma. Desde su puesta en marcha en 2010 han desarrollado varios índices de citas:

- El *Open Citations Corpus* (OCC) contiene en marzo de 2020 14 millones de citas a 7.5 millones de artículos. Muchos de ellos especializados en biomedicina y provenientes del Pubmed Central.
- *Open Citations Index* of *Crossref open DOI-to-DOI citations* (COCI) es el de mayor tamaño y se crea en junio de 2018. Contiene exclusivamente citas entre dois, a consecuencia de la publicación en abierto de las citas de *Crossref*. Contiene más de 655 millones de citas y 55 millones de publicaciones. Este índice tiene la particularidad de que la cita es tratada como objeto y no como atributo, así que lo que contiene son pares de dois (transformado a un código numérico) de documentos que son citados (Heibi *et al*., 2019a).
- *Crowdsourced Open Citation Index* (CROCI) es el índice más reciente, creado en marzo de 2019 y se alimenta de aportaciones voluntarias de autores, identificados con ORCID, y editores de revistas de acceso abierto. Por su juventud se desconoce su tamaño, que presumiblemente es pequeño con relación a los anteriores.

Aparte de todas estas fuentes, es necesario mencionar algunos productos comerciales que han surgido basándose en las citas abiertas que suministran los servicios y proyectos anteriormente citados. Estos ejemplos demuestran cómo pueden surgir nuevos productos que enriquecen el mercado e incrementan la competencia a partir de fuentes abiertas.

The Lens

Nacido en 2000 como un servicio especializado en la búsqueda de patentes, en 2013 incorporó referencias de artículos científicos, ofreciendo cerca de 214 millo-

nes de registros (lens.org). Lo innovador de este producto es que enlaza citas de artículos con patentes y viceversa, mostrando el impacto de los resultados científicos en la innovación tecnológica y, al contrario. Las fuentes bibliográficas que alimentan este servicio son múltiples, aunque las principales son *Microsoft Academic* y *Crossref*.

Dimensions

Dimensions fue creada por *Digital Science* en 2018 (app.dimensions.ai), y ofrece 110 millones de publicaciones científicas junto a un menor número de patentes, ensayos clínicos, informes técnicos, etc. El núcleo de *Dimensions* lo constituye así 100 millones de publicaciones de *Crossref*. Sin embargo, la limitada calidad de las citas en *Crossref*, lleva a que estos datos sean procesados para recuperar citas perdidas o incompletas (Bode *et al*., 2019).

7. CITAS ABIERTAS FRENTE CITAS CERRADAS

Sin embargo, unas de las cuestiones más acuciantes respecto a las citas en abierto es saber su volumen, cuantas citas hay disponibles y en qué sentido son una verdadera alternativa a los índices de citas tradicionales. En 2013, una primera estimación arrojó que sólo un 4% de las citas estaban en abierto en diferentes y pequeños proyectos (Shotton, 2013). Cinco años después, en un estudio comparativo entre las referencias en abierto depositadas en *Crossref* y las disponibles en la Wos y *Scopus*, estimó que un 39.7% de las citas de la Wos y un 34.8% de las de *Scopus* estaban en abierto en *Crossref* (Van Eck *et al*., 2018). Un año más tarde, Anne Harzing (2019) mostró que la cobertura de citas de *Crossref* con respecto a la Wos y *Scopus* es muy similar, aunque su estudio está limitado al ámbito de la Administración de Empresas. Sin embargo, estos resultados no están mal encaminados, ya que en el mismo año Heibi *et al*. (2019b) confirmaron por primera vez que el volumen de citas en abierto ya superaba al total de citas por suscripción, de forma general y en todas las disciplinas. Visser *et al*. (2020), comparando índices de citas abiertos y cerrados, comprobaron que *Microsoft Academic* y *Dimensions* (abiertos) incluyen tantas citas como *Scopus* (cerrados). Sin embargo, los autores señalan que el principal problema de *Microsoft Academic* deriva de la extracción automática de datos, que en muchos casos no consigue identificar correctamente las citas. El trabajo más reciente al respecto y el más completo, ya que añade también buscadores académicos (*Google Scholar* y *Microsoft Academic*), señala el enorme peso de *Google Scholar* al encontrar más citas (<89%) que el resto de índices. El estudio señala las grandes posibilidades que existirían para las citas abiertas si Google permitiera el procesamiento de sus datos (Martin-Martin *et al*., 2021).

8. CONCLUSIONES

El desarrollo y consolidación del movimiento *Open Citations* ha venido marcado por dos hechos. El primero es la aparición de herramientas tecnológicas que permiten la extracción, almacenaje y enlace de citas bibliográficas de una forma asequible, fácil y abierta. Y, por otro lado, el apoyo de la filosofía ciencia abierta y la concienciación de que las referencias bibliográficas deben ser un patrimonio de la comunidad científica a su libre disposición. Son varias las ventajas que el movimiento *Open Citations* tiene para la comunidad científica.

La primera y fundamental es que permite poner a disposición de todo el mundo un volumen grande y valioso de datos bibliográficos. Ya que a partir de estos corpus de citas se pueden desarrollar bases de datos de literatura científica abiertas y accesibles a toda la comunidad científica, independientemente de sus medios, recursos y lugar. Esto favorece la democratización de la investigación científica, eliminando cualquier brecha asociada al acceso de la información.

Para la investigación científica supone la apertura de un enorme abanico de posibilidades de investigación. Por un lado, permite trabajar con más fuentes de citación, posibilitando un mayor detalle y precisión en cualquier estudio relativo al impacto. Por otro, y quizás más importante, es la reproducibilidad de los resultados de cualquier estudio bibliométrico, ya que permite el acceso libre a los mismos datos utilizados por estudios anteriores. Esto, sin duda, enriquece el debate científico y la discusión sobre la fiabilidad de las fuentes bibliométricas.

Otra de las ventajas es la posibilidad de desarrollar nuevos productos comerciales a partir de fuentes abiertas. La reciente explosión de buscadores académicos (*Semantic Scholar*, *Baidu Scholar*) e índices de citas (*Dimensions*, *The Lens*) se explica en gran medida por la disponibilidad de citas en abierto. Esta realidad favorece la competencia entre los distintos actores, los cuales deben esforzarse en incluir mejoras e innovaciones que aporten más valor añadido a los datos.

Desde el punto de vista de la evaluación científica, las citas en abierto abren la posibilidad de desarrollar plataformas de información académica destinadas es-

ARBOR Vol. 197-799, enero-marzo 2021, a592. ISSN-L: 0210-1963 https://doi.org/10.3989/arbor.2021.799007

pecíficamente a la evaluación científica. Repositorios, CRIS u otro tipo de plataformas institucionales o regionales pueden incluir información de citas. Estos servicios orientados a la evaluación permitirían cuantificar el impacto de investigadores, departamentos, líneas u organizaciones, a través de indicadores específicos ajustados a cada realidad científica. Estos sistemas serían abiertos y supondría una enorme reducción de costes en las evaluaciones de autores y proyectos.

Otra ineludible ventaja viene por parte de los editores científicos. Las citas bibliográficas no dejan de ser vínculos que asocian publicaciones y por las que se puede navegar buscando nuevos contenidos. Muchos editores son conscientes de que gran parte de las visitas y lecturas de sus artículos vienen a través de citas externas. De esta forma, cuanto más abiertas sean las citas, más tráfico puede generarse en las plataformas de cada editorial, pudiendo incrementar las suscripciones y el interés de nuevo autores.

Sin embargo, la principal desventaja que el uso de citas abiertas está generando es su falta de estandarización y la deficiente información que algunos editores suministran (Van Eck, *et al*., 2018). Puesto que las citas son publicadas por cada editorial, cada paquete de citas puede contener formatos propios, carecer de algunos elementos o no disponer de vínculos entre publicaciones que citan y publicaciones citadas.

Otro problema es de carácter técnico, y está relacionado con la publicación de citas de forma enlazada. Este tipo de publicaciones requiere conocimientos técnicos en métodos y normas para la preparación, almacenaje y publicación de estos datos, lo que implica un esfuerzo y coste adicional.

Finalmente, y a modo de crítica a este movimiento, debemos enmarcar este compromiso por las citas dentro de la idea general de ciencia abierta. Es necesario reconocer que existen aún muchos metadatos asociados a las publicaciones (afiliaciones, agradecimientos, etc.) que tampoco son accesibles y que permitirían incrementar el acceso y explotación de la literatura científica. La reciente iniciativa por el libre acceso a los resúmenes I4OA evidencia que queda mucho camino por recorrer en lo que al libre acceso se refiere. Más aún, una crítica importante a estas iniciativas centradas en elementos específicos de la publicación, es que ocultan cierta claudicación ante la principal aspiración de un libre acceso a la literatura científica. La libre disposición de citas y resúmenes debe ser entendido sólo como un paso previo a una total apertura de la información científica.

REFERENCIAS

Berners-Lee, Tim (2006). Linked Data – Design Issues [en línea]. [8 de junio de 2020]. Disponible en: https://www.w3.org/DesignIssues/LinkedData.html

Bode, Christian; Herzog, Christian; Hook, Daniel; McGrath, Robert (2019). A Guide to the Dimensions Data Approach. Dimensions Report [en línea]. [8 de junio de 2020]. Disponible en: http://doi.org/10.6084/m9.figshare.5783094

Byrd, Gary D. (1990). An economic commons tragedy for research libraries: scholarly journal publishing and pricing trends. *College & Research Libraries*, 51(3), pp. 184-195.

Clarivate (2021). Web of Science Journal Evaluation Process and Selection Criteria [en línea]. [8 de junio de 2020]. Disponible en: https://clarivate.com/webofsciencegroup/journal-evaluation-process-and-selection-criteria/

Delgado López-Cózar, Emilio; Robinson-García, Nicolás; Torres-Salinas, Daniel (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, 65(3), pp. 446-454.

Elsevier (2021) Content Policy and Selection: Scopus [en línea]. [8 de junio de 2020]. Disponible en: https://www.elsevier.com/solutions/scopus/how-scopus-works/content/content-policy-and-selection

Harzing, Anne. W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, 120(1), pp. 341-349.

Heibi, Ivan; Peroni, Silvio; Shotton, David (2019a). Software review: COCI, the Open Citations Index of Crossref open DOI-to-DOI citations. *Scientometrics*, 121(2), pp. 1213-1228.

Heibi, Ivan; Peroni, Silvio; Shotton, David (2019b). Crowdsourcing open citations with CROCI-An analysis of the current status of open citations, and a proposal. *arXiv preprint arXiv:1902.02534*.

Hutchins, B. Ian; Baker, Kirk L.; Davis, Matthew T.; Diwersy, Mario A.; Haque, Ehsanul; Harriman, Robert M.; Santangelo, George M. (2019). The NIH Open Citation Collection: A public access, broad coverage resource. *PLoS biology*, 17(10), pp. e3000385.

Initiative for Open Citations (I4OC) (2020). I4OC: Initiative for Open Citations [en línea]. [8 de junio de 2020]. Disponible en: https://i4oc.org/

Jacsó, Peter (2010). Metadata mega mess in Google Scholar. *Online Information Review*, 34(1), pp. 175-191

Joyanes Aguilar, Luis (2016). *Big Data, Análisis de grandes volúmenes de datos en organizaciones*. Madrid: Alfaomega Grupo Editor.

Lauscher, Anne; Eckert, Kai; Galke, Lukas; Scherp, Ansgar; Rizvi, Syed Tahseen Raza; Ahmed, Sheraz; Klein, Annette (2018). Linked open citation database: Enabling libraries to contribute to an open and interconnected citation graph. En: Chen, J., Gonsalves. M. A., Allen, J. M. (ed.).

Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. New York: ACM, pp. 109-118.

Martín-Martín, Alberto; Orduña-Malea, Enrique; Thelwall, Mike; Delgado López-Cózar, Emilio (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), pp. 1160-1177.

Martín-Martín, Alberto; Thelwall, Mike; Orduña-Malea, Enrique; Delgado López-Cózar, Emilio (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and Open Citations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126(1), pp. 871–906.

Merton, Robert K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago press.

Moed, Henk F.; Bar-Ilan, Judit; Halevi, Gali (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of informetrics*, 10(2), pp. 533-551.

Mongeon, Philippe y Paul-Hus, Adele (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106(1), pp. 213-228.

Ortega, José Luis (2014). *Academic Search Engines: A Quantitative Outlook*. Oxford, UK: Chandos Publishing (Elsevier Group).

Peroni, Silvio y Shotton, David (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, 17, pp. 33-43.

Peroni, Silvio y Shotton, David (2020). Open Citations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, 1(1), pp. 428-444.

Plume, Andrew (2020). Advancing responsible research assessment. *Elsevier Connect* [en línea]. [8 de junio de 2020]. Disponible en: https://www.elsevier.com/connect/advancing-responsible-research-assessment

Price, Derek. J. de Solla (1970). Citation Measures of Hard Science, Soft Science, Technology, and Nonscience. En: Nelson, C. E. y Pollock, D. K. (ed.). Communication among Scientists and Engineers, Lexington, MA: D.C. Heath and Company, pp. 3-22.

Regier, Ryan (2019). The longer Elsevier refuses to make their citations open, the clearer it becomes that their high profit model makes them anti-open. *Medium* [en línea]. [8 de junio de 2020]. Disponible en: https://medium.com/@ryregier/the-longer-elsevier-refuses-to-make-their-citations-open-the-clearer-it-becomes-that-their-high-78576a48e64e

Romanello, Matteo y Colavizza, Giovanni (2018). The Scholar Index: A collaborative Citation Index for the Arts and Humanities. Open Citations Workshop. University of Bologna.

Shotton, David (2013). Publishing: open citations. *Nature News*, 502(7471), pp. 295.

Tay, Aaron (2018). Understanding the implications of Open Citations — how far along are we? *Academic Librarians on open access* [en línea]. [8 de junio de 2020]. Disponible en: https://medium.com/a-academic-librarians-thoughts-on-open-access/understanding-open-citations-f31b2f3a2533

Thelwall, Mike y Kousha, Keivan. (2017). ResearchGate versus Google Scholar: Which finds more early citations? *Scientometrics*, 112(2), pp. 1125-1131.

Thelwall, Mike. (2018). Does Microsoft Academic find early citations? *Scientometrics*, 114(1), pp. 325-334.

Van Eck, Nees. J.; Waltman, Ludo; Larivière, Vincent; Sugimoto, Cassidy. (2018). Crossref as a new source of citation data: A comparison with Web of Science and Scopus. CWTS Blog [en línea]. [8 de junio de 2020]. Disponible en: https://www.cwts.nl/blog?article=n-r2s234&sthash.IlnLf4Uz.mjjo

Van Leeuwen, Thed N.; Moed, Henk. F.; Tijssen, Rober J.; Visser, Martijn. S.; Van Raan, Antony. F. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics*, 51(1), pp. 335-346.

Visser, Martijn; Van Eck, Nees. J.; Waltman, Ludo (2020). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *arXiv preprint arXiv:2005.10732*.

Waltman, Ludo (2020). Q&A about Elsevier's decision to open its citations. *Leiden Madtrics* [en línea]. [8 de junio de 2020]. Disponible en: https://leidenmadtrics.nl/articles/q-a-about-elseviers-decision-to-open-its-citation