CÓMO ANALIZAR EL IMPACTO DE LOS DATOS DE INVESTIGACIÓN CON MÉTRICAS: MODELOS Y SERVICIOS

How to analyze the impact of research data with metrics: Models and services

Antonia Ferrer-Sapena, Enrique-Alfonso Sánchez-Pérez, Rafael Aleixandre-Benavent y Fernanda Peset

Antonia Ferrer-Sapena es licenciada en geografía e historia contemporánea por la *Universidad de Valencia* (*UV*), y doctora en técnicas y métodos de información y documentación por la *Universidad Politécnica de Valencia* (*UPV*). Es directora académica del *Master en Gestión de la Información MUGI* y directora de la *Cátedra de Transparencia y Participación* de la *Consellería de Transparencia, Responsabilidad Social, Participación y Cooperación*. Es profesora titular de la *UPV* y miembro del *Grupo ThinkEPI: estrategia y prospectiva de la información*. Su interés se centra en la investigación sobre datos abiertos y sus implicaciones en la empresa, la administración y la ciencia que aborda el proyecto de I+D del *Ministerio de Economía y Competitividad Datasea, datos abiertos de investigación, open research data*. Editora en España del repositorio de acceso abierto *e-Lis*, *E-prints in library & information science*.
http://orcid.org/0000-0001-6432-917X

Universitat Politècnica de València, Departament de Comunicació Audiovisual, Documentació i Història de l'Art
Camino de Vera, s/n. 46022 Valencia, España
anfersa@upv.es

Enrique-Alfonso Sánchez-Pérez es catedrático del *Departamento de Matemática Aplicada* de la *Universidad Politécnica de Valencia* (*UPV*) y miembro del *Instituto Universitario de Matemática Pura y Aplicada* de la misma universidad. Es licenciado en matemáticas, ciencias químicas, filosofía, y ciencias de la educación, y doctor en matemáticas por la *UPV*. Es profesor de la *Escuela de Ingenieros de Caminos, Canales y Puertos* de la *UPV*. Su investigación se centra en el análisis matemático, particularmente en temas de análisis funcional, como la teoría de operadores, la integración vectorial y la topología. Ha participado en numerosos proyectos de investigación en estos temas y también en colaboración con otros grupos científicos (física y biomedicina, principalmente), aplicando técnicas matemáticas avanzadas de tratamiento de señales. Ha publicado un centenar de artículos de investigación en revistas científicas de prestigio, principalmente de matemáticas.
http://orcid.org/0000-0001-8854-3154

Universitat Politècnica de València, Institut Universitari de Matemàtica Pura i Aplicada
Camino de Vera, s/n. 46022 Valencia, España
easancpe@mat.upv.es

Rafael Aleixandre-Benavent es científico titular del *Consejo Superior de Investigaciones Científicas* (*CSIC*) y catedrático acreditado de biblioteconomía y documentación. Es doctor en medicina, especialista en documentación médica por la *Universitat de València* (*UV*) y codirector de la *Unidad de Información e Investigación Social y Sanitaria* (*Uisys*) de la *UV*. Ha coordinado el proyecto *OpenDataScience, Centro de recursos para la preservación y gestión de datos abiertos de investigación* (*Odasci*) y participa en la *Red española sobre datos de investigación en abierto-Maredata*. Sus principales líneas de trabajo se centran en la evaluación de la investigación y de las publicaciones científicas, y en los estudios sobre el acceso abierto a los datos científicos.
http://orcid.org/0000-0002-6678-8844

Ingenio (*CSIC-UPV*). *UISYS* (*Universitat de València - CSIC*)
Plaza Cisneros, 4. 46003 Valencia, España
rafael.aleixandre@uv.es

Artículo recibido el 12-04-2016
Aceptación definitiva: 07-06-2016

Fernanda Peset es profesora titular de la *Universidad Politécnica de Valencia* (*UPV*), *Departamento de Comunicación Audiovisual, Documentación e Historia del Arte*. Trabajó en el *Servicio de Información Bibliográfica* de la *Universidad de Valencia* hasta 1999. Doctor por la *Universidad de Murcia* en 2002, es coordinadora del programa de doctorado *Industrias de la comunicación y culturales* de la *UPV*. Es evaluadora para revistas científicas y organismos nacionales de acreditación y evaluación. Su docencia y publicaciones se orientan a la comunicación científica, el acceso abierto y la implantación del protocolo OAI-PMH, normalización de la información, descripción de documentos, sistemas de documentación de museos, y datos de investigación abiertos y enlazados. Participa en proyectos como *IraLIS* y red *Maredata*, así como en el gobierno de *E-LIS* y del *Grupo Ciepi*, el inventario *ODiSEA* o *transparencyscience.es*. Dirige los proyectos de I+D *Datasea* y *Datasea Extended*.
http://orcid.org/0000-0003-3706-6532

Universitat Politècnica de València, Departament de Comunicació Audiovisual, Documentació i Història de l'Art
Camino de Vera, s/n. 46022 Valencia, España
mpesetm@upv.es

Resumen

Revisión de las vías de publicación de datos de investigación, las métricas de evaluación de la publicación y reutilización de datos, y los servicios existentes para medir la reutilización de los datos de investigación. Los datos abiertos de investigación aún no se encuentran incluidos en los indicadores de evaluación de la actividad científica. Éste es uno de los motivos por lo que los investigadores no los tienen incluidos en sus rutinas de trabajo. Igualmente, sus métricas se encuentran en sus primeros pasos. Es necesario incrementar los estudios para que estas métricas sean válidas para los sistemas de evaluación.

Palabras clave

Investigación; Datos; Datos de investigación; Métricas; *Data citation index*; Publicación científica; Publicación de datos de investigación.

Abstract

This article is a summary of the different pathways to publish research data, of metrics for evaluating publishing and reuse, and of the existing services for measuring the reuse of research data**.** It is not yet possible to accurately analyze the metrics for open research data and this is one of the reasons why, in general, researchers do not include them in the assessment of their work. In order to further the use of these tools, it is necessary to increase the number of studies about these evaluation metrics.

Keywords

Research; Data; Research data; Metrics; *Data citation index*; Scientific publication; Publishing open research data.

Ferrer-Sapena, Antonia; **Sánchez-Pérez, Enrique-Alfonso**; **Aleixandre-Benavent, Rafael**; **Peset, Fernanda** (2016). "Cómo analizar el impacto de los datos de investigación con métricas: modelos y servicios". *El profesional de la información*, v. 25, n. 4, pp. 632-641.

http://dx.doi.org/10.3145/epi.2016.jul.13

1. Introducción

En el contexto social actual existe un sentimiento generalizado entre la comunidad científica internacional de la necesidad de compartir los datos de investigación. Son conocidas las ventajas que aporta (**Tenopir** *et al.*, 2011), pero no siempre es aconsejable dejar en abierto todos los datos que se generan en una investigación. Debemos tener en cuenta, tal y como señala el grupo de trabajo *RDA* (*Research Data Alliance*) (*RDA*, sin fecha) que no todos los datos son susceptibles de ser compartidos. Existen áreas de conocimiento donde es mayor el coste de ponerlos en abierto que los beneficios que puede aportar su reutilización. Para la selección de los datos a compartir hay que tener en cuenta la facilidad en la repetición del experimento o prueba para la selección de los datos.

Es rentable hacerlo en aquellas áreas donde se producen datos derivados de grandes volúmenes de experimentación, existen datos derivados de series acumuladas y se emplean modelos generados utilizando el enfoque de sistemas.

En ciencias de la vida, este intercambio de datos es necesario para el beneficio de la comunidad científica, y lo mismo ocurre en campos como genética o cinética enzimática, donde se recomienda la reutilización, ya que en ellos se tiene mayor facilidad para la repetición de los experimentos o pruebas (*RDA*, s.f.). Compartir los datos de investigación ahorra costes e incrementa el valor de la investigación, pero no es una tarea fácil. Requiere un alto esfuerzo que consiste en una inversión en tiempo y recursos por parte de los investigadores, aunque en general el balance entre costes y beneficios es positivo.

El uso de métricas de datos puede crear nuevos incentivos que apoyen su intercambio e incrementar la velocidad en la difusión de la información en numerosas disciplinas. Las métricas ayudan a estudiar cómo se evalúa el resultado de un solo investigador o de un grupo, contribuyendo a una definición más amplia de los logros y la reputación en el ecosistema científico, más allá de la publicación en artículos científicos. El uso de métricas contribuirá a concienciar a las instituciones e infraestructuras para promover la utilización de datos de una manera responsable y que permita la promoción del bien público global (**Lin** *et al.*, 2014).

Uno de los supuestos más importantes que pueden potenciar el crecimiento de las métricas de datos es que el intercambio de éstos sea una fuente de potencial reconocimiento científico para los investigadores. El cuidado y difusión adecuada de los datos se puede considerar una actividad científica que tenga cabida en la evaluación de la investigación y que sirva por tanto como méritos para la contratación, promoción, obtención de fondos, etc.

> No todos los datos son susceptibles de ser compartidos; hay áreas donde es mayor el coste de ponerlos en abierto que los beneficios de su reutilización

El reconocimiento de los artículos publicados proviene generalmente del número de citas recibidas. Varios estudios muestran que los trabajos que comparten los datos reciben mayor número de citas que los estudios similares que no lo hacen (**Piwowar**; **Day**; **Fridsma**, 2007; **Piwowar**; **Vision**, 2013) y son una forma de fomentar la erudición responsable (**Mooney**; **Newton**, 2012). Aunque muchos expertos señalan las ventajas que conllevaría el uso de métricas de datos, su implantación en la evaluación de la actividad científica también puede contribuir a pervertir el sistema, ya que es probable que algunos investigadores prioricen esta práctica para conseguir el máximo número de puntos y lograr el reconocimiento frente a invertir en investigaciones innovadoras (**Ball**, 2015).

Hasta ahora el uso de indicadores bibliométricos ha sido clave en la evaluación de la actividad científica. El Factor de impacto de los *Journal Citation Reports* de *Thomson Reuters*, el cómputo de citas del *Science Citation Index* o el Índice h (índice de Hirsch), han sido utilizados por agencias de financiación, de evaluación y universidades para ponderar el resultado académico de instituciones y científicos. La idea básica de la bibliometría es evaluar el impacto que tienen las publicaciones científicas dentro de la comunidad. A pesar de las numerosas críticas recibidas, los indicadores basados en las citas han sido adoptados mayoritariamente en la evaluación de la ciencia (**Aleixandre-Benavent**; **Valderrama-Zurián**; **González-Alcaide**, 2007; *RDA-WDS*, s. f.). Por este motivo se tiende a buscar indicadores similares para valorar el impacto de los datos, lo que evitaría las posibles reticencias de los productores de datos para compartirlos.

Aunque todavía no hay métricas adecuadas para medir el impacto de los datos de investigación, existe una creciente preocupación para que se creen. Uno de los proyectos que ha abordado el desarrollo de estas métricas es *Making data count: A data metrics pilot project* (**Lin** *et al.*, 2014) en el que han participado *The California Digital Library*, *PLoS*, y *DataONE*. En este proyecto se están estudiando posibles métricas relacionadas con la actividad de los datos, creando mecanismos de testeo automático y un prototipo flexible para la gestión de la vida de los datos (*data life model* – DLM).

2. Estándares de reutilización de datos de investigación

Uno de los principales problemas de los investigadores es cómo encontrar los datos que necesitan, ya que aún no se ha elaborado ningún buscador (**Lin** *et al.*, 2014). En algunos casos los datos se pueden localizar en la Red, mientras que en otros están incluidos en el artículo al que están asociados (**Brase**; **Farquhar**, 2011), o depositados en repositorios específicos de datos (**Ferrer-Sapena** *et al.*, 2016).

Para que los investigadores puedan encontrar de forma exacta y precisa los datos que necesitan, es necesario que éstos vayan acompañados de los metadatos adecuados. Los metadatos deben proporcionar todos los detalles sobre el origen y la manipulación que han experimentado para evitar un uso inadecuado o una interpretación errónea. Además, es necesario utilizar estándares en los formatos que faciliten que el intercambio se produzca de manera eficaz.

Los métodos aplicados para el intercambio y reutilización pueden tener diferentes enfoques. Cada investigación debe disponer de su propia estrategia en función de para quién deban estar disponibles. En general, existen dos formas de hacer posible el intercambio:

- depositando los datos en repositorios, ya sea el de la propia institución o de terceros cuando la organización no dispone de ellos;
- depositándolos en las webs de las revistas.

El intercambio directo de datos entre investigadores suele producirse cuando no existen repositorios específicos en sus áreas temáticas. Asimismo, puede ser adecuado el intercambio entre comunidades cerradas cuando existan condicionantes éticos o de confidencialidad que así lo aconsejen.

> Uno de los principales problemas de los investigadores es cómo encontrar los datos que necesitan, ya que aún no se ha implementado ningún buscador

El *Biotechnology and Biological Sciences Research Council* (*Bbsrc*, 2007) indica que en su campo de conocimiento es recomendable guardar los datos de investigación por un período de diez años. Aconsejan que sean compartidos de manera inmediata una vez sean publicados los primeros resultados y, si se liberan previamente, las normas éticas recomiendan que no se publique ningún estudio hasta que los propios autores de los datos hayan publicado algunos resultados de estos datos recogidos. La escala de tiempo para

liberarlos depende de cada comunidad científica. El *Bbsrc* propone los siguientes plazos para liberarlos en función de las áreas temáticas de biotecnología (*Bbsrc*, 2007):

> "en el área de cristalografía (*Protein Data Bank*), se ha acordado un plazo máximo de 12 meses desde la publicación del primer documento sobre una estructura; y en los estudios sobre secuenciación (base de datos de secuencias de nucleótidos *EMBL*), los datos pueden ser retenidos hasta la publicación de los resultados, pero no después".

La reutilización de determinados tipos de datos puede conllevar problemas derivados de la propiedad intelectual y de la comercialización potencial a partir de las ideas patentables generadas con el nuevo conocimiento. Sin embargo, el *Bbsrc* considera que este tipo de negocio no debe demorar el intercambio de datos ni excluirlo. De la misma manera considera que los datos que provengan del estudio de grandes magnitudes deben ser puestos en abierto de manera inmediata a medida que se encuentren disponibles o se publiquen.

3. Modelos de publicación de datos

De acuerdo con **Parsons** y **Fox** (2013), se pueden señalar los siguientes cinco modelos:

Modelo basado en la publicación científica

Este enfoque es semejante al de la publicación científica, es decir, su métrica está basada en el recuento de citas. El conjunto de datos debería estar bien descrito y con un cierto nivel de control de la calidad o de revisión por pares. Aunque es uno de los modelos más maduros, cuenta con numerosas limitaciones, ya que no tiene claramente definido qué se entiende por publicación de datos (**Parsons**; **Fox**, 2013). Este modelo también se está analizando en el *Bibliometrics Working Group* de la *Research Data Alliance* (*RDA*) (**Callaghan**; **Lehnert**, s.f.), donde se trabaja en el fomento de más y mejores citas de datos para incrementar su disponibilidad y calidad. Al igual que en la publicación científica, para que la publicación de datos sea eficaz, los datos deben satisfacer los siguientes criterios:

- persistencia
- longevidad
- sostenibilidad
- calidad.

Otros problemas se relacionan con la granularidad, control de las versiones, y los aspectos legales.

Big iron

Este enfoque proviene de la cultura de ingeniería y, por lo general se ocupa de grandes volúmenes de datos que son relativamente homogéneos, bien definidos y dinámicos. Necesita de una gran infraestructura, sofisticada y bien controlada. Involucra a centros de supercomputación y sus interfaces son especializadas. Dispone de estándares de metadatos y suele utilizar estructuras de datos relacionales y jerárquicas con esquemas organizativos. Algunas instituciones que aplican este modelo son la *NASA* y la *Agencia Europea del Espacio* (*ESA*).

Tabla 1. Modelos de publicación de datos y principales características.

	Modelo basado en la publicación científica	***Big iron***	***Science support***	***Map making***	***Linked data***
Analogía	Publicación académica	Producción industrial	Producción artesanal del trabajo	Cartografía	Creación en la WWW
Características de los datos	Pequeño volumen, diversas formas, escalas y temas,	Alto volumen y más homogéneos en la forma	Pequeños y diversos	Geoespaciales, características y atributos	Entidades iguales nombradas de distintas maneras
Modelos organizativos de los datos	Jerárquico o relacional	Jerárquico	Geoespacial, jerárquico y relacional	Geospacial y relacional	Basados en mapas y datos enlazados
Fin primario	Calidad de los datos, certificación y preservación	Rendimiento y acceso manejable	Síntesis de los datos y reproducibilidad	Visualización basada en mapas e intercomparación	Interoperabilidad e interconexión
Normas	Cita de datos	Formatos de datos, control de versiones	Procesos locales	Sistemas de referencia de coordenadas, transformaciones espaciales	Ontologías
Ejemplos en la ciencia	*Pangea*, repositorios universitarios *Eosdis*	*Eosdis*	*LTER*	*Inspire*, *Geodata.gov*	*Integrated Ocean Drilling Program - IODP*, *MyGrid*, *Linked Open Government Data*
Terminología	Datos de autor, editor, citación de datos	Productor de datos, nivel de procesamiento, versión de lanzamiento	Colector de datos, personal de apoyo	Fuente de datos, capas	Proveedor de datos, nombre, enlace, recurso
Contexto cultural	Bibliotecas y grupos de investigación universitarios (por ejemplo, la *NSF*)	Ingeniería de sistemas y gestión de proyectos (por ejemplo, la *NASA*, el *Departamento de Defensa*)	Basado en el lugar de investigación (por ejemplo, *NSF*)	Uso y manejo de la tierra (por ejemplo, *United States Geological Survey -USGS*, e *Integrated Ocean Drilling Program - IODP*)	Aplicaciones informáticas y comerciales (por ejemplo, *National Science Fundation - NSF* y *Common Information Sharing Environment - CISE W3C*)

Fuente: **Parsons** y **Fox** (2013)

Science support

Se denomina así la estructura de apoyo operativo que se encuentra en un centro de investigación o en un laboratorio. Dado que se trata de estaciones o centros que realizan investigaciones de larga duración, la gestión de los datos se considera una función necesaria de la infraestructura para el apoyo a la investigación. Este modelo lo aplican organizaciones como la red *Long Term Ecological Research* (*LTER*).

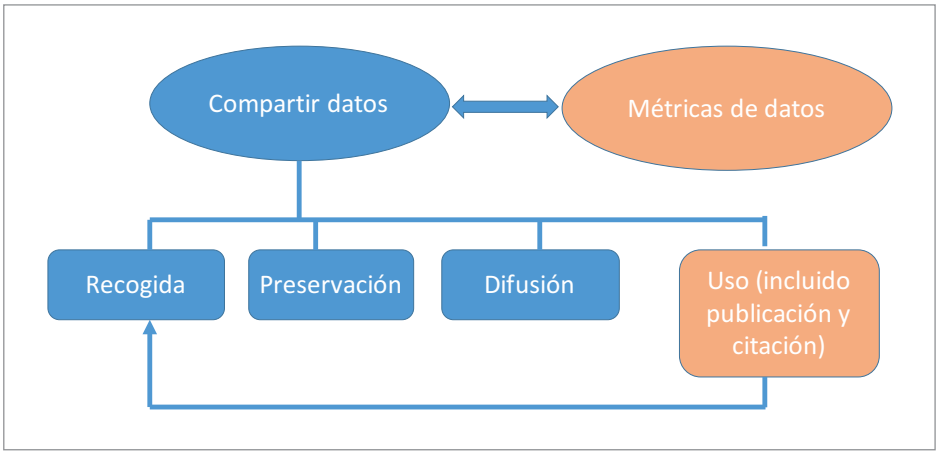

Figura 1. Esquema de las relaciones entre intercambio de datos y sus métricas. Fuente: **Costas** *et al.*, 2013

Map making

Son las infraestructuras de datos espaciales y los sistemas de información geográfica (SIG) asociados. La realización de un mapa puede ser visto como un subconjunto de la publicación de datos, pero en este contexto se publica un mapa o atlas.

Linked data

Se basa en la web de los datos y se fundamenta en el principio del diseño de la web semántica. El objetivo de este enfoque es más la interoperabilidad y la capitalización de la web interconectada que la preservación, la curación o la calidad. Aquí el énfasis se pone en la representación de los datos.

4. Situación actual

En la actividad académica se han valorado hasta el momento únicamente los artículos, pero no la calidad de los datos. Los datos pueden ir unidos o no a los artículos pero éstos deben encontrarse debidamente estructurados. Son heterogéneos y varían desde pequeñas unidades, como tablas individuales, a grandes conjuntos de datos. Este hecho dificulta su cita, que puede hacerse del conjunto o de una parte, lo que depende también del campo científico del que se trate.

Es necesario disponer de un marco abierto para la validación de los datos y poder compararlos, así como permitir la implementación de programas que agreguen valor a los datos en bruto. Las métricas deben estar abiertas tanto para los artículos como para los conjuntos de datos y permitir la reutilización en la mayor medida posible.

Existen formas tradicionales para medir el impacto de los datos como *Data Citation Index* de *Thomson Reuters*. En esta base de datos no se incluye la utilización y reutilización de los conjuntos de datos, sino únicamente su cita, por lo que su alcance tiene limitaciones. También tiene un retraso significativo de las métricas que proporciona debido a que los datos citados dependen de la publicación del artículo.

En el estudio *The value of research data. Metrics for datasets from a cultural and technical point of view* (**Costas** *et al.*, 2013), en el que se exponen las métricas disponibles, las percepciones de los usuarios sobre las mismas, los problemas existentes y las posibles soluciones, se indica que para identificar claramente las métricas de datos se debe precisar la terminología. Así "métrica de datos" comprende todo lo relacionado con la publicación de los datos y sus citas, incluyendo también algunos indicadores altmétricos. Tanto la publicación como la cita pueden ser consideradas como signos indicativos de uso de los datos. El uso puede generar nuevos datos que pueden alimentarlos de nuevo. En la figura 1 se muestran las relaciones entre el intercambio de datos y sus métricas.

En el proceso de publicación de los datos y sus métricas se encuentran implicados varios agentes. En la figura 2 se pueden ver los más representativos.

Cada agente tiene, por motivos distintos, interés en conocer las métricas de datos:

- las organizaciones financiadoras para conocer el impacto de lo que han financiado;
- las estructuras de investigación para potenciar el acceso y la preservación;
- los investigadores, como una medida más del reconocimiento e impacto de su trabajo;

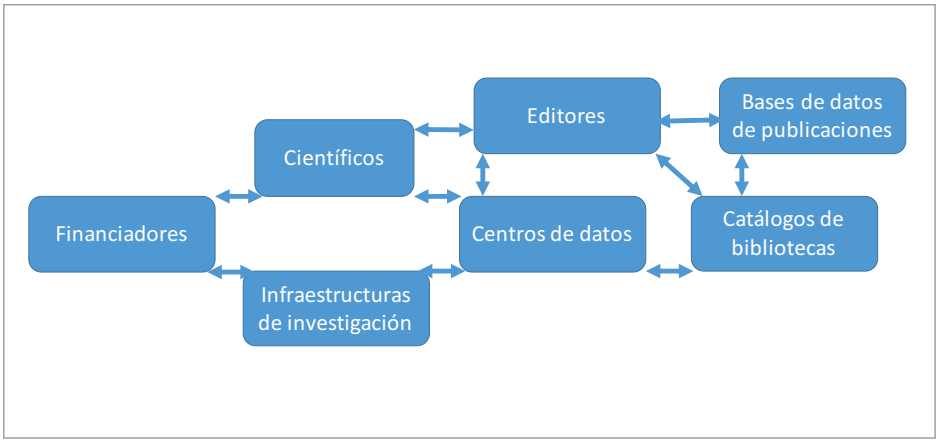

Figura 2. Agentes implicados en la publicación de datos. Fuente: **Costas** *et al.*, 2013

- los centros de datos, para estandarizar el almacenamiento y crear metadatos armonizados a nivel mundial, efectuar un seguimiento de la reutilización y promover buenas prácticas, ofrecer datos sobre la curación y recomendar la cita a los conjuntos de datos;
- los editores, para hacer frente a los datos que se adjuntan a los artículos, facilitar políticas y normas para la cita y métrica de los datos;
- las bibliotecas, para hacer los datos accesibles e identificables y establecer una coordinación entre los investigadores y los centros de datos;
- las bases de datos de publicaciones, para unir las publicaciones con las citas de los datos y habilitar indicadores y recuento de citas.

Otra revisión interesante es la guía publicada en 2015 por el *Digital Curation Center* (*DCC*), *How to measure the impact of research data* (**Ball**; **Duke**, 2015), que quiere servir de referencia para medir el impacto de los datos de investigación de las instituciones. Ofrece una panorámica general de los conceptos clave necesarios para medir el impacto, así como los servicios y programas disponibles.

> Como los investigadores únicamente obtienen reconocimiento por los artículos publicados, no invierten tiempo en otras actividades que pueden beneficiar a largo plazo su campo de estudio

Otro grupo que está abordando esta problemática junto al *Bibliometrics Working Group* de *Research Data Alliance* (**Callaghan**; **Lehnert**, s.f.) es el *Grupo de Trabajo de Datos* del *Consortium Advancing Standards in Research Administration Information* (*Casrai*), que está trabajando en proyectos de gestión de datos de investigación, abordando los temas de terminología, vocabularios e indicadores. Los grupos de interés detectados, al igual que en otros estudios, son: editores, organismos de financiación, administradores de las universidades y repositorios de datos.

Otros dos estudios a destacar son *Joint declaration of data citation principles* (**Martone**, 2014) y *Data citation implementation pilot* (*DCIP*, s.f.). Los principios de la *Joint declaration of data citation principles* reconocen la doble necesidad de crear prácticas de citación que sean comprensibles para los investigadores y para los ordenadores. No pretenden crear estándares para implementaciones específicas, sino alertar a las comunidades de la necesidad realizar prácticas y aplicaciones que incorporen estos principios. Por su parte, el objetivo principal del *DCIP* es proporcionar coordinación entre editores, repositorios y servicios de metadatos y de identificadores para abordar el problema de los investigadores que están empezando a citar datos, así como guías y grupos de consulta que ayuden a lograr este propósito.

5. Modelos de métricas de datos

Como se ha dicho antes, uno de los problemas es que si los investigadores únicamente obtienen reconocimiento y recompensa por los artículos publicados, eso les disuade de invertir tiempo en otras actividades que pueden beneficiar a largo plazo su campo de estudio. Ésta es una de las razones por las que se están examinando otras métricas para evaluar el impacto de la investigación que vayan más allá del mero recuento de citas.

Los investigadores pueden tener impacto en su disciplina no únicamente por las citas que han recibido sus artículos, sino también por otros reconocimientos a su labor como (**Ball**; **Duke**, 2015):

- reutilización de los datos que comparten;
- recomendación de la lectura de sus publicaciones;
- invitaciones a intervenir en los medios de comunicación;
- uso que se hace del software que han escrito.

A la hora de medir el impacto se debe tener en cuenta el alcance de lo que se quiere medir. Las métricas pueden servir para indicar o sugerir el impacto que ha tenido un conjunto de datos o *dataset*, pero esta medida puede que no sea del todo fiable, dado el elevado número de factores que intervienen en la consideración del impacto. Por ello, al igual que ocurre en la interpretación del impacto de las publicaciones, es necesaria la interpretación humana para evaluar la calidad de los *datasets*, ya que no siempre el hecho de que se cite mucho significa que sea bueno, sino que puede ser citado por lo contrario. Así, que un *dataset* sea muy mencionado en los medios sociales puede ser debido a que tiene un impacto positivo o negativo en su disciplina (**Ball**, 2015).

Existe muy poca bibliografía acerca de métricas de datos, hecho ya señalado por **Costas** *et al.* (2013), que indican que todavía no existen indicadores de uso. A continuación se presentan posibles métricas basadas en modelos existentes que pueden ser útiles para el desarrollo de las de datos. La tabla 2 muestra las métricas y su relación con el actual modelo de publicación científica.

Como puede observarse, uno de los modelos se basa en la publicación y cita de datos. Aunque es uno de los modelos más maduros (**Ball**; **Duke**, 2015), plantea numerosos problemas, sobre todo debidos a las estructuras necesarias para la revisión por pares que garanticen la calidad de los *datasets* y la cita formal de los conjuntos de datos (**Lawrence** *et al.*, 2011). Hay que recordar que la cita es uno de los principales elementos de reconocimiento académico y es un elemento clave para la identificación, recuperación, replicación y verificación de los datos en los que se basan los estudios publicados (**Mooney**; **Newton**, 2012).

Como hemos señalado, un aspecto muy importante en el depósito en repositorios es que se garantice la calidad de los datos depositados. Sin embargo, existen repositorios donde no existe un protocolo de evaluación de la calidad de los datos que incluyen (**Brase** *et al.*, 2009). Esta evaluación puede llegar a ser muy importante cuando el usuario potencial no tiene la suficiente experiencia y capacidad para evaluarlos. Se ha propuesto la revisión por pares como sistema de control de calidad de los conjuntos de datos (**Lawrence** *et al.*, 2011); no obstante, esta revisión humana tradicional puede ser apropiada para ciertos conjuntos de datos, pero es demasiado lenta para evaluar la avalancha actual de datos.

Este modelo tiene, como se ha comentado, numerosas limitaciones que todavía no han sido resueltas (**Costas** *et al.*,

Tabla 2. Escenarios de métricas de modelos de publicación de datos (comparativa con el actual modelo de publicación científica)

Tipos de métricas	Herramientas disponibles con las posibles métricas de datos	Dimensiones de las métricas	Modelos			
			Publicación científica	Publicación de datos	Publicación de datos	
					Publicación de datos solos	Revistas de publicación de datos
Publicación de datos y métricas basadas en citas	*Data Citation Index* (*Web of Science*)	Tamaño dependiente	Sí	Difícil	Sí	Sí
	Google Scholar	Tamaño independiente				
	Scopus	Rendimiento promedio directo	Sí	No	Sí	Sí
	Microsoft Academic Search *DataCite*	Rendimiento basado en la fuente	Sí	No	Sí	Sí
Métricas basadas en altmétricas	*ImpactStory* *Twitter* *Facebook*	Indicadores social media	Sí	Sí	Sí	Sí
	Mendeley *CiteULike*	Número de lectores	Sí	No	Sí	Sí
	Repositorios *Data Journals*	Descargas, número de lecturas (Métricas *DUI - data usage index*)	Sí	Difícil	Sí	Sí

Fuente: **Costas** *et al.*, 2013

2013). Por un lado el concepto de cita de datos no se encuentra estandarizado y aceptado. Por otro, no se encuentra generalizada la cita directa a los conjuntos de datos, de manera que en algunas citas se describen los conjuntos de datos y sus colecciones sin sacar conclusiones científicas, mientras que en otros casos los datos se publican en una sección especial de una revista o en una revista de datos, como *Journal of open archaeology data* (**Ball**; **Duke**, 2015). En muchas disciplinas predomina el enfoque de citar el artículo que hace uso de los datos, confiando en que el artículo indique si se han compartido los datos y cómo se ha hecho. Generalmente no es posible sin un esfuerzo manual considerable, identificar las citas a estos artículos que deben contarse a través de su presencia en el texto. En estas disciplinas por lo tanto, el número de citas es de poca ayuda como indicador del impacto de los datos y por ello deben encontrarse métricas alternativas.

> El concepto de cita de datos no se encuentra estandarizado y aceptado

No existe aún una teoría bibliométrica asentada sobre el valor de la cita de los datos que sea útil para la evaluación. Además, existen importantes limitaciones tecnológicas que restringen el desarrollo de la publicación y cita de datos y por consiguiente sus métricas. Estas limitaciones se deben a la incompatibilidad entre los softwares y los repositorios, las estructuras de los datos, su almacenamiento y su gestión.

Aunque este modelo es el más desarrollado, tal y cómo se ha comentado, no aporta grandes avances en este campo; ni siquiera han sido implementados por parte de infraestructuras tan conocidas como *DataCite*, aunque se encuentran trabajando en métricas y modelos para el reconocimiento de la publicación y *data sharing*.

Las métricas que se establecen para este modelo son semejantes a las de la bibliometría tradicional y tienen en cuenta el impacto de cada uno de los autores, de las instituciones a las que pertenecen, del proyecto que desarrollan y de la institución que financió la investigación. También se pueden sacar conclusiones sobre el impacto de la entidad frente a sus competidores.

Ball (2015) distingue tres tipos de métricas basadas en citas:

- número total de citas de un investigador o de un grupo;
- índice h;
- rendimiento promedio: promedio del número de citas en relación al promedio de la población analizada.

6. Altmétricas

De acuerdo con la *NISO* (2016b):

> "*Altmétricas* es un término amplio que incluye la recogida, creación y uso de múltiples formas de evaluación obtenidas a partir de la actividad y la relación entre los diversos actores y la producción científica en el ecosistema de la investigación".

En las métricas tradicionales no se tiene en cuenta el comportamiento del lector online, la interacción con los contenidos en la Red o las referencias en los medios sociales. Tampoco contemplan las nuevas formas de producción científica, como los conjuntos de datos publicados en repositorios, los algoritmos, las estructuras moleculares o

el software que se comparte en *GitHub*. Estas nuevas producciones son difíciles de evaluar a través de las métricas tradicionales, pues en ellas no existe una cultura de la cita (*NISO*, 2016b).

Los indicadores altmétricos pueden aplicarse a los repositorios de datos, ya que tienen en cuenta el número de visualizaciones, descargas, menciones en las redes sociales y otros indicadores que pueden adaptarse bien.

Existe una preocupación acerca de si las técnicas para citar y medir el impacto de los artículos de revistas científicas pueden no ser las más apropiadas para medir el impacto de otros tipos de productos de la investigación, software o datos.

Dentro del apartado destinado a las altmétricas, el *Grupo de Trabajo B* de la *NISO* estudia los productos de investigación no tradicionales y sus identificadores. Han publicado recomendaciones dirigidas a instituciones, administradores de repositorios, organizaciones de investigación internacional y agencias financiadoras sobre cómo citar datos y sobre las métricas de datos. Entre otras, se indica que éstas deberían estar disponibles tan pronto como sea posible, deberán estar relacionadas con la cita, usar identificadores persistentes, metadatos y aportar una página de destino (*NISO*, 2016a).

A pesar de las críticas del sistema basado en el cómputo de citas, se han utilizado métricas basadas en él (**Ball**, 2015), como:

- indicadores de las redes sociales: menciones o *likes*;
- contabilización de lectores;
- descargas y visualizaciones, es decir, el número de veces que se accede a un registro.

Por otra parte, las estadísticas de uso no se correlacionan directamente con las citas, ya que reflejan diferentes comportamientos de lectura que pueden no tener nada que ver con la cita a un trabajo de investigación. Se ha observado que el número de lectores de *Mendeley* parece correlacionarse mejor con el número de citas que recibe un trabajo; sin embargo, la correlación no es lo suficientemente fuerte como elemento de predicción de futuras citas.

Entre las principales críticas de las altmétricas deben mencionarse las siguientes:

- el concepto no está claramente definido y el término tiene significados distintos según quien lo trate;
- se están convirtiendo en objeto de muchos estudios, pero de momento no en una alternativa;
- se encuentran más bien relacionadas con las métricas de los social media.

A pesar de las críticas, las altmétricas se consideran un elemento clave para las métricas de los datos de investigación, aunque aún no se pueden observar en la mayoría de webs de los principales repositorios de datos de investigación.

7. Servicios de medida de impacto

Existen muy pocos servicios de medida del impacto de los datos de investigación. Citaremos el *Data Citation Index* y los repositorios de datos.

Data Citation Index

En octubre de 2012, *Thomson Reuters* lanzó el *Data Citation Index* (*DCI*) como parte de los servicios de *la Web of Science*. El *DCI* recoge las citas de los datos de los artículos indexados por la *Web of Science*. La información que contiene proviene de los repositorios de datos y de las prácticas de citación de datos, las cuales son inconsistentes en muchos casos (**Robinson-García**; **Jiménez-Contreras**; **Torres-Salinas**, 2015). En la actualidad el *DCI* se basa en la información ofrecida por el repositorio de datos con respecto a publicaciones en las que se cita el conjunto de datos o el estudio de los datos. En cada registro el *DCI* muestra el número de veces que el objeto ha sido citado en la *Web of Science*. El *DCI* admite tres tipos de tipos de documentos:

- *data sets*
- *data studies*
- repositorios.

Como con los artículos, existen diferencias entre las áreas de conocimiento: las citas son mayores en ingeniería y tecnología que en ciencias sociales, arte y humanidades, aunque en éstas el estudio de datos es importante. Las citas de datos se encuentran más extendidas en disciplinas como la cristalografía y la genómica. Los *datasets* que contiene el *Data Citation Index* se encuentran muy sesgados hacia las ciencias puras y experimentales. El 94% de ellos provienen del 75% de las bases de datos *Gene Expression Omnibus*, *UniProt Knowledgebase*, *Pangaea* and *U.S. Census Bureau Tiger/Line Shapefiles* (**Torres-Salinas**; **Martín-Martín**; **Fuente-Gutiérrez**, 2014).

El *DCI* sigue un proceso de selección con el fin de mantener ciertos estándares de calidad. Los criterios seguidos para la indexación de los repositorios incluyen factores como:

- normas de publicación;
- contenido editorial;
- diversidad internacional de la autoría;
- procedencia geográfica;
- citas recibidas.

Repositorios de datos

Aportan poca información sobre métricas de los *datasets* depositados. Se han consultado algunos y únicamente ofrecen información sobre visualizaciones, descargas y en algún caso estadísticas de carácter general. Por ejemplo:

- *Figshare*: visualizaciones y descargas;
- *Dryad*: descargas;
- *Open Aire*: información estadística general sobre visualizaciones, bancos de datos más citados, publicaciones por tipo, país, idioma y proveedor.

Otros productos que dan información sobre métricas de datos (**Ball**; **Duke**, 2015) son *ResearchGate* e *ImpactStory*, que además proporcionan información sobre las publicaciones y sobre *datasets* estadísticos. *PLoS* por su parte aporta estadísticas de los artículos que ha publicado, ofreciendo tanto información sobre citas en *Web of Science* y *Scopus*, como sobre visualizaciones y descargas, comentarios y calificaciones en los medios sociales, en algunos blogs o en *Wikipedia*.

Otro producto de estas características es *PlumX*, que fue adquirido por *Ebsco* a principios de 2014. Lo produce *Plum Analytics*, el principal proveedor de altmétricas de investigación. Ofrece una imagen completa del impacto de la investigación, incluyendo conjuntos de datos y códigos fuente, así como las publicaciones tradicionales. El producto va más dirigido a instituciones que a personas. Ofrece métricas de cinco categorías:

- uso: contabiliza el número de veces que un recurso ha sido visto o descargado, enlaces recibidos, clics, usuarios que contribuyen en *GitHub*, bibliotecas que tienen una copia;
- captura: veces que el recurso ha sido marcado como de interés (por ejemplo, un libro marcado en *Delicious*, añadido a una biblioteca en *Mendeley*, seguido, enlazado o visto en *GitHub*);
- menciones: posts en blogs, comentarios hechos (en *Facebook*, *SlideShare*, *YouTube*, etc.), o recibidos (en *Amazon* o *GoodReads*);
- medios de comunicación social: veces que se ha recomendado el recurso (por ejemplo "me gusta" en *Facebook*, "+ 1" en *Google+*, votos en *Reddit*, tweets...).
- citas: número de citas que ha recibido en la *WoS*, *Scopus* y otras bases de datos.

> "Es necesario encontrar métricas que permitan salir del círculo vicioso en el que nos encontramos"

8. Conclusiones

Actualmente aún es escasa la reutilización de datos en todas las áreas de conocimiento, y las métricas todavía no se encuentran lo suficientemente desarrolladas.

Un problema con el que se enfrenta la evolución de las métricas basadas en la cita de datos es la escasa conciencia que existe para incluir su cita en las publicaciones. La citación se incrementará si se reconoce en la evaluación de la actividad científica o si los editores ven un modelo de negocio en ello. Puede influir también favorablemente el que los organismos financiadores lo tengan en cuenta a la hora de financiar proyectos o como elemento obligatorio para incrementar su impacto y transferencia.

Otros problemas que hay que resolver a la hora de proponer métricas es el gran número de bancos de datos existentes y su dispersión (más de 1.300 indexados por *re3data*), y la necesidad de elaborar estándares y protocolos de interoperabilidad entre los distintos actores.

Los indicadores altmétricos, que podrían ser adecuados para medir el impacto de los *datasets*, no se encuentran lo suficientemente desarrollados y aceptados por la comunidad académica.

Es necesario encontrar métricas adecuadas que permitan salir del círculo vicioso en el que nos encontramos, en el que los investigadores no comparten sus datos porque esta práctica no está reconocida y, por lo tanto, no les vale la pena invertir tiempo. Implementar métricas es difícil, y sin métricas cuesta estimular el intercambio de datos (**Ball**, 2015).

9. Bibliografía

Aleixandre-Benavent, Rafael; **Valderrama-Zurián, Juan-Carlos**; **González-Alcaide, Gregorio** (2007). "El factor de impacto de las revistas científicas: limitaciones e indicadores alternativos". *El profesional de la información*, v. 16, n. 1, pp. 4-11.
http://eprints.rclis.org/9489

Ball, Alex (2015). "Better data-level metrics: Quality and impact". En: *Workshop held as part of the BioMedBridges Symposium 'Open bridges for life science data' at the Wellcome Genome Campus*, Hinxton, UK, 17-18 November 2015, pp. 17-18.
http://opus.bath.ac.uk/48224
http://dx.doi.org/10.7490/f1000research.1111029.1

Ball, Alex; **Duke, Monica** (2015). "How to track the impact of research data with metrics". *DCC how-to guides & checklists*. Edinburgh: Digital Curation Centre.
http://www.dcc.ac.uk/resources/how-guides
http://www.dcc.ac.uk/resources/how-guides/track-data-impact-metrics#sthash.U85HBEIb.dpuf

Bbsrc (2007) "Bbsrc data sharing policy: version 1.2". *Bbsrc, bioscience for the future*.
http://www.bbsrc.ac.uk/datasharing

Brase, Jan; **Farquhar, Adam** (2011). "Access to research data". *D-lib magazine*, v. 17, n. 1-2.
http://dx.doi.org/10.1045/january2011-brase

Brase, Jan; **Farquhar, Adam**; **Gastl, Angela**; **Gruttenmeier, Herbert**; **Heijne, Maria**; **Heller, Alfred**; **Piquet, Arlette**; **Rombouts, Jeroen**; **Sandfaer, Mogens**; **Sens, Irena** (2009). "Approach for a joint global registration agency for research data". *Information services and use*, v. 29, n. 1, pp. 13-27.
http://orbit.dtu.dk/files/4261926/Publication%20DataCite%20J.B..pdf
http://dx.doi.org/10.3233/ISU-2009-0595

Callaghan, Sarah; **Lehnert, Kerstin** (s.f.). *Bibliometrics working group. Publishing Data Bibliometrics*.
https://rd-alliance.org/sites/default/files/attachment/1_PublishingDataBibliometrics_SarahCallaghan.pdf

Costas, Rodrigo; **Meijer, Ingeborg**; **Zahedi, Zohreh**; **Wouters, Paul** (2013). *The value of research data. Metrics for datasets from a cultural and technical point of view*. Copenhagen: Knowledge Exchange.
http://www.knowledge-exchange.info/datametrics

DCIP (s.f.). *Data Citation Implementation Pilot*
https://www.force11.org/group/dcip

Ferrer-Sapena, Antonia; **Aleixandre-Benavent, Rafael**; **Peset, Fernanda**; **Vidal-Infer, Antonio**; **Alonso-Arroyo, Adolfo** (2016). "Los científicos ante los datos abiertos. Red Maredata". En: *II Encuentro de datos abiertos de la Universidad de Alicante*.
http://es.slideshare.net/uadatos/presentacion-alicante-59894628

Lawrence, Bryan; **Jones, Catherine**; **Matthews, Brian**; **Pepler, Sam**; **Callaghan, Sarah** (2011). "Citation and peer review of data: Moving towards formal data publication". *The international journal of digital curation*, v. 6, n. 2, pp. 4-37.

http://dx.doi.org/10.2218/ijdc.v6i2.205

Lin, Jennifer; **Cruse, Patricia**; **Fenner, Martin**; **Strasser, Carly** (2014). *Making data count: A data metrics pilot project*. University of California. California Digital Library.
http://escholarship.org/uc/item/9kf081vf

Martone, Maryann (ed.) (2014). *Data Citation Synthesis Group: Joint declaration of data citation principles*. San Diego CA: Force11.
https://www.force11.org/group/joint-declaration-data-citation-principles-final

Mooney, Hailey; **Newton, Mark P.** (2012). "The anatomy of a data citation: Discovery, reuse, and credit". *Journal of librarianship and scholarly communication*, v. 1, n. 1, eP1035.
http://dx.doi.org/10.7710/2162-3309.1035

NISO (2016a). *Alternative outputs in scholary communications: Data metrics*. National Information Standards Organization.
http://goo.gl/BHe7qq

NISO (2016b). *Altmetrics definitions and use cases*. National Information Standards Organization.
http://goo.gl/WcsBHr

Parsons, Mark; **Fox, Peter** (2013). "Is data publication the right metaphor?". *Data science journal*, v. 12, pp. WDS32–WDS46.
https://www.jstage.jst.go.jp/article/dsj/12/0/12_WDS-042/_pdf

Piwowar, Heather A.; **Day, Roger S.**; **Fridsma, Douglas B.** (2007). "Sharing detailed research data is associated with increased citation rate". *PLoS one*, 21 March.
http://dx.doi.org/10.1371/journal.pone.0000308

Piwowar, Heather A.; **Vision, Todd J.** (2013). "Data reuse and the open data citation advantage". *PeerJ*, 1:e175.
https://doi.org/10.7717/peerj.175

RDA (s.f.). "Data re-use, share your experiences". *Research Data Alliance*.
https://rd-alliance.org/groups/data-re-use-share-your-experiences.html

RDA-WDS (s.f.). *Cost recovery for data centres Working Group*, *Case statement*. RDA-WDS Publishing Data Interest Group. Bibliometrics Working Group.
https://rd-alliance.org/sites/default/files/case_statement/RDA_WDS_IG_Publishing_Costs.pdf

Robinson-García, Nicolás; **Jiménez-Contreras, Evaristo**; **Torres-Salinas, Daniel** (2015). "Analyzing data citation practices using the data citation index". *Journal of the Association for Information Science and Technology*.
https://arxiv.org/abs/1501.06285
http://dx.doi.org/10.1002/asi.23529

Tenopir, Carol; **Allard, Suzie**; **Douglass, Kimberly**; **Aydinoglu, Arsev U.**; **Wu, Lei**; **Read, Eleanor**; **Manoff, Maribeth**; **Frame, Mike** (2011). "Data sharing by scientists: Practices and perceptions". *PLoS one*, v. 6, n. 6, p. e21101.
http://dx.plos.org/10.1371/journal.pone.0021101

Torres-Salinas, Daniel; **Martín-Martín, Alberto**; **Fuente-Gutiérrez, Enrique** (2014). "Analysis of the coverage of the data citation index - Thomson Reuters: disciplines, document types and repositories". *Revista española de documentación científica*, v. 37, n. 1, e036.
http://dx.doi.org/10.3989/redc.2014.1.1114