bid textos universitaris de biblioteconomia i documentació
ISSN 1575-5886
DL B-19.675-1998
número 29
desembre de 2012
Facultat de Biblioteconomia i Documentació
Universitat de Barcelona

ODiSEA: *International Registry on Research Data*

[Versió catalana]

Alicia García-García
Estudiante predoctoral. Instituto Universitario de Investigación Dr. Viña Giner
Universidad Católica de Valencia
alicia.garcia@ucv.es

Josep-Manuel Rodríguez-Gairín
Profesor del Departament de Biblioteconomia i Documentació
Universitat de Barcelona
rodriguez.gairin@ub.edu

Tomás Saorín
Profesor de la Facultad de Comunicación y Documentación
Universidad de Murcia
tsp@um.es

Luís-Millán González
Profesor del Departamento de la Actividad Física y Deportiva
Universidad de Valencia
luis.m.gonzalez@uv.es

Xavi García-Massó
Profesor del Departamento de Fisioterapia
Universidad de Valencia
xavier.garcia@uv.es

Antonia Ferrer Sapena, Fernanda Peset
Profesoras del Departamento de Comunicación, Documentación e Historia del Arte
Universidad Politécnica de Valencia
mpesetm@upv.es, *anfersa@upv.es*

Resumen [Abstract] [Resum]

El artículo revisa los temas principales en la preservación y reuso de los datos de investigación (beneficios, ciclo de vida, proyectos, normativas, etc.) e identifica la falta de un registro mundial de bancos, repositorios y bibliotecas de datos. Expone la creación de una herramienta web que recoja este tipo de depósitos y los clasifique por áreas disciplinares: ODiSEA, *International Registry on Research Data*. Ofrece resultados sobre número y tipología temática de este tipo de depósitos a escala mundial. Esta aportación facilita el descubrimiento de nuevos conjuntos de datos cuya recombinación, desde una perspectiva multidisciplinar, fomentará la innovación y la rentabilidad de la inversión en ciencia.

1 Introducción

En los últimos años, en la bibliografía se ha debatido el beneficio del acceso y la preservación de

las publicaciones científicas y académicas a través de repositorios de acceso abierto. Pero el producto de la actividad investigadora incluye también datos, identificados como materiales suplementarios, en la Declaración de Bethesda y de Berlín, ambas de 2003, sobre acceso abierto al conocimiento en ciencias y humanidades. Estos materiales deben ser preservados y puestos a disposición de la sociedad como recursos que pueden ser de utilidad social para el procomún (*commons*). De hecho, para Abella (2011) "los datos deben ser la infraestructura de la economía digital", y tienen que permitir la creación de empleo e innovación. Esto es lo que pretende el movimiento *open data*, refrendado por la Directiva europea 2003/98/CE, de reutilización de información en el sector público, y la Ley 37/2007: liberar el valor social y económico de los contenidos, datos y documentos que se encuentran en poder de las administraciones públicas (Ferrer-Sapena; Peset; Aleixandre-Benavent, 2011).

La reutilización de la información del sector público es un elemento cada vez más importante en las estrategias de *open data* y *open government* (Marcos-Martín; Soriano-Maldonado, 2011) y puede extenderse a la puesta en disposición de los datos científicos, que no sólo pertenecen a las administraciones, sino que son generados por grupos de investigación atomizados. El potencial de su reutilización como herramienta de validación de la ciencia y como motor de innovación es incuestionable (Peset; Ferrer-Sapena; Subirats-Coll, 2011).

Los beneficios de compartir y reutilizar los datos han sido expuestos por numerosos autores, que resaltan principalmente un ahorro en los costes, una mayor rentabilidad de la inversión pública en los proyectos de investigación y un considerable aumento de citas, añadidas a las citas que reciben del artículo. El libre acceso y el intercambio de los datos refuerzan la investigación científica abierta, y fomentan la diversidad de análisis y de opinión y, por ende, la aparición de nuevas investigaciones. Explorar temas no previstos por los investigadores iniciales hace posible la comprobación de hipótesis alternativas y métodos de análisis, lo cual facilita la formación de nuevos investigadores, así como la creación de nuevos conjuntos de datos cuando recombina fuentes múltiples (Fienberg; Martin; Straf, 1985).

El éxito de un proyecto de investigación se mide en la actualidad no sólo por las publicaciones que produce, sino también por los datos que pone a disposición de la comunidad en general. Archivos pioneros como GenBank han demostrado el potencial de estos conjuntos de datos para la generación de nuevos descubrimientos, especialmente cuando se combinan datos de muchos laboratorios y se analizan de maneras no previstas por los investigadores originales (Guralnick; Constable; Wieczorek [et al.], 2009). En definitiva, verificación y reutilización, junto a su preservación, son factores que, en nuestra consideración, van a influir en el futuro de los esquemas de financiación de la ciencia.

Este tipo de datos científicos incluye gran variedad de tipologías dependiendo del área de conocimiento. Entre ellos se encuentran, por ejemplo, los conjuntos de datos primarios, los materiales fuentes, las representaciones digitales de materiales gráficos y pictóricos, las bases de datos de estructuras genéticas y cristalográficas, secuencias de proteínas, *microarrays*, las neuroimágenes, etc.

En el contexto de preservación de datos surgen nuevos fenómenos que regulan esta actividad, tal y como sucedió en el campo de las publicaciones científicas. Observamos cómo se instituyen los depósitos para los datos (bancos de datos y repositorios); cómo se desarrollan o adaptan las tecnologías y los paquetes existentes; cómo comienzan a ser mencionados expresamente en las normativas y regulaciones en el ámbito nacional e internacional; y, por último, cómo nace un estado de opinión colectiva con respecto a los derechos de acceso y reutilización tanto por parte de científicos productores y usuarios de estos datos, de las agencias que los producen o financian a través de programas de investigación, como de las editoriales que, en ocasiones, los difunden desde sus plataformas de edición junto a los artículos y cuyas políticas de reutilización no siempre son explicitadas.

También se ha debatido el papel que pueden jugar los profesionales de la información y las bibliotecas académicas en la distribución de estos materiales (Martínez-Uribe; Macdonald, 2008). Uno de los conceptos clave en el acceso a los datos de investigación es la preservación digital, *digital curation*, definido por el Digital Curation Centre como "las acciones necesarias para mantener, preservar y añadir valor a los datos digitales de investigación a lo largo de todo su ciclo de vida". El ciclo de vida de los datos (*data life cycle*) se refiere a todo el proceso por el cual los datos son creados, analizados y gestionados, es decir, la recogida de datos experimental y de observación, su limpieza e integración, el análisis, la publicación, y su conservación en un depósito. Este ciclo de vida de los datos es necesario para su intercambio. Y cuenta con una serie de etapas que se inicia con la preservación digital de los datos (*digital curation lifecycle*). En primer lugar, se

evalúan los datos, ya que su valor y tipología determinarán la necesidad de preservarlos a largo plazo. En este proceso de selección, hay que destacar la importancia de entender el contexto en el que se generan los datos porque cada disciplina y subdisciplina tienen sus propias características particulares, y diferentes niveles de complejidad en los datos (ARL, 2006). Las acciones que se realizan en conjunto han de asegurar una preservación duradera en el tiempo, su acceso y reutilización, en función de las políticas y requisitos legales definidos por cada institución; y, por último, su transformación, para crer nuevos objetos digitales a partir del original.

El objetivo de esta preservación es favorecer la compartición de los datos para reutilizarlos en nuevos trabajos de investigación, lo que se denomina *data sharing*. En el ámbito científico, compartir datos ha sido una práctica habitual entre las instituciones científicas desde hace tiempo. Tradicionalmente, se ha llevado a cabo por vías informales —los investigadores facilitan a sus colegas los datos en bruto—, pero existe una tendencia a que los datos estén liberados para ser usados por investigadores anónimos, siguiendo el modelo de acceso abierto a las publicaciones. Por esta razón, se han desarrollado vías formales para el depósito de estos materiales, como los bancos de datos y repositorios de acceso abierto, junto a otra infraestructura tecnológica que está ganando presencia: las sedes o plataformas editoriales (Torres-Salinas, 2010b).

En los últimos años las editoriales han promovido la recepción de estos datos y la mayoría de ellas contemplan en su política para el autor unas pautas para el material complementario. En áreas como la medicina o las ciencias naturales las editoriales especifican los repositorios públicos en los que se deben depositar los conjuntos de datos para que el artículo pueda publicarse.

Pero no todos los investigadores comparten sus datos por vías formales a pesar de sus ventajas. Según apunta Torres-Salinas (2010a), tal vez sea como consecuencia de la falta de repositorios específicos, de la desconfianza en su preservación o del temor a que no se reconozca su autoría. Los datos resultantes de la investigación constituyen un capital intelectual de gran valor, por lo que algunas instituciones de investigación perciben como ineficiente compartir el resultado de la actividad investigadora si no extraen el valor económico total de los datos que comparten (Hammond; Moritz; Agosti, 2008).

Entre las iniciativas para regular la gestión de los datos de investigación destaca el *Data Audit Framework* (DAF) del Digital Curation Centre, que proporciona a las organizaciones los medios para identificar, localizar, describir y evaluar cómo están gestionando sus propios datos de investigación para asegurar su futura reutilización. Otro proyecto también financiado por el Joint Information Systems Committee (JISC), *DISC-UK DataShare*, pretende elaborar un modelo para el depósito de datos en repositorios institucionales en el Reino Unido.

Por su parte, las instituciones gubernamentales, a través de normativas y políticas para el almacenamiento y preservación de datos, establecen que los trabajos subvencionados con fondos públicos sean depositados en repositorios de acceso abierto (Arzberger; Schroeder; Beaulieu, 2004). Sirvan como ejemplo los *Principles and guidelines for access to research data from public funding* de 2007 de la Organización para la Cooperación y el Desarrollo Económicos (OCDE). En España, la Ley 14/2011, de 2 de junio de 2011, de la ciencia, la tecnología y la innovación, de momento no hace referencia a los materiales complementarios, mientras que en Estados Unidos los National Institutes of Health (NIH) obligan, desde 2003, a depositar los datos generados por la investigación financiada en bancos específicos en función de las áreas y subáreas de conocimiento. Los NIH desarrollaron varios repositorios y bases de datos suplementarios específicos, entre los que se encuentran el GenBank, Gene, Genome, Protein Cluster y PubChem, que facilitan la labor de preservación de los datos generados a los investigadores.

Las normativas y orientaciones de los organismos de investigación que recomiendan preservar los datos en depósitos creados específicamente para su almacenamiento se erigen de facto en una guía para su correcta gestión y conservación.

En definitiva, el acceso formalizado a los datos procedentes de la investigación aumenta la eficiencia de la inversión, por lo que es vital conocer las fuentes de datos disponibles: bancos de datos, repositorios, sedes editoriales y bibliotecas de datos. Los investigadores, como consumidores de literatura científica, necesitan acceder también a los datos generados (Martínez-Uribe; Macdonald, 2008).

En los últimos años, se han creado una gran cantidad de repositorios de datos institucionales, pero se observa en la bibliografía que no existe una identificación sistemática de las fuentes de preservación de los datos de investigación. La descentralización de los depósitos de almacenamiento de datos en los repositorios de las propias instituciones y, por tanto, la ausencia de un sistema central de búsqueda, requiere un registro que permita la identificación de todos

estos depósitos.

Al no encontrar un registro mundial de depósitos de datos de investigación, tal y como existe para repositorios —ROAR y OpenDoar—, nuestro grupo de investigación planteó la creación de una herramienta que los aglutinara de forma clasificada por disciplinas. El objetivo de este proyecto es facilitar la identificación de las fuentes de almacenamiento de datos de investigación para permitir, como mínimo, a los profesionales de la información conocer de forma fácil y fiable dónde los investigadores deben depositar sus datos y si existen lagunas disciplinares.

2 Material y método

En primer lugar, se han realizado búsquedas de información generales en Internet con el fin de identificar un directorio similar al nuestro.

En segundo lugar, se han consultado varias fuentes para recopilar los depósitos y repositorios de datos que existen. Se han revisado estudios bibliográficos previos interrogando las bases de datos de la Web of Knowledge, Scopus, CSIC y LISA, combinando las palabras clave: *data sharing*, *reuse*, *data curation*, *research data* y *data repositories*. Estos trabajos citaban depósitos como DART (Treolar, 2006), ARROW (Payne; Treolar, 2006), DRYAD (Greenberg, 2009), Protein Data Bank y GenBank (Martínez-Uribe; Macdonald, 2009) y otros mencionaban conjuntos de ellos (Torres-Salinas; Robinson-García; Cabezas-Clavijo, 2012). Se han examinado las políticas de copyright de las editoriales científicas con respecto al material suplementario de los artículos, pues en ellas se citan en ocasiones los depósitos recomendados. También se ha realizado la consulta de los registros de repositorios de acceso abierto ROAR (*Registry of Open Access Repositories*) y OpenDoar (*Directory of Open Access Repositories*) e identificado los archivos digitales que contienen datos de investigación.

De este análisis se han obtenido numerosos repositorios de datos.

En tercer lugar, se ha utilizado Drupal para construir el web ODiSEA y la base de datos. Utilizar un gestor de contenidos ha permitido incorporar un sistema de búsqueda en los registros aplicando distintos tipos de filtros. El uso del módulo Google chart API facilita al usuario la visualización en tiempo real de distintos tipos de gráficos (barras, sectores, etc.).

El registro contiene campos de directorio y campos de análisis. Se obtuvo información sobre la institución responsable del depósito, el área geográfica, el tipo de datos y la cantidad que almacena, el formato, y el grado de cumplimiento del protocolo OAI-PMH.

La clasificación de los bancos de datos está basada en las áreas de conocimiento del Essential Science Indicators de la Web of Knowledge: Agricultural Science, Biology and Chemistry, Chemistry, Clinical Medicine, Computer Science, Economics and Bussines, Engineering, Environment Ecology, Geoscience, Immunology, Material Science, Mathematics, Microbiology, Molecular Biology and Genetics, Multidisciplinary, Neuroscience and Behaviour, Pharmacology and Toxicology, Physics, Plant and Animal Science, Psychiatry / Psychology, Social Science General, Space Science.

3 Resultados

El resultado de este trabajo es un inventario de depósitos especializados en la preservación de datos de investigación a nivel mundial llamado ODiSEA: International Registry on Research Data. Recoge los depósitos que conservan conjuntos de datos, material adicional a los artículos y materiales gráficos y multimedia.

Imagen 1. Interfaz de ODiSEA

Es un proyecto conjunto entre cinco universidades españolas: Universidad Politécnica de Valencia, Universidad de Valencia, Universitat de Barcelona, Universidad Católica de Valencia y Universidad

de Murcia.

Sus funcionalidades incluyen la búsqueda por disciplina y por tipos de datos que almacena el depósito.

Imagen 2. Ejemplo de interfaz de búsqueda

Permite conocer los datos de depósitos que responden a los criterios de búsqueda, de manera que ayuda a la identificación, por ejemplo, de las lagunas disciplinares o geográficas que actualmente existen.

Imagen 3. Lista de resultados

Actualmente cuenta con 176 depósitos, entre los que encontramos bancos especializados, bibliotecas de datos, repositorios que aceptan conjuntos de datos y bancos de imágenes.

Imagen 4. Distribución por materias

Recoge tanto depósitos específicos como GenBank, y genéricos como Dataverse Network, entre otros muchos. Permite conocer los datos individuales de cada depósito, si bien la base de datos recoge otros para el análisis.

Imagen 5. Interfaz de un registro completo

El análisis por países muestra que el mayor número de depósitos se encuentran en Estados Unidos y Reino Unido. La situación en España, es similar a Japón, Canadá y Australia con un número de depósitos superior a Francia, Italia, Dinamarca o Alemania.

Imagen 6. Distribución por área geográfica

La inclusión de un campo que indica si el depósito es compatible con el protocolo OAI-PMH, que por el momento se está investigando en cada uno de los depósitos, nos permite saber si estos conjuntos de datos van a ser cosechados por máquinas que ofrecen servicios de búsqueda masivos, como son OAister-OCLC, o Google. Por último, consideramos que uno de los indicadores más importantes de este registro es el que muestra si es de acceso abierto, es decir, si los datos son reutilizables o no. Con ello tendremos el verdadero panorama del retorno de la inversión en investigación.

4 Conclusiones

Para facilitar que prospere la investigación, es necesaria la conservación de los datos. Ello incluye su selección, la conservación en función de las políticas y requisitos legales de cada institución, y

su reutilización atendiendo a los derechos de propiedad intelectual y de patentes. Aunque cabe destacar que aún existe una cierta carencia en los marcos técnicos e institucionales para regular la normalización del acceso abierto a los datos de investigación, de forma gradual están emergiendo iniciativas internacionales que favorecen el poder compartirlos. Numerosos investigadores, agencias de investigación y centros superiores están interesados en hacer que los datos científicos sean reutilizables a través del desarrollo e implementación de repositorios digitales, que faciliten su gestión y el acceso.

La proliferación de estos depósitos específicos de preservación de los datos en distintas disciplinas ha conducido a que surja una nueva necesidad: la existencia de un registro global que recopile y clasifique estos depósitos. Para cubrir esta nueva demanda se ha creado ODiSEA, que registra y clasifica por disciplinas los depósitos existentes de datos de investigación.

El análisis de los datos que contiene muestra un predominio geográfico de Estados Unidos y Reino Unido y del tipo banco especializado y repositorio digital. Las disciplinas que más depósitos reúnen son biología molecular y genética, y biología y química, frente a farmacología y toxicología y ciencias de las plantas y los animales. Hay que destacar que el mayor número de depósitos se aglutinan bajo la categoría multidisciplinar.

En cuanto al acceso abierto y reutilización, los resultados se están investigando, pues muchos de los depósitos no lo definen específicamente en sus políticas.

La aportación que supone este registro facilita el descubrimiento de nuevos conjuntos de datos cuya recombinación desde una perspectiva multidisciplinar fomentará la innovación y la rentabilidad de la inversión en ciencia.

Bibliografía

ARL (Association of Research Libraries) (2006). *To stand the test of time: long-term stewardship of digital data sets in science and engineering*. Arlington (Va.): The Association. <http://www.arl.org/bm~doc/digdatarpt.pdf>. [Consulta: 28/08/2012].

Abella, Alberto (2011). *Reutilización de información pública y privada en España. Avance de situación para agentes públicos y privados. Una oportunidad para los negocios y el empleo*. Madrid: Rooter. <http://rooter.es/documents/PAPER_REUTILIZACION_INFORMACION_PUBLICA_PRIVADA_OPENDATA.pdf>. [Consulta: 25/08/2012].

Arzberger, P.; Schroeder, A.; Beaulieu, G. [et al.] (2004). "Promoting access to public research data for scientific economic and social development". *Data science journal*, vol. 3, no. 29 (Nov.), p.135–152. <https://www.jstage.jst.go.jp/article/dsj/3/0/3_0_135/_article>. [Consulta: 25/08/2012].

Data Information Specialist Comitee-UK (DISC-UK) (2007). *DataShare Project*. <http://www.disc-uk.org/datashare.html>. [Consulta: 25/08/2012].

Ferrer-Sapena, Antonia; Peset, Fernanda; Aleixandre-Benavent, Rafael (2011)."Acceso a los datos públicos y su reutilización: *open datay open government*". *El profesional de la información*, vol. 20, n.° 3 (mayo–junio), p. 260–269. <http://elprofesionaldelainformacion.metapress.com/link.asp?id=92741636q145x727>. [Consulta: 25/08/2012].

Fienberg, Stephen E.; Martin, Margaret E.; Straf, Miron L. (ed.) (1985). *Sharing research data*. Washington, D.C.: National Academy Press.

Greenberg, Jane (2009). "Theoretical considerations of lifecycle modeling: an analysis of the dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption". *Cataloging & classification quarterly*, vol. 47, no. 3, p. 380–402.

Guralnick, Robert; Constable, Heather; Wieczorek, John [et al.] (2009)."Data's shameful neglect". *Nature*, vol. 461, no. 145 (Sept.). <http://www.nature.com/nature/journal/v461/n7261/full/461145a.html>. [Consulta: 25/08/2012].

Hammond, Tom; Moritz, Thomas D.; Agosti, Donat (2008)."The conservation knowledge commons: putting biodiversity data and information to work for conservation". En: *Proceedings of the Twelfth Biennal Conference of the International Association for the Study of Commons*.

<http://dlc.dlib.indiana.edu/dlc/bitstream/handle/10535/2132/Moritz_129701.pdf?sequence=1>. [Consulta: 25/08/2012].

Marcos-Martín, Carlos; Soriano-Maldonado, Salvador-Luis (2011). "Reutilización de la información del sector público y *Open data* en el contexto español y europeo. *Proyecto Aporta*". *El profesional de la información*, vol. 20, n.° 3 (mayo–junio), p. 291–297. <http://es.scribd.com/doc/57214418/Reutilizacion-de-la-informacion-del-sector-publico-y-open-data-en-el-contexto-espanol-y-europeo>. [Consulta: 12/07/2012].

Martinez-Uribe, Luis; Macdonald, Stuart (2009)."User engagement in research data curation". *Lecture notes in computer science*, vol. 5.714, p. 309–314. <http://www.era.lib.ed.ac.uk/bitstream/1842/3206/1/Martinez_Macdonald_ECDL09.pdf>. [Consulta: 02/11/2012].

— (2008). "Un nuevo cometido para los bibliotecarios académicos: *data curation*". *El profesional de la información*, vol. 17, n.° 3 (mayo–junio), p. 273–280. <http://www.elprofesionaldelainformacion.com/contenidos/2008/mayo/03.pdf>.[Consulta: 15/07/2012].

Payne, Geoff; Treloar, Andrew (2006)."The ARROW Project after 2 years: are we hitting our targets?". En: *Proceedings of VALA*. Melbourne. <http://www.valaconf.org.au/vala2006/papers2006/57_Treloar_Final.pdf>. [Consulta: 28/08/2012].

Peset, Fernanda; Ferrer-Sapena, Antonia; Subirats-Coll, Imma (2012). "*Open data* y *linked open data*: su impacto en el área de bibliotecas y documentación". *El profesional de la información*, vol. 20, n.° 2 (marzo–abril), p. 165–173. <http://www.elprofesionaldelainformacion.com/contenidos/2011/marzo/06.pdf>. [Consulta: 02/08/2012].

Torres-Salinas, Daniel (2010a). "Compartir datos (*data sharing*) en ciencia: contexto de una oportunidad". *Anuario ThinkEPI*, vol. 4, p. 258–261. <http://www.thinkepi.net/compartir-datos-data-sharing-en-ciencia-el-contexto-de-una-oportunidad>. [Consulta: 24/07/2012].

— (2010b). "Hacia la gestión de datos de investigación en las universidades: la *Data asset framework*". *Anuario ThinkEPI*, vol. 4, p. 262–265.

Torres-Salinas, Daniel; Robinson-García, Nicolás; Cabezas-Clavijo, Álvaro (2012). "Compartir los datos de investigación: introducción al *data sharing*". *El profesional de la información*, vol. 21, n.° 2 (marzo–abril), p. 173–184. <http://hdl.handle.net/10760/16786>. [Consulta: 24/07/2012].

Treloar, Andrew (2006). "The Dataset Acquisition, Accessibility, and Annotatione-Research Technologies (DART) Project: building the new collaborative e-research infrastructure". En: *Proceedings of AusWeb06, the Twelfth Australian World Wide WebConference*. Southern Cross University Press. <http://ausweb.scu.edu.au/aw06/papers/refereed/treloar/paper.html>. [Consulta: 08/07/2012].

Treloar, Andrew; Groenewegen, David; Harboe-Ree, Cathrine (2007)."The data curation continuum: managing data objects in institutional repositories". *D-Lib magazine*, vol. 13, no. 9–10. <http://www.dlib.org/dlib/september07/treloar/09treloar.html>. [Consulta: 08/07/2012].

Fecha de recepción: 05/09/2012. Fecha de aceptación: 28/10/2012.

Articles del mateix autor a Temària

García Palomeque, Rebeca Pérez Campos, Rafael

[més informació]