Biblios
ISSN 1562-4730 (online)

Infraestructuras y políticas internacionales de desarrollo para gestión de los datos de investigación

Fabiano Couto Corrêa da Silva
Universidad de Barcelona - UB, España

REVIEW

Resumen

Existen enormes limitaciones tecnológicas para atender adecuadamente la tarea de facilitar el acceso a los datos producto de la investigación científica. Han surgido como una respuesta de las instituciones, en especial las académicas, repositorios que sirven para preservar y poner a disposición de su comunidad académica, los datos de investigación que hacen parte de su patrimonio intelectual. Por lo tanto, es necesaria la adopción de formatos de intercambio de datos y la normaización de los mecanismos de almacenamiento, para poner a disposición estos datos científicos a nivel local y a nivel internacional. Nuestra investigación constituye un análisis de infraestructuras y políticas de desarrollo para la gestión de los datos de investigación considerando los mecanismos adoptados a nível internacional.

Palabras clave
Datos de investigación ; Datos científicos ; Información Científica

Infrastructure and international development policies for data management research

Abstract

There are big technological limitations to attend properly the task to facilitate the access to the data product of scientific research. They have emerged as a response of institutions, especially from universities, repositories that serve to preserve and make available to the academic community, research data that are part of their intellectual heritage. Therefore, is necessary the adoption of data exchange formats and standarization storage mechanisms to make available this scientific data locally and internationally. Our study is an analysis of infrastructure and development policies for managing research data considering the mechanisms adopted internationally.

Keywords
Research data ; Scientific data ; Scientific Information

1 Introducción

La explosión de información que tras la de la Segunda Guerra Mundial provocó, sobre todo en los países desarrollados, un gran interés en las actividades de ciencia y tecnología, conllevó a un aumento considerable de las nuevas técnicas de gestión documental. Este fenómeno se caracteriza por un crecimiento exponencial de los conocimientos registrados aunque trajo con el un problema de fondo: la tarea de hacer más accesible la creciente documentación.

Actualmente estamos presenciando una segunda gran transformación en el desarrollo en ciencia y tecnologia, por causa de la impactante cantidad de datos producidos en la actividad de investigación científica, especialmente entre los formatos digitales. El progreso incurrido hasta día de hoy en el traspaso de datos ha llevado a alcanzar una velocidad compatible al del volumen que en nuestro presente se produce y se consume, convirtiéndose necesario que transite de forma automática y con la menor interferencia posible de quienes especifícalo. Si anteriormente los investigadores producían conocimiento buscando documentos en distintos repositorios, hoy son los datos de investigación los que se requieren para avanzar en la producción científica (CONICYT, IDER (2010). Por consiguiente, es así cómo tuvo lugar una tendencia en todo el mundo homogénea

en cuanto a la cuestión del almacenamiento de resultados científicos. Asimismo, cabe añadir la latente demanda por parte de los investigadores hacía el acceso a esos datos pues son la base del análisis de los datos resultantes.

Todo ello también viene motivado principalmente debido a que la tecnología digital se ha convertido en un elemento cada vez más omnipresente en los procesos de construcción del conocimiento científico, ya sea mediante el aumento de la capacidad de los instrumentos científicos, mediante la reconstrucción de la realidad a través de la simulación, o mediante la apertura de nuevas formas de colaboración para compartir datos de investigación. Además, el avance de las herramientas de preservación y el procesamiento de datos, en constante evolución, han abierto nuevas aplicaciones para las fuentes básicas de la investigación -los datos de investigación- dando un nuevo impulso a las investigaciones científicas. Ya un informe de la Organización para la Cooperación y el Desarrollo Económico (OCDE) datado en el año 2007 hacía hincapié en esta condición destacando algunas situaciones en las que los datos de investigación se convierten en un factor esencial: la cadena de la innovación, la cooperación internacional, la promoción de nuevas investigaciones y pruebas de nuevas hipótesis o alternativas en estudios de diversidad y opiniones; la formación de nuevos investigadores en la exploración de temas no previstos originalmente, para generar nuevos conjuntos de datos de múltiples fuentes y, sobre todo, en la promoción de actividades científicas más abiertas y transparentes, que tienen como principio producir conocimiento a disposición del público.

El vicepresidente de *Oracle Europa*, Malhar Kamdem, asegura que los datos tendrán este siglo "el mismo poder que la electricidad en el siglo XX"1. La excomisaria europea Neelie Kroes de la Agenda Digital ya advertía hace años que los datos crudos son "el nuevo petróleo". Pero para refinarlos, advierte Kamdem, se necesitan profesionales especializados. Es justamente la capacidad de estructurar los datos adecuadamente lo que el progreso científico pretende hacer factible en consecuencia de la diversidad del ecosistema de tecnologías y métodos actuales de gestión de los datos de investigación. Pero lo sorprendente no es sólo la cantidad de datos, sino también lo que podemos hacer con ellos actualmente, pues los nuevos avances en la minería y visualización de datos nos dan formas de extraer información útil a partir de conjuntos de datos cada vez más grandes.

Como dijo Tim Berners-Lee2, director del Consorcio de la *World Wide Web*: "Los datos son preciosos y van a durar más que los propios sistemas." Él tiene razón. Aunque, cabe recordar que los datos son tan útiles como su plan de almacenamiento, los datos que se encuentran dispersos en PCs, notebooks, tablets, móviles u otros dispositivos personales suelen ser difíciles de manejar y susceptibles a perderlos. Sin embargo, si se accede on-line, las posibilidades de comprender algo más a fondo puede presentar resultados de manera más amplia y completa, representando un enorme potencial en el avance científico. En concreto, facilita la recopilación de los resultados de las investigaciones científicas y permite la aplicación de los datos antiguos en nuevos contextos. Siendo así, no es de extrañar que el intercambio de los datos de investigación encuentre un amplio apoyo entre los actores académicos. La Comisión Europea, por ejemplo, proclama que el acceso a los datos de investigación aumentará la capacidad de innovación de Europa (European Commission, 2012). Al mismo tiempo, asociaciones nacionales de investigación están uniéndose para promover el intercambio de datos en el mundo académico. Dentro de este ámbito, por un lado, el Knowledge Exchange Group (2015), en un esfuerzo conjunto de las cinco principales agencias de financiación europea, es un buen ejemplo de esfuerzo transfronterizo para fomentar una cultura del intercambio y la colaboración. Por otro lado, revistas como *Nature*, *PLoS One* o *Atmospheric Chemistry and Physics* adoptan cada vez más las políticas de intercambio de datos con el objetivo de promover el acceso público a los datos.

Limitaciones del estudio y metodología

Aunque existan muchas demandas relacionadas con la gestión de los materiales digitales, en este apartado daremos énfasis a las necesidades de acceso, gestión y conservación de los datos por parte de la comunidad científica. Desde esta perspectiva, analizaremos el proceso de registro de los datos de investigación y el papel de sus integrantes desde su recogida hasta la publicación.

La reflexión que ofrecemos tiene como objetivo investigar la relación, a nivel conceptual y práctico, entre el proceso de investigación y el tratamiento de los datos, teniendo en cuenta el estado de la cuestión sobre las infraestructuras disponibles y las posibilidades de uso de los recursos en las diferentes áreas del conocimiento. Aunque más específicamente se refiere a una construcción teórica sobre la importancia de los datos de

investigación: una revisión de la literatura sobre la infraestructura global de gestión de los datos de investigación. El análisis de la literatura fue realizado en informes de las políticas y las prácticas vigentes en los principales proyectos de investigación. Para localizar los principales informes y políticas de preservación digital y el uso de los datos de investigación, investigamos las referencias apuntadas por el Digital Curation Centre, que es es un centro líder en la curaduría de información digital, que hace análisis del escenario global, y el Programa Horizon 2020, iniciativa destinada a asegurar la competitividad global de Europa. Además, hicimos un rescate histórico de las políticas de acceso a los datos de investigación en nivel internacional.

2 El valor de los datos

Para definir los datos de manera sucinta podemos decir que el conocimiento es el motor del avance científico. Y los datos es su combustible. Para los investigadores, una gestión de datos de investigación adecuada permite nuevas formas de comparación y descubrimientos, es decir, permite nuevos campos enteros de investigación. Estamos en el comienzo de un cambio de paradigma. Y empezamos a ser capaces de almacenar, procesar y analizar estas enormes cantidades de datos. Esto puede cambiar la manera en que tomamos decisiones y ejecutamos nuevas investigaciones.

Nuevas investigaciones generan datos que necesitan ser organizados y entendidos, por lo que requieren de un proceso continuo que busque identificar las dimensiones, categorías, tendencias, patrones y relaciones, revelando su significado. Este proceso es complejo e implica un trabajo de reducción, organización e interpretación de los datos que se inicia previamente en la fase exploratoria y que continua durante todo el ciclo de investigación.

Teniendo en cuenta que los datos son la esencia que hacen que la información se encuentre un constante movimiento, es difícil de fijarlos en una forma estática permanente. Las fronteras entre el conocimiento tácito y un terreno común también están cambiando, ya que múltiples áreas del conocimiento negocian cómo se entienden los datos en todas las disciplinas y los dominios científicos. Los avances del conocimiento científico son aún más difíciles de establecer en una época que grandes conjuntos de datos están disponibles en todas las áreas. Es difícil replicar el intercambio de información entre colaboradores desconocidos, sobre todo entre distintas comunidades científicas y durante largos períodos de tiempo. Algunas transferencias pueden ser mediadas por la tecnología, pero muchas dependerán de la experiencia de los mediadores, ya sean investigadores, bibliotecarios u otros actores involucrados en la gestión de los datos de investigación.

La gestión de los datos de investigación tiene sus aplicaciones presentadas ante la perspectiva del desarrollo de los procesos de preservación, uso y reutilización de los datos. El dominio de estos aspectos es fundamental para que los investigadores planifiquen su trabajo desde la concepción de su proyecto hasta la ejecución, uso y archivamiento de los datos.

3 Adiós al viejo mundo

Las posibilidades de reutilización de los datos se amplia cada vez más. De acuerdo con Falcone (2011), el volumen, se refiere a que en la actualidad nos encontramos escalando desde terabytes a zettabytes; administrar la complejidad de múltiples fuentes de información estructurada y no estructurada; gestionar datos y eventos en tiempo real junto a volúmenes masivos de información almacenada. El autor va más allá y apunta que en los próximos años la cantidad de información se multiplicará por 44. Eso significa que alcanzaremos a 35 zettabytes en 2020. Su análisis estima cifras donde el alrededor del 90% de los datos existentes en el mundo estarían siendo creados cada 2 años; el 80% de los datos del mundo en la actualidad no está estructurado y tan sólo el restante 20% estaría almacenado en bases de datos que posibilitan analizar la información de forma estructurada.

Contra la necesidad de estructurar el creciente volumen de datos, en el proceso de gestión de datos de investigación la confianza, fiabilidad y la facilidad de uso de los datos son fundamentales, pero en general aún es un reto implementar los requisitos de gestión necesarios. La adopción colectiva por la comunidad científica podría provenir del "Horizon 2020", el nuevo programa del marco europeo de investigación e innovación para el período 2014-2020. Según la Comisión Europea, "un plan de gestión de datos es un documento que describe cómo los datos de investigación recopilados o generados serán tratados durante un proyecto de investigación y después de que se haya completado, describe cuales serán recogidos siguiendo metodología y estándares específicos, y cómo estos datos serán compartidos, si serán abiertos y cómo ocurrirá la curaduría digital". El establecimiento de los Planes de Gestión de Datos aún es una realidad distante para muchos países.

Expresiones como "Datos Abiertos" o "*Big Data*" son muy populares, pero muy pocas instituciones tienen una política de gestión de datos.

Es evidente que muy pocos investigadores se sienten preocupados por los registros de los datos de investigación. Por lo general, los mantienen hasta que ya no los necesitan, pero gracias a los planes de gestión de datos de investigación podrían asegurarse de que sus objetivos sean compatibles con la preservación y el acceso, mientras que con los datos de investigación son una producción cada vez más importante pero costosa en el proceso de la investigación académica, en todas las disciplinas. Los datos de investigación son una parte esencial de las pruebas necesarias para evaluar los resultados de investigación y para reconstruir los hechos y procesos que conducen a ellos. Su valor aumenta a medida que se agregan en colecciones y están más disponibles para su reutilización en nuevas cuestiones. Pero solo vamos a entender el valor de datos si nos movemos más allá de las políticas de investigación, prácticas y sistemas de apoyo desarrollados en un momento anterior diferente. Necesitamos nuevos enfoques para la gestión y el acceso a los datos de investigación puesto que al intentar desarrollar infraestructuras para archivar y disponer los datos de investigación todas las partes deben trabajar en colaboración y asegurar que es sensible a las necesidades de los investigadores y a los diferentes contextos en los que trabajan. También se deben tener en cuenta los avances técnicos y de formulación surgidas de políticas procedentes del ámbito nacional y extranjero.

4 El papel de los investigadores

Hay un par de maneras con el cual los investigadores pueden apoyar este proceso. La primera es creando una organización adecuada con respecto al ciclo de vida de los datos. De hecho, la planificación desde el momento de la captura hasta el almacenamiento es la única manera de garantizar la continuidad de nuevas investigaciones. La segunda manera es el depósito en los repositorios y revistas que presentan infraestructura para indexar y recuperar los conjuntos de datos.

Hay muchas áreas de actuación donde los datos de investigación son importantes, así como existen varios grupos de individuos y organizaciones que pueden beneficiarse de esta disponibilidad, incluido el propio Gobierno. A su vez, es imposible saber exactamente cómo y dónde los datos abiertos serán mejor valorados, ya que al permitir que haya innovación en diversas áreas, pueden surgir nuevas maneras de utilización.

Además del Gobierno, esta nueva configuración del desarrollo científico global se encuentra cada vez más insertada entre los análisis de agencias científicas y las personas involucradas con el desarrollo científico de manera general. En este sentido, un reporte de gran repercusión publicado en 2010 bajo el título "*A Surfboard for Riding the Wave: how Europe can gain from the rising tide of scientific data*" nos permite conocer las valoraciones de un grupo de expertos que fue convocado por la Comisión Europea para evaluar los posibles beneficios de la puesta en marcha de una e-infraestructura global de datos. Ello es muy importante porque permitiría a investigadores procedentes de diferentes áreas de conocimiento y centros compartir datos y reutilizarlos. Este informe parte de la constatación de que el avance de nuevas infraestructuras para la gestión de los datos de investigación puede acelerar descubrimientos y cambiar la manera de realizar las investigaciones dentro de un escenario completamente nuevo. La razón de esto es evidente: la revolución digital ha hecho que sea mucho más fácil almacenar, compartir y reutilizar datos. Los datos de investigación de todas las disciplinas están ahora casi universalmente en formato digital. El amplio acceso e intercambio de datos aumenta el retorno de las grandes inversiones que se realizan en la investigación y tiene el potencial para avanzar de manera exponencial el conocimiento. Por ejemplo, en el análisis de la genética, la tecnología para capturar enormes cantidades de datos y la conexión entre ellos posibilita generar nuevas informaciones que están mostrando la composición y la comprensión de numerosas enfermedades y síndromes. Pronto seremos capaces de analizar las complejidades tales como la predisposición a la enfermedad en las poblaciones de animales y plantas teniendo en cuenta su base genética, las condiciones ambientales y las condiciones demográficas, por lo que todos estos factores pueden convertirse en parte de las estrategias de prevención en la fabricación de medicamentos. Con la capacidad de acceso e integración de los datos recopilados (los datos no son personas, no tienen capacidad, faltaría un sujeto o mayor concreción) en los diferentes campos, nuevos regímenes de conocimiento se están abriendo en formas que han sido históricamente imposibles.

A diferencia de las formas tradicionales de archivo, con los datos de investigación no se trata el mantenimiento de registros para temas de efectos legales, históricos o culturales sino que se intenta satisfacer las necesidades de los investigadores que operan en el entorno digital de hoy en día. La misión principal de un archivo de datos de investigación no es solamente conservar la memoria grabada de un grupo, organización o nación, sino que también lo es el proporcionar un servicio de vital interés para la comunidad investigadora.

Para tener una utilidad aceptada por la comunidad científica, los datos necesitan de una estructura y organización jerárquica, ofreciendo colecciones informativas relacionadas y registradas en un formato adecuado según el tema del cual se trate, es decir, en el contexto de la comunicación científica deseada. De esta manera, de los resultados generados se obtiene un conjunto de datos que puede almacenarse (correctamente) y ser reutilizado al distribuirse a otros investigadores, incluso puede ampliarse a áreas distantes a priori a los objetivos iniciales de la investigación.

5 Las infraestruturas de gestión de datos

Las agencias de fomento para la planificación de los datos de investigación ofrecen una amplia gama de políticas de gestión a sus respectivas comunidades de investigación, incluyendo el acceso a catálogos y bases de datos a través de Internet, protocolos de retención, la creación de metadatos, la migración de datos a través de software y sistemas de hardware, y, la formación y el desarrollo de las normas internacionales. Al ofrecer estos servicios las agencias desempeñan un papel activo y estratégico en la formulación de nuevos métodos y técnicas dentro de los intercambios de los datos de investigación y el desarrollo de nuevos estándares en todos los aspectos de la conservación de datos.

Los estudios, informes y declaraciones que destacan la importancia de la gestión de los datos de investigación son relativamente recientes. Sin embargo, son numerosas y variadas las referencias destacadas. A continuación, procederemos a comentarlas en orden cronológico. Si tratan de iniciativas que hacen eco de políticas anteriores en esta área que fueron desarrolladas para el establecimiento de un consenso sobre las mejores prácticas para el acceso a los datos de investigación.

En 2004 se reunieron en París los Ministros de la Ciencia y Tecnología de los países integrantes de la OCDE, conjuntamente con China, África del Sur, Israel y Rusia, para discutir sobre las directrices internacionales respecto al acceso de los datos de investigación. Ellos realizaron un análisis sobre las múltiples posibilidades de acceso libre a los datos para la promoción del desarrollo científico -principalmente para obtener un retorno del financiamiento público invertido en investigaciones-. En consecuencia, se aprobó la *Declaration on Access to Research Data from Public Funding* (OCDE, 2004).

También ese mismo año se creó la *Open Knowledge Foundation* (OKF) para promover el acceso a los contenidos y datos abiertos. Esta fundación inició proyectos como el *Comprehensive Knowledge Archive Network* y la iniciativa *Open Data Commons*, así como soluciones jurídicas para la apertura y reutilización de datos de investigación. En 2008 se inauguró la *Public Domain Dedication and License* (PDDL), una licencia pensada para el uso de bases de datos (Open Data Commons, 2008). Con la supervisión de la OKF, el Manual de Periodismo de Datos3 nació en un taller de 48 horas encabezado por la *European Journalism Centre* y la *Open Knowledge Foundation* en la MozFest4 celebrado en Londres el año 2011. Posteriormente, la MozFest se amplió convirtiéndose en un proyecto de difusión internacional que actualmente cuenta con la participación de los principales representantes del periodismo de datos.

La OCDE quiso dar continuidad a la declaración de 2004 y mediante el *Committee for Scientific and Technological Policy* (OECD's) designó un equipo de expertos con el encargo de proponer un conjunto de normas para la promoción y el desarrollo de los datos de investigación procedentes de la financiación pública. Para obtener dicho consenso entre la investigación y los representantes, ellos contactaron con instituciones de investigación y representantes políticos de los países miembros de la OCDE. Como resultado de ello, en 2007 se presentó el documento Principles and guidelines for access to research data from public funding (OECD, 2007) unas directrices que facilitan el acceso a los datos de investigación generados con financiación pública. Los principios y las directrices contenidos en este documento pretenden orientar a los gobiernos, las organizaciones de financiación, las instituciones de investigación y a los propios investigadores. Sobre todo a estos últimos, quieren servir de ayuda en el trato con los obstáculos y desafíos surgidos a raíz del intercambio internacional de los datos de investigación, como por ejemplo los siguientes: los problemas tecnológicos, de gestión institucional o financiera, las políticas legales y, las cuestiones relacionadas con la financiación, producción, administración y uso de los datos.

En 2007 el *Research Information Network* (RIN, 2007) publicó un informe de las políticas y las prácticas vigentes en los principales proyectos de investigación del Reino Unido. Aunque la política del momento del estudio se había enfocado principalmente a la difusión mediante artículos de revistas y actas de congresos, se reconoció que algunos consejos de investigación tenían una muy buena infraestructura con la política asociada a los datos de curaduría. Este estudio proporciona un panorama comparativo sobre cómo los diferentes grupos de

financiación esperan que los investigadores que apoyan se ocupen de gestionar y facilitar el acceso a los datos de sus investigaciones. Se examinaran las políticas y la práctica de una selección alrededor de 25 de los mayores financiadores de investigación, entre los cuales se incluyeran: los siete consejos de investigación del Reino Unido, siete universidades, una selección de departamentos del Gobierno, de organizaciones benéficas de investigación y de empresas e industrias que invierten significativamente en I+D.

También en 2007, la Comisión Europea publicó la comunicación *Scientific Information in the Digital Age* (European Comision, 2007), en la que examinaba cómo las nuevas tecnologías digitales pueden ser utilizadas para aumentar el acceso a las publicaciones de investigación y también cómo los datos son un importante motor de la innovación. Así se propone poner en marcha un marco a nivel de la Unión Europea para apoyar nuevas formas de promover un mejor acceso a la información científica en línea, preservando los datos de investigación. En cuanto a medidas concretas, la Comisión apoya la difusión en acceso abierto de los proyectos de investigación (mediante, por ejemplo, el reembolso de los costes de publicación) y ha destinado recursos para el desarrollo de infraestructuras en el almacenamiento de los datos de investigación y la investigación en preservación digital en Europa.

Dos años más tarde, la Comisión Europea encargó a un grupo de expertos un informe sobre su visión al acceso, uso, re-uso y calidad de los datos de investigación, con vistas al año 2030. El objetivo principal era conocer los beneficios y los costes de la puesta en marcha de una infraestructura global de datos fiable y estable, que permitiera a los investigadores y a otras partes interesadas de la educación, la sociedad y los negocios la utilización, reutilización y explotación de los datos de investigación de cara al máximo beneficio de la ciencia y la sociedad. El informe titulado *A Surfboard for Riding the Wave* no fue publicado hasta 2011, y está siendo utilizado como referencia europea en la construcción de una infraestructura que maximice los beneficios del acceso a la información científica.

De forma similar, el informe llamado Report on integration of data and publications (Reilly, 2011), procedente del proyecto ODE (Opportunities Data Exchange) sobre la integración de datos primarios y publicaciones demuestra la importancia de esta cuestión. Según dicho informe, se estima con un 58% el número de investigadores que desearían utilizar datos primarios ajenos, mientras que aproximadamente un 25% indica problemas para compartir los suyos. El informe también menciona repositorios y editoriales como las vías de almacenamiento preferidas por los investigadores y pone de manifiesto el deseo de reutilizar datos ajenos aunque con cierta reticencia a compartir los propios, aduciendo problemas legales. La ruta verde sería el depósito de datasets en repositorios o bancos específicos por disciplinas; y la via dorada, por su parte, consistiría en almacenarlos en plataformas editoriales junto a la publicación. Se apunta que relacionar los datos con las publicaciones puede aportar dos ventajas añadidas: contextualizar e interpretar los datos y, además, proporcionar valor tanto a los investigadores que los comparten como a las propias publicaciones. Cabe mencionar que en cuanto al papel de las editoriales, la validación y la preservación son los mayores problemas que se detectan al estar bajo la responsabilidad de estas.

En 2011, el *Knowledge Exchange*, una asociación con miembros de instituciones dedicadas a la creación de e-infraestructuras para la investigación y la enseñanza superior de cuatro países europeos, presentó una visión general de la situación actual respecto a los datos de investigación en Dinamarca, Alemania, los Países Bajos y el Reino Unido, proponiendo líneas generales para el desarrollo de una infraestructura de datos por medio de un programa de acción conjunto entre ellos.

Otro informe llamado *Riding the wave* (Van der Graaf; Waaijers, 2011), reuniendo la experiencia de expertos en el desarrollo de infraestructura de datos de investigación, ha hecho recomendaciones para el gobierno e iniciativas del ámbito privado con el fin de conseguir explorar los datos generando al mismo tiempo que generase beneficios para la sociedad y la ciencia. Señala que para que un plan tan ambicioso obtenga éxito se necesita la participación de todos los interesados de la comunidad científica fomentando así la capacidad iniciativa de los investigadores en su papel de productores y usuarios de las infraestructuras de información de datos. El informe también recomienda que las competencias básicas en materia de manejo de datos deben convertirse en una competencia académica básica. Algunos de los países como Dinamarca, Alemania, Países Bajos y el Reino Unido que están desarrollando la infraestructura de datos colaborativa elaboraron su proprio esquema de programa de acción y recordaron que este necesita la participación de todos los interesados en la comunidad científica. Asimismo son identificados cuatro factores clave para el éxito de esta infraestructura: (1) los incentivos, (2) la formación de investigadores, tanto en su papel de productores como de usuarios de las infraestructuras de información de datos, (3) la infraestructura técnica y de organización, y (4) el financiamiento de infraestructura para los nuevos desarrollos en la logística de datos. Si ejercemos una visión general de la

situación, el informe ofrece tres objetivos estratégicos a realizar a largo plazo: 1. El intercambio de datos será parte de la cultura académica; 2. Los datos logísticos serán un componente integral de la vida profesional académica; 3. La infraestructura fijará el ritmo, tanto a nivel operativo como financiero.

Continuando con el informe, este señala que para la comunidad científica, todavía ligada a los viejos tópicos de difundir la ciencia tradicionalmente en libros o artículos, la sistematización de los datos está muy lejos de materializarse. Sin embargo, entre la comunidad no-tradicional que por lo contrario desarrolla innovadores diseños destaca que es posible avanzar en lo que se denomina datos de investigación globales.

En este sentido, el informe *Science as an open enterprise* de la Royal Society (2012) aporta un importante estudio sobre el uso de la información científica, ya que afecta tanto a los científicos como a la sociedad. Este no sólo identifica las oportunidades o desafíos de compartir y divulgar la información científica, sino también nos cuestiona en la manera de conseguir exprimir el máximo potencial de los datos de investigación, todo ello con la finalidad de apoyar una investigación innovadora y productiva beneficiosa con la sociedad. Además, el informe reconoce las ventajas fundamentales del acceso abierto a los datos de investigación y toma en consideración las condiciones específicas en que la "apertura" es la más beneficiosa a la comunidad de investigación, la política y la sociedad en general.

En España se presentó en 2012 el informe “Depósito y Gestión de datos en Acceso Abierto” elaborado por el grupo de trabajo del repositorio Recolecta, el cual adjunta consideraciones notables de tenerse en cuenta. Algunas de ellas hacen mención al diseño e implementación de la política de gestión de datos de investigación, otorgando especial atención a la situación del país con respecto a otros. Pero también son identificadas la variedad de tipos de datos de investigación y los actores implicados en su gestión (los repositorios institucionales y temáticos, las agencias de financiación, los centros de datos existentes, los investigadores, los bibliotecarios y los expertos en la gestión de datos, etc.). Asimismo, se reflexiona sobre los aspectos económicos derivados de la creación de una infraestructura interoperable de gestión de datos.

El informe europeo *Towards better access to scientific information: Boosting the benefits of public investment in research* (European Commission, 2012) constata el hecho de que hasta ahora los resultados de investigación científica se han difundido sobre todo mediante artículos. Asimismo advierte de que no hay una práctica bien establecida para la publicación de los datos subyacentes. De ello se concluye que los investigadores son, a menudo, reticentes a compartir sus datos con la falsa creencia de que otros puedan beneficiarse injustamente de su trabajo. El proceso de preparación de datos para compartir también se percibe como mano de obra intensiva y con la ausencia de mecanismos de metadatos. A su vez, la ausencia de un claro incentivo para el intercambio de datos es otro obstáculo importante. El informe se ocupa de esta cuestión en la recomendación del año 2012, denominándolo de ajuste de la contratación y evaluación del sistema de carrera para los investigadores y, de sistema de evaluación para la concesión de becas de investigación en los que participen en la cultura de compartir sus resultados como recompensa.

Los resultados presentados en el informe *Report on integration of data and publications* (Reilly, 2011) indican que la citación de datos es un sector de oportunidad tanto para los investigadores como para las bibliotecas y sirvió como base para el informe intitulado *Report on best practices for citability of data and on evolving roles in scholarly communication* (2012). El resultado de la nueva investigación de la OCDE expone el pensamiento actual sobre las mejores prácticas de citación de datos y presenta los resultados de una encuesta a los bibliotecarios preguntando cómo podrían y deberían desarrollarse nuevas funciones de apoyo. En ese informe también se hace referencia a la Conferencia LIBER 2011 de Barcelona. Su taller, basado en las conclusiones preliminares sobre la integración de datos y publicaciones, reveló que, a pesar de las bibliotecas observar un emergente y oportuno paisaje de los datos de investigación, había una verdadera necesidad de definir las orientaciones futuras y el alcance de la función de las bibliotecas en el intercambio de datos. El tema de la citación de datos también fue abordado como una cuestión fundamental que debe tratarse. Contodo, este trabajo es descrito aquí con mayor información obtenida a través de una amplia investigación documental, entrevistas estructuradas y una encuesta en línea de los miembros LIBER para analizar las más adecuadas prácticas en la cita de los datos y la evolución de las funciones de apoyo a las bibliotecas. Dicha encuesta fue diseñada para reunir pruebas sobre las funciones actuales y deseadas de las bibliotecas en lo que respecta a la gestión de datos con el fin de prescribir medidas para la evolución de estos roles.

En 2013, se publicó el informe *European Landscape Study of Research Data Management* que integra los documentos generados por el Proyecto SIM4RDM, que es resultado de una investigación desarrollada por seis socios Europeos: JISC (Reino Unido), HEA (Irlanda), NIIF (Hungría), NordForsk (Noruega), CSC (Finlandia) y

SURF (Países Bajos). El estudio presenta los resultados obtenidos de una encuesta para conocer cuáles son las actuaciones de los organismos de financiación, instituciones de investigación, organismos nacionales y editores de los estados miembros de la Unión Europea y de otros países para mejorar la capacidad y las habilidades de los investigadores en el uso efectivo de las infraestructuras de los datos de investigación. También, incluye recomendaciones a las organizaciones para ayudar a sus investigadores con el análisis de los programas paneuropeos, las donaciones internacionales existentes y las políticas de las instituciones. Y el informe también examina si tales programas o políticas incluyen las intervenciones de apoyo a los investigadores que incentiva la obtención de los conocimientos, las habilidades y el apoyo necesarios para la gestión de datos.

Así pues con este breve repaso a los estudios e informes encontrados sobre la gestión de los datos de investigación hemos dejado constancia de que las cuestiones técnicas son uno de los elementos que aparecen con mayor frecuencia. De todas formas, seguramente sea la actitud de los investigadores con respecto de la divulgación de sus datos de investigación una de las cuestiones clave y, a su vez, más polémicas (Borgman, 2012). Además de las preocupaciones acerca de la idea de compartir libremente datos de investigación, muchos investigadores están poco dispuestos a dedicar el tiempo necesario para conservar correctamente sus datos de investigación, especialmente debido a que muchos no han recibido entrenamiento de cómo hacerlo. A pesar de que políticas para la preservación y el compartimiento de datos se han establecido por la *National Science Foundation* (2015), *the American Geophysical Union* (2013), y la *US Office of Science and Technology Policy* (2013), entre otras, no está claro aún si estas directrices motivarán a los investigadores a cumplirlo en un futuro. Sin duda, este es uno de los retos principales para mapear los procesos de gestión de los datos de investigación.

6 Políticas de retención e intercambio de datos de investigación

Para las investigaciones financiadas con fondos públicos hay algunas tendencias generalizadas para los períodos de retención de datos. Por ejemplo, para los investigadores del Reino Unido, el tiempo de retención de los datos es más largo. Es decir, no todo consejo de investigación determina un periodo de retención de datos necesario, y los que lo hacen, a menudo, requieren su retención durante al menos diez años (Cambridge University, 2010). Por ejemplo, la política de la *Engineering and Physical Sciences Research Council* (EPSRC) sobre la retención de datos reglamenta que las organizaciones de investigación se asegurarán de que los datos de la investigación financiada por la EPSRC sean conservados de forma segura durante el tiempo mínimo de 10 años a partir de la fecha que el investigador obtenga los datos o, si la obtención de los datos fue realizada por más investigadores, a partir de la última fecha en la que el acceso a los datos fue solicitado por un tercero (Engineering and Physical Sciences Research Council, 2014).

El tiempo mínimo de retención para la investigación financiada por el gobierno federal en los Estados Unidos es de tres años, según lo establecido por la *White House Office of Management and Budget* (2013). Los registros financieros, documentos de apoyo, registros estadísticos y todos los demás registros de entidades no federales pertinentes a un premio federal deben ser retenidos por un período de tres años a partir de la fecha de presentación del informe final de gastos o, para las concesiones federales que se renuevan trimestral o anualmente, a partir de la fecha de la presentación del informe financiero trimestral o anual, respectivamente, según ha informado la agencia federal de adjudicación (White House Office of Management and Budget, 2013).

En Australia, el período de retención recomendado es de cinco años, según lo establecido por el *National Health and Medical Research Council*, el *Australian Research Council* y las Universidades de Australia en el Código Australiano responsable por la conducta de investigación. En general, el periodo mínimo recomendado para la retención de los datos procedentes de científicos es de cinco años a partir de la fecha de publicación. Sin embargo, en casos particulares el período por el cual se deben conservar los datos debe ser determinado por el tipo específico de investigación. Por ejemplo:

1. Para los proyectos de investigación a corto plazo que sirven solamente para fines de evaluación, como los proyectos de investigación realizados por estudiantes, de retención de datos de investigación de 12 meses después de la finalización del proyecto puede ser suficiente.

2. Para la mayoría de los ensayos clínicos, puede ser necesario la retención de datos de investigación durante 15 años o más.

3. Para áreas como la terapia genética, los datos de investigación deberán ser conservados de forma permanente (por ejemplo, registros de pacientes)

4. Si el trabajo tiene valor para la comunidad o el patrimonio, los datos de investigación deben mantenerse permanentemente en esta etapa, de preferencia dentro de una colección nacional. (National Health and Medical Research Council et al 2007).

Así que la tendencia general de tiempo de retención suele oscilar entre los tres, cinco y diez años a partir de la finalización del proyecto o publicación para los Estados Unidos, Australia y el Reino Unido, respectivamente. Sin embargo, es recomendable consultar las políticas del fondo de investigación en particular para saber el punto de inicio y la duración exacta del resguardo obligatorio de los datos.

También es importante reconocer que algunos tipos de datos pueden requerir períodos de retención más largos, como se ejemplifica en la cita anterior del Código Australiano responsable por la conducta de investigación. En los Estados Unidos, el *Office of Research Integrity* señala que los datos involucrados en investigaciones confidenciales y patentes requieren períodos de retención más largos. También puede haber requisitos especiales en función del tema en cuestión. Por ejemplo, en el caso de investigaciones confidenciales con financiamiento de los *National Institutes of Health*, se deben conservar los registros durante seis años después de la fecha final de la resolución del caso. Como se señaló anteriormente, también es importante mantener los datos de investigación pertinentes a invenciones patentadas (Office of Research Integrity, 2014).

Al igual que en los Estados Unidos y Australia, los períodos de retención más largos se pueden encontrar para determinados tipos de datos en el Reino Unido. El *Medical Research Council* (MRC), por ejemplo, exige períodos de retención más largos para los datos de la investigación clínica. Las expectativas del MRC para la retención de datos de la investigación son:

- - Los datos de investigación y material conexo deben conservarse durante un mínimo de diez años después de que el estudio se haya completado.

- - Para la investigación clínica realizada en unidades e institutos de investigación del MRC, deben conservarse durante 20 años después de que el estudio se ha completado antes de permitir un período de seguimiento apropiado.

- - Los estudios que proponen periodos de retención más allá de 20 años deben incluir una justificación válida, por ejemplo, los datos de investigación relacionados con los estudios longitudinales suelen ser retenidos indefinidamente y archivados y gestionados en consecuencia. (Medical Research Council 2014).

Por lo tanto, los datos médicos, los datos de patentes o los datos importantes para la investigación longitudinal tienen períodos de retención más largos que los datos normales.

Un último punto a tener en cuenta sobre las políticas nacionales es que, a menudo, las políticas de intercambio y de retención de datos están intrínsicamente unidas. En el Reino Unido, por ejemplo, hay un mayor énfasis en el intercambio de datos que en la retención de datos, lo que conlleva a períodos de retención más largos; la expectativa es que el intercambio de datos se produzca a través de un repositorio de terceros como soporte para una retención más larga de los datos. Para un periodo de retención más largo, la recomendación es utilizar un repositorio de datos que conserve sus datos al mismo tiempo que los disponga al público. Otra opción es utilizar las estrategias contempladas en el capitulo 7 para gestionar sus propios datos, aunque esto no sea ideal para el intercambio de datos con largos tiempos de retención.

El lugar de trabajo es otra fuente común para las políticas de retención que se aplican a sus datos. Para los investigadores que trabajan en la industria, la empresa en que actúan es la propietaria de los datos y por lo tanto va a determinar el período de retención. Algunas compañías tienen normativas explícitas y otras no, lo principal es que la empresa sea responsable en última instancia por el mantenimiento a largo plazo de los datos.

Para los investigadores que trabajan en el ámbito académico, las expectativas son menos claras. Algunas universidades tienen una política explícita sobre la retención de datos, pero la mayoría no lo hacen. Tenemos aquí un ejemplo de una política universitaria de Harvard:

> "Los registros de investigación deben ser conservados, en general, por un período de no menos de siete años después del final de una actividad en los proyecto de investigación. Con este fin, un proyecto o actividad de investigación debe ser considerado como terminado después de (a) la presentación de informes final al patrocinador de la investigación, (b) cierre final de salida financiera de un premio de investigación patrocinado, o (c) la publicación final de los resultados de investigación, o (d) cese de la actividad académica o científica en un proyecto de investigación específico, independientemente de si se publican sus resultados, lo que sea más tarde. (Harvard University, 2011).

Las políticas universitarias a menudo reflejan las políticas de retención de datos nacionales, a veces requieren períodos de retención más largos y suelen llevar especificado cuánto tiempo se deben conservar los datos. Sin embargo, estas políticas no son específicas en relación a los requisitos para retener los datos, aplicado en términos generales a los datos de investigación generados en el marco de la política.

En relación a los requisitos de intercambio, los organismos de financiación son la principal fuente de consulta. En los EE.UU., los *National Institute of Health* (NIH) y la Fundación Nacional de Ciencia (NSF) han sido los principales motores para el intercambio de datos, y los datos que comparten las políticas se han extendido a otras agencias federales de financiamiento. En 2013, la Oficina de Ciencia y Tecnología de la Casa Blanca publicó un memorandum sobre el acceso público (Holdren, 2013) requiriendo que los principales organismos de financiación federal de los Estados Unidos promulgasen un método para la gestión de datos incluyendo los requisitos de intercambio. Esta nota se ha tornado como parámetro a seguir en el intercambio de datos, además de ser un criterio para recibir fondos federales en los Estados Unidos. En el Reino Unido, el *Research Councils UK* y el *Wellcome Trust* han sido los principales impulsores de la política de intercambio de datos desde mediados de la década de 2000 (Wellcome Trust 2010; Research Councils UK 2011). Sus políticas son generalmente más fuertes que las de donantes de Estados Unidos, llamando a "investigadores para maximizar la disponibilidad de los datos de investigación, con el menor número de restricciones posibles" (Wellcome Trust 2010). En su conjunto, la tendencia general de los proveedores de fondos federales de investigación es la adopción de los requisitos de intercambio de datos con muchos proveedores de fondos más pequeños siguiendo su ejemplo.

7 Conclusiones

La recopilación, creación, análisis, interpretación y gestión de los datos requiere de experiencia en el ámbito de la investigación y ello sugiere una nueva composición de la información científica, conforme análisis que presentamos en el apartado "Políticas de retención e intercambio de datos de investigación". Thanos (2013) apunta que esta nueva ciencia de datos dará lugar a una nueva forma de organizar y llevar a cabo las actividades de investigación, hecho que podría desembocar en un replanteamiento de los enfoques a la hora de resolver problemas de investigación y conducir descubrimientos fortuitos. La reciente disponibilidad de poder acceder a grandes cantidades de datos, junto con las herramientas avanzadas de análisis exploratorio de datos como la minería y la visualización de datos, también comportará un importante cambio dentro de la metodología científica. Uno de los puntos de vista que se viene planteando es que el método científico tradicional impulsado mediante hipótesis, lo que consiste básicamente en un método deductivo, se complementará con un método basado en datos, lo que quiere decir, esencialmente inductivo, conforme hemos visto en el apartado sobre “el valor de los datos”.

Con el objetivo de ser capaces de explotar estos grandes volúmenes de datos, se necesitan nuevas técnicas y tecnologías. Para que estas sean de utilidad entre la comunidad científica, se requiere de una estructura y organización jerárquica, deben constituirse colecciones relacionadas entre sí y registradas en un formato adecuado al objetivo por el cual se han recogido, así como siempre deben ir acompañadas de un cuerpo descriptivo (los metadatos), que incluya, entre otras cosas, la autorización legal para acceder y difundir sus

contenidos. Es lo que hemos tratado en los apartados sobre "El papel de los investigadores" y "Las infraestruturas de gestión de datos"

Referências

AMERICAN GEOPHYSICAL UNION. **AGU publications data policy**. 2013. Disponíble en: < http://publications.agu.org/author-resource-center/publication-policies/data-policy/ >. Acceso en: 11 jul. 2015.

BORGMAN, Christine L. The conundrum of sharing research data. **Journal of the American Society for Information Science and Technology** . n. 63, p. 1059–1078. 2012.

CONICYT, IDER. **Estado del arte nacional e internacional en materia de gestión de datos de investigación e información científica y tecnológica y recomendaciones de buena prácticas**. Santiago: Gobierno de Chile. 2010.

DIGITAL Curation Centre. **DCC Curation Lifecycle Model**. Disponíble en: < http://www.dcc.ac.uk/resources/curation-lifecycle-model >. Acceso en: 13 abril 2016.

Engineering and Physical Sciences Research Council. **Expectations**. Disponible en: < https://www.epsrc.ac.uk/about/standards/researchdata/expectations/ >. Acceso en: 22 nov. 2015.

European Commission. **Scientific data**: open access to research results will boost Europe's innovation capacity. 2012. Disponible en: < http://europa.eu/rapid/press-release_IP-12-790_en.htm >. Acceso en: 20 nov. 2015.

EUROPEAN COMISSION. **Scientific information in the digital age**: Ensuring current and future access for research and innovation. Brussels. Feb. 2007. Disponible en: < http://europa.eu/rapid/press-release_IP-07-190_en.htm?locale=en >. Acceso en: 10 jul. 2015.

FALCONE, Andres Araya. **Big Data**: El 90% de los datos existentes en el mundo han sido creados en los últimos 2 años. Disponíble en < http://andresarayafalcone.blogspot.com.es/2011/10/big-data-el-90-de-los-datos-existentes.html >. Acceso en: 10 dic. 2013.

Grupo de Trabajo de "Depósito y Gestión de datos en Acceso Abierto" del proyecto RECOLECTA. **La conservación y reutilización de los datos de investigación en España**. Informe del grupo de trabajo de buenas prácticas. Madrid: Fundación Española para la Ciencia y la Tecnología, FECYT, 2012. Disponíble en: < https://universoabierto.com/2016/01/02/la-conservacion-y-reutilizacion-de-los-datos-cientificos-en-espana/>. Acceso en: 08 jun. 2013.

HOLDREN, John P. **Increasing Access to the Results of Federally Funded Scientific Research**. Office of science and technology policy. Disponible en: < https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf >. Acceso en: 20 nov. 2015.

European Commission. **Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020**. v. 1.0. Dec. 2013.

NATIONAL SCIENCE FOUNDATION. **Dissemination and Sharing of Research Results**. 2015. Disponible en: < http://www.nsf.gov/bfa/dias/policy/dmp.jsp >. Acceso en: 11 jul. 2015.

OECD. **Principles and Guidelines for Access to Research Data from Public Funding**. Paris, 2007. Disponible en: < http://www.oecd.org/sti/sci-tech/38500813.pdf >. Acceso en: 27 ene. 2016.

OPEN DATA COMMONS. **Public Domain Dedication and License (PDDL)**. 2008. Disponible en: < http://opendatacommons.org/licenses/pddl/1-0/ >. Acceso en: 20 nov. 2014.

KNOWLEDGE EXCHANGE. Disponible en: < http://www.knowledge-exchange.info/ >. Acceso en: 09 sep. 2015.

KOTARSI, Rachael Kotarski; REILLY, Susan Reilly; SCHRIMPF, Sabine; SMIT, Eefke; WALSHE, Karen. **Report on best practices for citability of data and on evolving roles in scholarly communication**. 2012.

LIBER 40th Annual Conference. Universitat Politècnica de Catalunya. Barcelona, 29 June. 2011. Disponíble en: < http://liber2011.upc.edu >. Acceso en: 30 nov. 2014.

MOSSINK, Wilma; BIJSTERBOSCH, Magchiel; NORTIER, Joeri. SIM4RDM. **European Landscape Study of Research Data Management**. 2013. Disponible en: < http://www.sim4rdm.eu >. Acceso en: 23 nov. 2014.

REILLY, Susan, et. al. Report on integration of data and publications. **Opportunities for data exchange**. Oct. 2011. Disponíble en: < http://epic.awi.de/31397/ >. Acceso en: 30 nov. 2014.

THANOS, C. A vision for global research data infrastructures. **Data Science Journal**, Sept. 2013. 71–90.

University of Cambridge, 2010. UK Funding Councils: Data Retention and Access Policies. Disponible en: < http://find.jorum.ac.uk/resources/bitstream/377746 >. Acceso en: 22 nov. 2015.

VAN DER GRAAF, Maurits; WAAIJERS, Leo. A Surfboard for Riding the Wave. Towards a four country action programme on research data. A Knowledge Exchange Report. 2011. Disponible en: < http://eprints.gla.ac.uk/80155/ >. Acceso en: 11 jul. 2015.

Wellcome Trust, 2010. Policy on Data Management and Sharing. Disponible en: < http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002753.htm >. Acceso en: 12 nov. 2015.

White House Office of Management and Budget. Uniform Administrative Requirements, Cost Principles, and Audit Requirements for Federal Awards. 2013. Disponible en: < https://www.federalregister.gov/articles/2013/12/26/2013-30465/uniform-administrative-requirements-cost-principles-and-audit-requirements-for-federal-awards >. Acceso en: 18 nov. 2015.

Datos del autor

Fabiano Couto Corrêa da Silva

Profesor del Curso de Biblioteconomia de la Universidade Federal do Rio Grande (Brasil). Doctorando del Curso Información y Documentación en la Sociedad del Conocimiento de la Universidad de Barcelona. k fabianocc@gmail.com

Recibido - Received: 2016-01-27

Aceptado - Accepted: 2016-05-07

1 Conferencia realizada en Barcelona durante el evento BigDataCoe. Disponible en: < Oracle y Barcelona Digital abren un centro 'big data' de referencia europea >. Acceso en: 18 sep. 2015.

2 Entrevista Disponible en: < http://www.bcs.org/content/ConWebDoc/3337 >. Acceso en: 19 mar. 2016.

3 Una de las iniciativas de la Open Data Commons es la distribución del Manual de Periodismo de Datos. Disponible en: < http://interactivos.lanacion.com.ar/manual-data >. Acceso en: 20 nov. 2014. En los últimos años, cambiaran los conceptos en favor de los datos abiertos, u Open Data. Si el término Open Data es reciente, el concepto y la práctica es un fenómeno antiguo. El movimiento de los datos de investigación abiertos establece que determinados datos estén disponibles de manera gratuita, sin restricciones de copyright, patentes o mecanismos de control.

4 La MozFest es un evento anual organizado por la Fundación Mozilla que reúne interesados en discutir el futuro de la web y en compartir experiencias innovadoras.