

## Compartir datos de investigación: una realidad necesaria



Por [Sandra Sánchez García](#)<sup>i</sup>

Directora Técnica de la Biblioteca General del Campus de Cuenca, UCLM  
Coordinadora editorial de Ocnos: revista de estudios sobre lectura

**Palabras clave:** Ocnos, Open Science, Open Data, datos de investigación, publicación científica

*Publicado el 17 de julio de 2020.*

Hace casi cuatro años comenzaba mi colaboración en el blog Aula Magna 2.0 con un artículo dedicado a la necesidad de compartir los datos de investigación como una práctica cada vez más necesaria pero en la que aún nos quedaba un largo camino por recorrer. En ese trabajo se describían los “beneficios e incertidumbres” (Sánchez-García, 2016) que presentaba en ese momento el *data sharing*, en un contexto en el que aún no éramos del todo conscientes de su necesidad y en el que aún faltaba mayor desarrollo normativo a nivel nacional e institucional.

Ha tenido que ser necesario el paso de los años y la consolidación de las políticas de *Open Science*, para que sobre todo en el ámbito académico, seamos conscientes de sus implicaciones en los avances de la investigación científica. Tener acceso a los datos de investigación supone un mayor rendimiento de dinero invertido en ciencia, y permite a otros investigadores replicar y reproducir los estudios, gracias a la reutilización de los datos. Pero ha sido una crisis sanitaria a nivel global, y la difusión de estas prácticas en los medios de comunicación, las que han favorecido que todos fuésemos conscientes de las utilidades de los repositorios de datos abiertos. Como señala Martínez-Samper (2020) los resultados científicos “para entender el coronavirus SARS-CoV-2 y frenar la pandemia han ocupado un espacio relevante a la vez que inusual en conversaciones, redes sociales y telediarios”. Esto ha permitido visibilizar y concienciar a toda la población sobre la importancia de movimientos como el *Open Access*, *Open Science* y *Open Data*. Movimientos vinculados directamente con el ámbito académico y científico, que llevaban más de 20 años fraguándose y que han necesitado muchos apoyos e inversiones para que hoy puedan ser una realidad.

**Políticas sobre datos de investigación abiertos**

Varios han sido los hitos que nos han llevado a construir este escenario en el que, ya no cabe la menor duda, todos debemos contribuir como agentes de cambio.

A nivel internacional, el desarrollo de la Agenda 2030 para el Desarrollo Sostenible (ONU, 2015), señalaba los grandes retos a los que nos debíamos enfrentar en los próximos años, presentando el acceso abierto como una de las metas a la hora de garantizar el acceso público a la información y proteger las libertades fundamentales; además del acceso a los datos, para construir instituciones eficaces y transparentes capaces de rendir cuentas a la sociedad.

En el ámbito científico, es importante destacar el papel que ha jugado la Comisión Europea a partir de las convocatorias de los proyectos Horizonte 2020. El Programa Horizonte2020, establece en su artículo 29 la obligatoriedad de publicar en abierto los resultados de las investigaciones, y desde enero de 2017 incluye la obligatoriedad de publicar el Programa de Gestión de datos en los primeros seis meses del proyecto, además del depósito de los mismos en abierto para todas las áreas científicas. Entre los requisitos señala la necesidad de asegurarse de que terceros puedan acceder, extraer, explotar, reproducir y diseminar los datos, a través del uso de licencias; e indicar claramente que herramientas se requieren para usar los datos en bruto para validar los resultados de investigación (o proporcionar las herramientas directamente).

Finalmente, la *Directiva (UE) 2019/1024 del Parlamento Europeo y el Consejo relativa a los datos abiertos y la reutilización de la información del sector público*, impone a los Estados miembros la obligación de adoptar políticas de acceso abierto con respecto a los datos de la investigación financiada públicamente y garantizar que dichas políticas son ejecutadas por todas las organizaciones que realizan actividades de investigación y las organizaciones que financian la investigación.

A nivel nacional, el primer gran hito hacia el acceso abierto fue la promulgación de Ley 14/2011 de la ciencia, la tecnología y la innovación, en la que se establece el mandato de depositar en abierto tan pronto como sea posible y no después de los 12 meses las publicaciones de los resultados de investigación financiados con fondos públicos. A esta ley le siguieron la Ley 19/2013 de transparencia, acceso a la información pública y buen gobierno y la Ley 18/2015, por la que se modifica la Ley 37/2007 sobre reutilización de la información del sector público. Este marco legislativo ha permitido el desarrollo de proyectos nacionales como datos.gob.es.

El documento estatal que se refiere de forma explícita al uso compartido de datos de investigación es el *Plan Estatal de Investigación Científica y Tecnológica y de Innovación* (Ministerio de Industria Economía y Competitividad, 2017). Este documento, que comprende la convocatoria de proyectos desde 2017 a 2020, menciona la importancia de compartir los datos de investigación, pero lo hace como una recomendación y siempre de manera opcional, aunque menciona posibles incentivos en la evaluación curricular de los investigadores por compartir.

Todo este marco legal y normativo ha conseguido consolidar una práctica cada vez más habitual entre los investigadores, si bien es cierto que todavía se limita a ciertas áreas y a un perfil determinado de investigador.

### **Características y requisitos de los datos abiertos**

El movimiento *Open Data* se integra dentro del nuevo paradigma de hacer ciencia, con el objetivo de que esta sea abierta, transparente, colaborativa, accesible y hecha con y para la sociedad. Conocido como *Open Science*, este movimiento promovido por los países de la OCDE e impulsado por la Comisión Europea, implica la apertura de las publicaciones (*Open Access*), la apertura de los datos de investigación (*Open Data* y *Linked Open Data*), de los procesos, de la metodología, del software, de los flujos de trabajo científicos.

*Open Data* trata de hacer los resultados de la investigación financiada con fondos públicos accesibles en formato digital para la comunidad científica que los produce, así como para la sociedad en general que los financia. Esta iniciativa mundial se engloba en un proceso más amplio que afecta también a los datos generados por los gobiernos y las administraciones públicas, para aportar transparencia a la gestión, favorecer la participación ciudadana, la innovación y para generar riqueza y bienestar.

Para favorecer la confluencia entre transparencia y reutilización los datos abiertos deben cumplir las siguientes características, conocidas por bajo las siglas FAIR:

- Findable: los datos y los metadatos deben ser encontrables mediante herramientas de búsqueda, y recuperables con identificadores persistentes.
- Accessible: los datos y los metadatos deben cumplir los estándares de accesibilidad y pueden ser descargados por otros investigadores utilizando sus identificadores.
- Interoperable: los datos y los metadatos deben estar descritos según estándares abiertos y normalizados que permitan el intercambio y la reutilización.
- Reusable: los datos y los metadatos deben ser susceptibles de ser utilizados por otros investigadores, con información de las condiciones de reutilización (procedencia, derechos y permisos de licencias).



Figura 1. Principios FAIR (Australian National Data Service)

Para el cumplimiento de estas características actualmente tenemos a nuestra disposición las tecnologías de la web semántica y el *Linked Open Data* (LOD), a partir del desarrollo de estándares para la descripción de recursos, datos y metadatos que permiten el acceso a datos abiertos y enlazados provenientes de diversas fuentes e instituciones.

Atendiendo a estas características Tim Berners-Lee desarrolla en 2010 el modelo LOD, que permite clasificar los datos abiertos en 5 niveles de calidad en función de la apertura de los formatos en los que los datos son compartidos.

- 1 estrella: datos abiertos no estructurados (documentos PDF, imágenes JPG, TIFF).
- 2 estrellas: formatos estructurados con software propietario (hojas de cálculo xls, documentos doc.).
- 3 estrellas: datos abiertos estructurados con software libre (OpenOffice, csv, xls, json).

- 4 estrellas: datos estructurados con URIS que identifican recursos y utilizan estándar RDF.
- 5 estrellas: además de lo anterior están enlazados con otros datos a través de formatos semánticos (RDF/XML, OWL/XML).

Los metadatos que describen estos datos son un elemento clave para facilitar el acceso y reutilización de los mismos, ya que identifican las características y estructura de los datos o *datasets* y sus medidas de preservación y acceso. Más allá de los metadatos mínimos recomendado por la *Guía de aplicación de la Norma Técnica de Interoperabilidad de Reutilización de Recursos de la información* (2016), en la que se recomienda el uso de los esquemas Dublin Core (DC), y Dataset Catalog (dcat), estándar desarrollado por la W3C, el documento *Conservación y reutilización de los datos científicos en España* (FECYT-REBIUN, 2012), recomiendan el uso de Data Documentation Initiative (DDI) y DataCite, diseñados específicamente para de descripción de conjuntos de datos de investigación.

Finalmente, para propiciar la reutilización de los datos de investigación estos deben compartirse bajo licencias abiertas que definan e identifiquen de forma clara las posibilidades de reutilización de los datos. Siguiendo el modelo de las LCC que definen 6 modelos de licencias para las publicaciones en abierto a partir de la combinación de 4 condiciones de uso (BY Atribución, NC Uso No Comercial, ND Obra No Derivada y SA Compartir Igual), la Open Knowledge Foundation, desarrolla en 2009 las Open Data Commons que contemplan 3 licencias de uso de datos abiertos:

- PDDL: datos de dominio público
- ODC-BY: exige la referencia a la autoría o fuente para su reutilización
- ODC-ODbl: además del reconocimiento de la fuente se debe compartir bajo la misma licencia.

Todos estos aspectos deben quedar recogidos en el Data Management Plan o Plan de Gestión de Datos (DMP o PDG), que se convierte en un documento clave dentro de los proyectos de investigación, ya que marcan la ruta que seguirán los datos, a partir de un documento vivo que describe cómo serán gestionados los datos recogidos o generados durante y después del proyecto de investigación.

La Comisión Europea proporciona orientación sobre cómo elaborar el plan de gestión de datos en *Guidelines on FAIR Data Management in Horizonte 2020*, traducido al español por el Consorcio Madroño, bajo el título *Directrices para la Gestión de Datos en Horizonte 2020*. Este documento deberá reunir los siguientes elementos:

- Quién se responsabiliza de la custodia de los datos de investigación.
- Tipos de datos que se van a recoger durante la investigación.
- Estándares para la recogida y preservación de los datos.
- Políticas de acceso y reutilización de los datos.
- Medidas de seguridad, propiedad intelectual, privacidad y confidencialidad.
- Archivo y preservación de los datos.

El PGD debe de tener en cuenta, además, el volumen de datos y los formatos de archivo, los costes previstos para cumplir con estas directrices y su preservación a largo plazo, la seguridad de los datos y las cuestiones éticas.

**10 pasos para elaborar un Plan de Gestión de Datos**

Un Plan de Gestión de Datos (PGD) o Data Management Plan (DMP) es un documento formal, que debe presentarse al inicio de la investigación, en el que se describe qué vas a hacer con tus datos durante y después de finalizar tu investigación y que puede modificarse si se producen cambios en el proceso de la misma.

**¿Por qué?**  
Es una buena práctica, es un elemento clave de Open Science y es obligatorio en los nuevos proyectos H2020.

**Herramientas gratuitas para elaborar un PGD**

- PGDonline (Consorcio Madroño) <http://dmp.consorcio.madroño.es/>
- DMPonline (Digital Curation Centre, UK) <https://dmponline.dcc.ac.uk/>

- 01 Revisa los requerimientos de la entidad financiadora (H2020).
- 02 Identifica los datos: tipología, procedencia, volumen, formatos y ficheros.
- 03 Define cómo se organizarán y gestionarán los datos: nombre de los ficheros, control de versiones, software necesario...
- 04 Explica cómo se documentarán los datos: identifica la información a procesar, consulta si hay estándares o esquemas de metadatos, identifica herramientas que permitan gestionarlos.
- 05 Describe los procesos que aseguran una buena calidad de los datos.
- 06 Prepara una estrategia de almacenamiento (durante el proceso) y de preservación de datos (repositorio).
- 07 Define las políticas de datos del proyecto: cuestiones sobre propiedad intelectual y cómo se tratarán los datos sensibles y personales.
- 08 Describe cómo se difundirán los datos: dónde, cuáles, cuándo se van a difundir. Si publicarás los datos en un repositorio, como información suplementaria del artículo o como un "data paper".
- 09 Asigna roles y responsabilidades para las personas y organizaciones participantes en el proyecto.
- 10 Prepara un presupuesto realista: la gestión de datos cuesta tiempo y dinero en términos de software, hardware, servicios y personal.

Adaptado de Michener, WK. Ten simple rules for creating a good data management plan. PLOS Comp Biol. 2015; 11(10)

El personal de tu Biblioteca te puede asesorar

Figura 2. 10 Pasos para elaborar un Plan de Gestión de Datos (REBIUN)

### Infraestructuras tecnológicas para el depósito abierto de datos

El despliegue de las políticas de acceso abierto a los datos ha sido posible gracias al desarrollo de una serie de plataformas tecnológicas que han permitido el depósito de estos datos. La evolución del desarrollo de estos repositorios se observa principalmente en los datos recogidos en el directorio OpenDOAR, donde se recogen actualmente 373 repositorios de *datasets*, doblando la cifra de los datos recogidos en el trabajo de 2016. Actualmente, existe un gran número de repositorios que aceptan el depósito de datos de investigación, entre ellos destacan, por un lado los repositorios temáticos, específicos de un área concreta, y, por otro, los repositorios multidisciplinares, que son de tipo generalista y sirven para cualquier área. La oferta es tan amplia que requieren de un análisis previo por parte del investigador para seleccionar aquel que se ajusta mejor a las características de sus proyectos. Sixto-Costoya et al. (2019) señalan los siguientes aspectos a tener en cuenta:

Lo que el investigador necesita de un repositorio	Lo que el repositorio necesita del investigador (o del usuario que vaya a depositar los datos)
<ul style="list-style-type: none"> <li>· Saber si acepta datos de su disciplina.</li> <li>· Que sea intuitivo.</li> <li>· Que ponga a disposición de usuarios guías y protocolos sencillos de seguir.</li> <li>· Que, si el investigador lo necesita, ofrezca diferentes grados de apertura y periodo de embargo.</li> <li>· Que garantice la preservación de los datos.</li> <li>· Que garantice mecanismos de seguridad de los datos.</li> <li>· Saber cómo gestiona la propiedad intelectual y los derechos de autor.</li> <li>· Saber si ofrece un identificador persistente.</li> </ul>	<ul style="list-style-type: none"> <li>· Que el investigador se informe previamente de si ese repositorio es adecuado para su trabajo.</li> <li>· Que los datos estén en un formato reutilizable, es decir, libre y abierto para que, dado el caso, otros puedan usarlos y se asegure el cumplimiento de los principios FAIR.</li> <li>· Que los datos vayan acompañados de metadatos, es decir, de un contexto que les dote de significado y los haga comprensibles.</li> </ul>

Figura 3. Tabla Aspectos para la selección de repositorios de datos (Sixto-Costoya et al., 2019)

En cualquier caso queremos destacar algunos de los repositorios multidisciplinares más utilizados por los investigadores:

- a. **Zenodo:** repositorio multidisciplinar alojado en CERN y financiado por la UE que permite a investigadores, científicos, proyectos europeos e instituciones:
  - Compartir los resultados de investigaciones de todos los campos científicos en una amplia variedad de formatos incluyendo texto, hojas de cálculo, audio, video e imágenes.
  - Mostrar los resultados de sus investigaciones y obtener reconocimiento al posibilitar su citación e integrarlos en líneas de información ya existentes con agencias de financiación como la Comisión Europea.
  - Acceder y reutilizar fácilmente los resultados de investigaciones compartidos.
  - Integrar los resultados de sus investigaciones en el portal OpenAIRE, a partir de la asignación de DOIs.
- b. **FigShare:** repositorio digital en línea de datos de ciencia abierta, apoyado por la Digital Science-MacMillan Publishers Company, para que los investigadores puedan preservar, compartir los datos y resultados de la investigación haciéndolos accesibles mediante la asignación automática del DOI. La gama de archivos que acepta es muy amplia, y va desde los propios artículos, a imágenes, videos, *datasets* o posters.
- c. **Dataverse:** proyecto colaborativo del Institute for Quantitative Social Science (IQSS) de la Universidad de Harvard. Es una aplicación web de código abierto para compartir, preservar, citar, explorar y analizar datos de investigación. Un elemento diferenciador con otros repositorios, es la posibilidad que ofrece tanto a investigadores como a instituciones y revistas de la instalación de un software que permite la creación de un espacio independiente.

El desarrollo a nivel internacional de estos repositorios ha facilitado enormemente la difusión de datos abiertos en los últimos años, convirtiéndose en las herramientas clave a la hora de compartir datos de investigación.

### Las revistas científicas en el uso compartido de datos

A la hora de compartir datos es imprescindible tener en cuenta el papel que juegan en este contexto las revistas científicas y sus editores. Es importante que las revistas adopten medidas para favorecer las prácticas de *data sharing* entre sus autores, algo que ya han comenzado a hacer los grandes grupos editoriales. Como señalan Sixto-Costoya et al. (2019), actualmente nos encontramos básicamente con tres formas o niveles a la hora de favorecer el uso compartido de datos.

1. En este primer nivel nos encontramos con revistas que permiten a los autores indicar que los datos pueden ser solicitados directamente al autor o autores, facilitando la vía de contacto.
2. En un segundo nivel nos encontramos con aquellas revistas que ofrecen la posibilidad de que los datos brutos vayan adjuntos como material suplementario al artículo. En esta modalidad los datos se almacenan en la propia plataforma digital de la revista.
3. El tercer nivel incluye a aquellas revistas que obligan o recomiendan a los autores depositar los datos en un repositorio de datos abiertos. De esta forma los datos están guardados a través de una modalidad que ofrece mejores garantías de preservación y accesibilidad. En este tipo de revistas, el editor generalmente selecciona el repositorio en el que se debe hacer el depósito y facilita el acceso a los mismos a través del DOI que genera la plataforma en la que se encuentran los datos.

Sin duda, el momento que estamos viviendo nos presenta una oportunidad excepcional para que las revistas científicas y todos los agentes implicados en la publicación y difusión científica demos el impulso necesario a una práctica que se ha ido consolidando en los últimos años pero que la Covid-19 ha señalado como inaplazable. Es el momento de dar un paso más allá y ser capaces de enriquecer los artículos y publicaciones de investigación con sus datos, lo que ayudará a una mejor comprensión y avance de la ciencia.

### Referencias bibliográficas:

- Consorcio Madroño (2013). *Directrices para la Gestión de Datos en Horizonte 2020*. [https://www.consorciomadrono.es/wp-content/uploads/2017/05/directrices\\_gestion\\_datos\\_horizon\\_2020\\_es.pdf](https://www.consorciomadrono.es/wp-content/uploads/2017/05/directrices_gestion_datos_horizon_2020_es.pdf)
- Dirección de Tecnologías de la Información y las Comunicaciones (DTIC) (2016). Guía de aplicación de la *Norma Técnica de Interoperabilidad de Reutilización de Recursos de la información*. Ministerio de Hacienda y Administraciones Públicas. [https://administracionelectronica.gob.es/pae/Home/pae\\_Estrategias/pae\\_Interoperabilidad\\_Inicio/pae\\_Normas\\_tecnicas\\_de\\_interoperabilidad.html](https://administracionelectronica.gob.es/pae/Home/pae_Estrategias/pae_Interoperabilidad_Inicio/pae_Normas_tecnicas_de_interoperabilidad.html)
- FECYT-REBIUN (2012). *Conservación y reutilización de los datos científicos en España*. <http://hdl.handle.net/10261/65317>
- Martínez-Samper, P. (2020). Conocimiento compartido para no dejar a nadie atrás. *Ctxt. Revista Contexto*, 261. <https://ctxt.es/es/20200601/Firmas/32587/agenda-2030-universidad-ciencia-abierta-covid19-pastora-martinez-samper.htm>
- Ministerio de Industria Economía y Competitividad (2017). *Plan Estatal de Investigación Científica y Tecnológica y de Innovación*. <https://www.ciencia.gob.es/stfls/MICINN/Prensa/FICHEROS/2018/PlanEstatallIDI.pdf>
- Parlamento Europeo (2019). Directiva (UE) 2019/1024 del Parlamento Europeo y el Consejo relativa a los datos abiertos y la reutilización de la información del sector público. *Diario Oficial de la Unión Europea*, 26/06/2019, L. 172, 56-78. <https://www.boe.es/doue/2019/172/L00056-00079.pdf>
- Sánchez-García, S. (2016). Compartir datos de investigación: beneficios e incertidumbres. Debates en torno a las métricas fiables de reputación académica. *Aula Magna 2.0*. [Blog]. <https://cuedespyd.hypotheses.org/2245>
- Sixto-Costoya, A., Aleixandre-Benavent, R., Vidal-Infer, A., Lucas-Domínguez, R., & Castello-Cogollos, L. (2019). *Data sharing: qué son y cómo se pueden compartir los datos de investigación. Manual de recomendación para gestores de la información*. SEDIC. [https://www.sedic.es/wp-content/uploads/2020/01/Data\\_Sharing-DEF.pdf](https://www.sedic.es/wp-content/uploads/2020/01/Data_Sharing-DEF.pdf)

### Cómo citar esta entrada:

Sánchez García, S. (2020). Compartir datos de investigación: una realidad necesaria. *Aula Magna 2.0*. [Blog]. Recuperado de: <https://cuedespyd.hypotheses.org/8398>

---

<sup>1</sup> Licenciada en Documentación y Doctora en la Filología Hispánica y Clásica, es coordinadora editorial de Ocnos: Revista de estudios sobre lectura, e investigadora del Centro de Estudios de Promoción de la lectura y literatura Infantil (CEPLI) de la UCLM.