Implementación de los repositorios de datos de investigación en las universidades públicas españolas: estado de la cuestión

Implementation of research data repositories in Spanish public universities: state of the art

Francisco Javier Martínez Méndez (1), Ana Alice Baptista (2), Rosana López Carreño (3), Ángel María Delgado Vázquez (4)

(1) (3) Facultad de Comunicación y Documentación de la Universidad de Murcia, javima@um.es, rosanalc@um.es
(2) Escola de Engenharia da Universidade do Minho, analice@dsi.uminho.pt
(2) Área de Biblioteconomía y Documentación, Universidad Pablo de Olavide, adelvaz@bib.upo.es

Resumen

La Ciencia Abierta se abre paso poco a poco en el seno de las instituciones de educación superior. Aspira a conseguir que los resultados de las investigaciones científicas sean accesibles, reutilizables y transparentes. Para ello, es especialmente importante el acceso abierto a los datos de investigación. En este trabajo se analiza cómo se está llevando a cabo el depósito de los conjuntos de datos en las universidades públicas españolas, su nivel de implementación, qué plataformas se están implementando y conocer si se está orientando y apoyando a los investigadores en esta tarea. Los resultados obtenidos muestran una realidad diversa e incipiente. Por un lado, están dos consorcios regionales de universidades (Cataluña y Madrid) que han apostado claramente por la gestión de los conjuntos de datos de investigación. En el otro, aparece un modelo estándar de depósito de esta información como si de colecciones de un repositorio institucional se tratara. Las universidades con más conjuntos de datos publicados pertenecen a uno de los consorcios, implementan un software más avanzado y son un modelo a considerar por las otras instituciones, especialmente aquellas que apenas han publicado unos pocos conjuntos de datos. El apoyo de las administraciones regionales es también un factor a considerar y una oportunidad para las comunidades con varias universidades.

Palabras clave: Ciencia abierta. Conjuntos de datos de investigación. Repositorios de datos de investigación. Universidades públicas. España.

Abstract

Open Science is gradually making its way into the heart of higher education institutions. This paradigm shift aims to make the results of scientific research accessible and reusable by other researchers. In this regard, open access to research data is particularly important, in order to facilitate their reuse and increase their level of transparency. Our paper aims to analyse how the deposit of these datasets is being carried out in Spanish public universities, their level of implementation, which platforms are being implemented and whether researchers are being guided and supported in this task. The results obtained show a diverse and incipient reality. On the one hand, there are two regional consortia of universities (Catalonia and Madrid) that are clearly committed to the management of research datasets. On the other hand, there is a standard model for depositing this information as if they were collections in an institutional repository. Universities with more published datasets belong to one of the consortia, implement more advanced software and are a model to be considered by the others, especially those that have only published a few datasets. Support from regional administrations is also a factor to consider and an opportunity for communities with several universities.

Keywords: Open science. Research datasets. Research data repositories. Public universities. Spain.

1. Introducción

En el transcurso de muchas investigaciones se generan datos que sirven para la obtención de resultados y la extracción de conclusiones. Dentro del campo de la Ciencia Abierta (UNESCO, 2021) viene cobrando impulso la idea de publicar, junto al artículo o informe derivado de la investigación el conjunto de datos empleados en la misma con vistas a favorecer la reutilización de los mismos por parte de otros investigadores, fomentando así a transparencia de la labor investigadora y garantizando la integridad de todo el proceso (Borghi & Van Gulick, 2019). La Ciencia Abierta representa un nuevo paradigma que incorpora una visión holística del proceso de generación de conocimiento, a partir del diseño inicial de un proyecto y su desarrollo hasta la comunicación, difusión y preservación de sus resultados (Abadal et al, 2023). Dentro de este entorno cobra especial importancia la gestión de los datos de investigación: "un proceso diseñado para gestionar y difundir conjuntos de datos de alta calidad, que cumplan con los requisitos académicos, legales y éticos establecidos" (Alonso-Arévalo,

2019): Esta gestión se encuentra entre las principales líneas de trabajo, no ya futuras sino actuales, de las bibliotecas académicas (Federer & Qin, 2019), si bien no está implantada de manera estándar (Ayris & Ignat, 2018; Ashiq, 2022). Asimismo, esta gestión constituye uno de los ejes estratégicos de sobre los que se estructura la Estrategia Nacional de Ciencia Abierta de España (2023), recientemente aprobada.

El acceso a los datos de investigación es imprescindible para la reproducibilidad de los resultados científicos, facilita la cooperación interdisciplinar, estimula el crecimiento económico a través de mejores oportunidades para la innovación, permite la reutilización de datos, aumenta la eficiencia de los recursos, mejora la transparencia, la rendición de cuentas y la confianza en los resultados de la investigación científica (OCDE 2021). Esta gestión se lleva a cabo sobre los conjuntos de datos de investigación ('datasets') que recogen información estructurada y organizada recopilada o generada en el desarrollo de una investigación. Los conjuntos contienen información relevante y detallada que se utiliza para respaldar los objetivos, análisis y resultados del estudio científico. Los datos pueden ser primarios o secundarios, cuantitativos o cualitativos, y pueden estar disponibles en diferentes formatos y tamaños. Es totalmente recomendable que los datos cumplan los principios FAIR ('findability, accessibility, interoperability, reuse') que establecen pautas para la correcta gestión de datos (Wilkinson et al., 2016).

Con la idea de cumplir estos principios, se han desarrollado diversas recomendaciones de buenas prácticas para llevar a cabo su gestión, tales como las recogidas en el documento *FAIR Data Maturity Model. Specification and Guidelines* (2020) que ofrece un conjunto de directrices y una lista de comprobación relacionada con el cumplimento de estos principios, o las que recogen Rocca-Serra et al. en su proyecto *FAIR Cookbook* (2023). En España, la red de bibliotecas universitarias elaboró su propio documento de recomendaciones para gestionar los datos de investigación (REBIUN, 2017). Los requerimientos técnicos, tecnológicos, así como la implicación bibliotecaria y de los responsables de la publicación en los repositorios, son una pieza fundamental para lograr una óptima implantación y, por tanto, para la visibilidad y reutilización de dichos datos en futuras investigaciones.

La implementación de los repositorios de datos de investigación en universidades y centros de investigación es un signo más de esta tendencia conducente a disponer en acceso abierto fuentes de información específicas y bien definidas en torno a los datos generados durante las investigaciones. Como se ha indicado anteriormente, esta publicación debe seguir los principios FAIR en los propios datos, los mismos deben estar depositados en repositorios fiables y han de adoptar licencias tipo CC-BY o CC-0 (1) para compartirlos.

Debemos recordar que esto se debe llevar a cabo no solo para cumplir la solicitud de financiación pública de la investigación, algo ya habitual tanto en EE.UU. como en la Unión Europea (Redkina, 2019), sino para contribuir a una ciencia cada vez más abierta. Si estos requisitos no se cumplen del todo, disponiéndose estos conjuntos de datos de cualquier forma y formato (como ocurre con otros tipos documentales habituales de los repositorios institucionales), no se terminará de alcanzar uno de sus principales propósitos: la reutilización.

La responsabilidad de la gestión de datos recae, inicialmente, en los investigadores, si bien parece claro que van a necesitar apoyo para adaptarse a los nuevos requisitos específicos de esta gestión (Marín-Arraiza et al., 2019). En este punto en particular, las bibliotecas académicas tienen una nueva oportunidad (y un reto) de mejorar sus actuales servicios de apoyo a la investigación y aumentar su presencia y relevancia en el seno de las instituciones universitarias y de investigación (Angelozzi, S. M., 2020; Sheikh et al., 2023). En algunos casos se han constituido consorcios regionales, como en Cataluña y Madrid, para "acometer de forma colectiva los retos derivados de la Ciencia Abierta y procurar que su adopción se realice con el menor esfuerzo por parte de las universidades" (Alcalá y Anglada, 2019). Por ello, las políticas de acceso abierto de I+D+i que se están implantando han de indicar de manera explícita los procedimientos, estándares, formatos, licencias y lenguajes a adoptar para una normalización e identificación de los datos de investigación (EC, 2023). También es importante la revisión de los propios contenedores de los conjuntos de datos con el objeto de que estén en condiciones de ofrecer una respuesta útil en su identificación, descripción, catalogación clasificación y métricas de uso.

La integración de este tipo de contenidos no termina de llevarse a cabo en la gestión editorial de las revistas científicas donde se publican los artículos, donde apenas tienen cabida los conjuntos de datos. Esto ocurre a pesar de las iniciativas de grandes editoriales que implementan buscadores especializados en repositorios de datos, como Data Citation Index de Web of Science Group, Mendeley Data de Elsevier y Google Data Search. Este interés evidencia la dimensión del volumen de los conjuntos de datos de investigación generados (preceptiva o voluntariamente) y muestra claramente la necesidad de su localización y curación, como es el caso del proyecto

Data Curation Network (Johnston et al., 2017; Johnston et al., 2018).

Es innegable que los conjuntos de datos son una fuente de información en continuo crecimiento. Alcanzar su efectivo desarrollo es uno de los retos a los que se enfrenta el movimiento hacia la Ciencia Abierta (UNESCO, 2021; Bethencourt-Aguilar, 2022). El momento actual parece el más adecuado para reconsiderar el proceso de creación y mantenimiento, ayudando a establecer líneas y estrategias de gestión y depósito institucional que formen parte del signo distintivo de calidad científica propio de las instituciones de investigación. En España, la mayor parte de la investigación se lleva a cabo en universidades y centros públicos de i+d+i (Fundación CYD, 2023). Son estas instituciones las que deben procurar implementar estos repositorios de conjuntos de datos siguiendo las mejores prácticas posibles. Otra tarea importante es hacer partícipes a sus comunidades investigadoras de los beneficios de disponer en abierto los datos de sus investigaciones, especialmente por la transparencia de los trabajos de investigación (De Giusti, 2020). Este punto en partícular ha cobrado una especial relevancia en los últimos meses por la proliferación de malas prácticas entre la comunidad científica, especialmente aquellas relacionadas con una hiperproducción científica incompatible con la calidad y la excelencia que se le presupone a la investigación.

Una vez expuesto el objeto de estudio, procede establecer el objetivo: analizar los repositorios de datos científicos de las universidades públicas españolas y establecer un diagnóstico de su nivel de implantación.

2. Metodología

Directamente relacionadas con la gestión de repositorios y datos de investigación están las organizaciones REBIUN y FECYT. En cuanto al total de universidades, en España, según los datos de Registro de Universidades, Centros y Títulos hay 86.

2.1. Universidades

De este total de universidades, 50 son públicas (47 presenciales, 1 no presencial y 2 especiales) y 36 privadas (31 presenciales y 5 no presenciales). La UNED es la universidad pública a distancia de gran implantación en todo el país y la Oberta de Catalunya, una universidad, pública, también de enseñanza a distancia, perteneciente al sistema universitario catalán y de gestión privada (si bien la consideramos como pública en el estudio). Las universidades "especiales" son la Menéndez Pelayo y la Internacional de Andalucía, especializadas en la organización de cursos de verano y conferencias por lo que no desarrollan investigación y no van a ser objeto de estudio.

El total de universidades a estudiar lo van a formar las 47 universidades presenciales más las dos universidades a distancia citadas, en total serán 49 las instituciones a analizar.

Las universidades de la comunidad de Madrid (menos la Complutense a nivel de conjuntos de datos) y la UNED forman el Consorcio Madroño para poner en marcha el repositorio colectivo e-Ciencia-Datos (2). Algo parecido ocurre en Cataluña, donde existe el repositorio CORA.RDR (3). La Universidad de las Islas Baleares también participa en el consorcio catalán (si bien no había publicado conjunto de datos alguno en el momento de la recogida de datos). Así, este segundo consorcio queda formando por ocho universidades.

2.2. Organizaciones

REBIUN es la red que reúne a las bibliotecas universitarias. Es un grupo de trabajo de CRUE (Conferencia de Rectores de las Universidades Españolas). Es la referencia para el desarrollo de los repositorios institucionales de las universidades españolas. Dentro de esta organización existe el grupo de trabajo de repositorios que tiene como objetivo la gestión de los datos de investigación en las universidades españolas. Este grupo es autor de una memoria de buenas prácticas de los servicios ofrecidos (2018) con el objetivo de presentar posibles soluciones para un mismo servicio, ofreciendo modelos a seguir por otras instituciones, recogiendo además ejemplos de iniciativas llevadas a cabo.

Casi todas estas buenas prácticas tienen que ver con las políticas y estrategias seguidas para concienciar a los investigadores de la importancia y necesidad de elaborar proyectos de gestión de datos y considerar casos de buenos ejemplos a seguir. En cuanto a los aspectos tecnológicos, este grupo ha trabajado en un esquema común de metadatos basado en Dublin Core (2017) verificando que "no existen muchos ejemplos de catálogo específico de datos institucional; algunas universidades han resuelto el depósito de datos en su propio repositorio institucional". Esto, como veremos en la discusión de resultados, es la causa de que casi todas las universidades hayan implementado sus repositorios de datos de investigación con la plataforma DSpace (4).

FECYT es la agencia del ministerio de Ciencia e Innovación que sirve de apoyo a universidades, centros de investigación y otras instituciones para el desarrollo del i+d+i. Su misión principal es ser catalizadora de la relación entre la ciencia y

la sociedad, impulsando el crecimiento de la cultura científica española y fomentando la transferencia de conocimientos. En su organigrama, cuenta con la Unidad de Acceso Abierto que aboga por la eliminación de las barreras que impiden el acceso a los resultados de la investigación científica, mayoritariamente financiada con fondos públicos. Esta unidad es uno de los pilares sobre los que asienta la transición hacia la Ciencia Abierta (UNESCO, 2021) en España y es la redactora de una guía para la evaluación de los repositorios institucionales de investigación que recoge un conjunto de ítems para medir la calidad de los metadatos y un conjunto de aspectos a evaluar de su interoperabilidad (FECYT, 2021). Esta unidad, además, ha tenido un papel decisivo en la elaboración de la Estrategia Nacional de Ciencia Abierta de España (2023).

Dato	Propósito
Tipo	Si se trata de una universidad individual o de un conjunto de universidades parte de un consorcio la que implementa el portal de datos de investigación.
URL consulta datasets	Dirección del portal en la que se pueden consultar los conjuntos de datos depositados.
URL informativa	Dirección donde se encuentra ayuda y documentación sobre el portal de datos de investigación.
Modalidad	Si los datos de investigación se depositan en un portal específicamente implementado para ello o si se depositan dentro del repositorio institucional.
Software	La aplicación informática con la que se ha implementado el portal de datos de investigación.
#datasets	El número de conjuntos de datos depositados.
Archivo	Si el depósito de los conjuntos de datos lo hacen directamente los investigadores o se delega en técnicos de la universidad.
Notas	Documentos de ayuda y guía al usuario. Tutoriales. Infografías. Acceso a herramientas informáticas para la gestión de los datos de investigación.
Fecha/as de recogida de datos	Se han recopilado los datos, en una primera revisión del 6 al 12 de marzo de 2023. Posteriormente, se han revisado datos y anotaciones hasta el día 25 de marzo de 2023.

Tabla I. Recopilación preliminar de información sobre el total de universidades públicas españolas

2.3. Recogida inicial de información

En una primera instancia se ha trabajado sobre las 49 universidades indicadas anteriormente, las que forman un consorcio (15 en total) y las que han dispuesto repositorios de datos de investigación de forma individual (34), tomándose los datos que se muestran en la Tabla I (a la izquierda).

La revisión inicial de los repositorios de conjuntos de datos de investigación muestra, en general, un nivel de implementación muy diverso, oscilando desde instalaciones mínimas hasta la implementación de repositorios colectivos de consorcios de instituciones de educación superior. El número de universidades con menos de una decena de conjuntos de datos depositados asciende a 16, lo que lleva a pensar, con cierta base, que la gestión para el depósito de estos conjuntos de datos se encuentra aún en una fase embrionaria. Este conjunto lo forman: Cantabria, Córdoba, Huelva, Jaén León, Oviedo, Pablo de Olavide, Politécnica de Cartagena, Castilla-La Mancha, Extremadura, La Rioja, Miguel Hernández, Pública de Navarra, Santiago y Vigo. Las siete últimas no disponían de ningún conjunto de datos publicado cuando se llevó a cabo el análisis de los repositorios, así que se excluyeron del estudio y se redujo el total de universidades analizadas de forma individual a 25.

2.4. Determinación de la muestra de estudio

Se han tomado las siguientes decisiones para formar la muestra objeto de estudio: se analizan los consorcios como si de una única institución se tratara, se excluyen del estudio aquellas universidades con cero conjuntos de datos depositados, y se excluyen del estudio a aquellas universidades que no implementen su repositorio de datos de investigación con software libre. De esta forma, la muestra queda determinada por el conjunto de universidades recogido en la *Tabla II.*

Nombre	#univ	Composición	#datasets
CORA	8	Autónoma de Barcelona; Barcelona; Girona; Lleida; Oberta de Catalunya; Politécnica de Catalunya; Pompeu Fabra; Rovira i Virgili e Illes Balears	467
e-Ciencia-Datos	6	Alcalá de Henares; Autónoma de Madrid; Carlos III, UNED; Politécnica de Madrid y Rey Juan Carlos	744
Resto	25	A Coruña; Alicante; Almería; Burgos, Cádiz, Cantabria, Complutense, Córdoba, Granada, Huelva, Jaén, Jaume I, León, Málaga, Murcia, Oviedo, Pablo de Olavide, País Vasco, Politécnica de Cartagena, Politécnica de Valencia, Salamanca, Sevilla, Valencia, Valladolid y Zaragoza	744

Tabla II. Composición de la muestra objeto de estudio y número de datasets

3. Discusión de resultados

A partir de los aspectos analizados en la revisión inicial de los repositorios de datos de investigación de las universidades que forman parte de la muestra, se obtienen los siguientes resultados:

3.1. Tipo de organización

Son dos los consorcios y 24 las universidades que han implementado repositorios de datos de investigación con, al menos, un conjunto de datos publicado. Los consorcios agrupan 14 universidades con conjuntos de datos publicados.

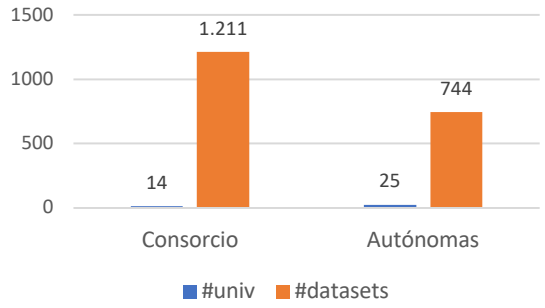

Figura 1. Total de repositorios y conjuntos de datos por tipo de organización (consorcio(/universidad)

Las diez universidades que más conjuntos de datos tienen publicados, en sus repositorios institucionales o en uno colectivo, son las siguientes:

Universidad	#datasets
Alcalá de Henares	375
Carlos III	259
Zaragoza	222
Granada	112
Barcelona	105
Politécnica de Catalunya	75
Lleida	59
Pompeu Fabra	44
Politécnica de Valencia	43
Politécnica de Madrid	41

Tabla III. Las 10 primeras universidades españolas por número de conjuntos de datos publicados

Estas diez universidades totalizan 1.334 conjuntos de datos, el 68,3% del total de conjuntos publicados por las universidades públicas españolas (5). Las otras 28 universidades analizadas publican un poco más del 30%. Existe, pues, una fuerte concentración en las universidades de la tabla anterior, mucho más manifiesta entre las que superan los 100 conjuntos publicados (cinco). Destaca el hecho de que las tres universidades politécnicas aparecen en la Tabla III. También es interesante el dato de que siete de estas universidades formen parte de consorcio (tres del madrileño y cuatro del catalán).

3.2. Año de publicación

La distribución de conjuntos de datos publicados año por año se ve en el siguiente gráfico.

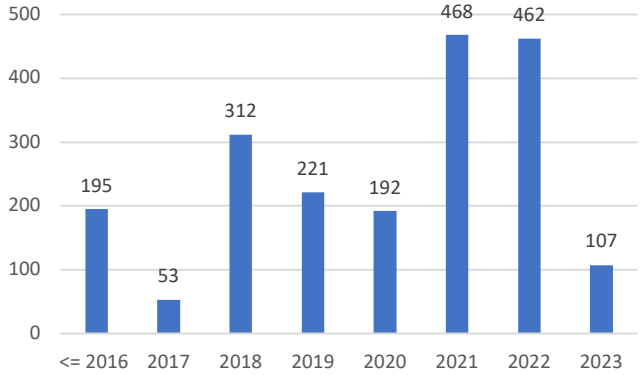

Figura 2. Conjuntos de datos por año de publicación

Los años 2021 y 2022 son los que se han publicado más conjuntos de datos en las universidades españolas, si bien ese dato está muy influenciado porque la Universidad de Zaragoza solo ha publicado en esos dos años (81 y 141 conjuntos de datos, todos ellos sobre la calidad del aire, lo que también influye en el análisis de las materias). Se ha incluido el total de conjuntos de datos de 2023 que, obviamente será mucho mayor cuando finalice el año en curso ya que muestra una proyección, de seguir el ritmo actual, de una cifra similar a la de los dos últimos años

Llama la atención el alto número de conjuntos de datos publicados en el año 2016 o antes. Esto se debe a un caso particular, el de la Universidad de Sevilla, en la que dieron de alta ese año un total de 124 conjuntos de datos de arquitectura y urbanismo del Protectorado Español en Marruecos, fondos documentales donados por el Ejército Español. El conjunto de datos se crea en esa fecha, si bien la información que contienen es de fechas muy anteriores (entre los años 1930 y 1950). En todo caso, queda pendiente confirmar. También destacan los conjuntos de datos de investigaciones bibliométricas de la Universidad de Granada, la mayor parte de ellas realizadas antes de ese año. Es previsible una tendencia positiva en la publicación de conjuntos de datos este año.

3.3. Materia

Se han revisado los conjuntos de datos publicados por las universidades y consorcios de la muestra y se han asignado a las cinco ramas tradicionales de investigación en el sistema universitario español: artes, salud, ciencias, sociales e ingenierías. Los resultados se visualizan en el siguiente gráfico:

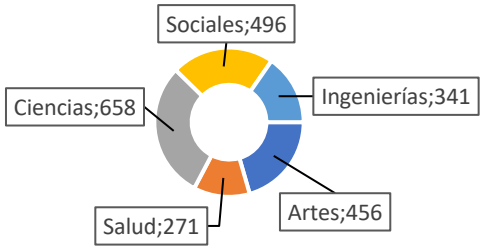

Figura 3. *Conjuntos de datos por año de publicación*

El 30% de los conjuntos de datos publicados son de ciencias, seguidos de ciencias sociales y artes y humanidades. En ciencias se nota la influencia de la Universidad de Zaragoza donde todos los conjuntos de datos son de la misma temática (medio ambiente) y están publicados solo en dos años. En artes y humanidades casi el 75% lo aporta el consorcio Madroño (Universidad de Alcalá de Henares). Recordemos también el caso de la Universidad de Sevilla y los datos de arquitectura y urbanismo de ciudades norteafricanas, casi un tercio del total. En ciencias sociales destaca la aportación del consorcio madrileño (269 conjuntos de datos) y, en particular, de la Universidad de Granada con un alto número de conjuntos de datos bibliométricos publicados por el grupo de investigación EC3 (la mayoría antes del año 2016).

3.4. Plataforma software

Los dos consorcios han implementado sus repositorios usando la plataforma Dataverse. Por regla general, las universidades que han acometido el desarrollo del repositorio de datos de investigación por su cuenta emplean el software DSpace. Una de ellas, las Palmas de Gran Canaria, emplea DSpace-CRIS, la distribución de DSpace específica para la gestión de datos de investigación (aunque solo se habían depositado dos conjuntos en el momento del estudio). Cuando se usa DSpace, el repositorio de los conjuntos de datos suele ser una colección específica del repositorio institucional. Algunas universidades están ahora migrando a la versión 7 de este sistema que aporta varias prestaciones para la gestión de conjuntos de datos de investigación.

La Universidad de A Coruña no ha implementado su repositorio, sino que redirige a los investigadores a Zenodo, el repositorio de acceso abierto de propósito general desarrollado por el CERN dentro de una iniciativa de la Comisión Europea. La Universidad de Zaragoza emplea la plataforma Invenio RDM para implementar su repositorio que se da la circunstancia de que este software es una distribución de Zenodo. Por último, la otra universidad canaria, La Laguna, emplea tecnología de pago (Digital Commons de Elsevier), por lo que ha quedado excluida del estudio.

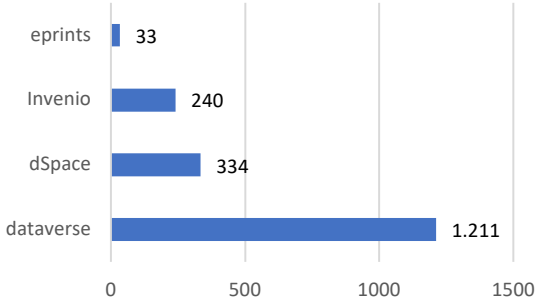

Figura 4. Conjuntos de datos distribuidos según la plataforma software en la que se han implementado

3.5. Archivo delegado o autoarchivo

Revisando las páginas de información de los repositorios de datos de investigación (también las de los repositorios institucionales cuando en las primeras no aparecía esa información), se han identificado tres modalidades: autoarchivo, archivo delegado en el personal del repositorio y ambas modalidades al mismo tiempo. No ha sido sencillo encontrar esta información y en algún caso ha resultado imposible. En el caso de los consorcios el archivo es delegado (es uno de los servicios ofrecidos). En el caso de las universidades que gestionan de forma individual el repositorio, 12 de ellas optan por el autoarchivo, 1 por el archivo delegado y 3 por la combinación de ambos. El predominio del autoarchivo es fruto, sin duda alguna, de la costumbre heredada del depósito en los repositorios institucionales En algunas universidades se informa a los investigadores de posibles futuras revisiones del depósito para verificar la calidad del mismo.

3.6. Ayuda y orientación

La apuesta decidida de Rebiun y FECYT por la Ciencia Abierta sí parece haber tenido éxito en las bibliotecas universitarias en lo relacionado con la ayuda y orientación a la comunidad investigadora por medio de biblioguías, infografías, información sobre los planes de gestión de datos,

aplicaciones para desarrollarlos, plantillas para hacerlo manualmente y aprendizaje de cómo deben ser citados.

Tras la revisión del total de universidades objeto del estudio, podemos afirmar que 36 de 39 orientan y ayudan de forma suficiente a su comunidad. Es previsible que las tres universidades restantes solucionen este problema pronto. Los temas más presentes en estos contenidos son los siguientes:

- La gestión de los conjuntos de datos de investigación. En qué consiste y cuál es su finalidad.
- Los planes de gestión de datos, qué son y para qué sirven. Guías para la elaboración de estos planes.
- Información sobre aplicaciones para la elaboración de los planes de gestión de datos. Se recomienda DMP Online, ARGOS y PGD Online (aplicación del consorcio madrileño).
- Orientación sobre dónde publicar los conjuntos de datos. Se citan los principales repositorios generalistas, en especial Zenodo y el repositorio institucional de cada universidad (con algo menos de frecuencia, parece darse por sentado que los investigadores saben que disponen de ese servicio).
- En algunas universidades se informa de la importancia del uso de licencias de libre acceso a los contenidos (Creative Commons generalmente).

4. Análisis y descripción de las plataformas software empleadas

En la revisión de los repositorios de datos de investigación en funcionamiento se han identificado dos plataformas software ampliamente utilizadas (Dataverse y DSpace) y otra tres de menor frecuencia de uso (que en realidad son dos porque Invenio RDM y/o Zenodo son distintas implementaciones del mismo software).

4.1. Dataverse

Es una aplicación web de código abierto para compartir, preservar, citar, explorar y analizar datos de investigación. Facilita poner los datos a disposición de otros y permite replicar el trabajo de otros más fácilmente. Este proyecto está liderado por el Instituto de Ciencias Sociales Cuantitativas de la Universidad de Harvard y se basa en la experiencia de un anterior proyecto del Centro Virtual de Datos que se desarrolló entre 1997 y 2006 como una colaboración entre centro de datos Harvard-MIT (ahora parte de este instituto) y la Biblioteca de la Universidad de Harvard. Cada colección contiene conjuntos de datos, y cada conjunto de datos contiene metadatos descriptivos y archivos de datos (incluida la documentación y el código que acompañan a los datos). Como método de organización, estas colecciones también pueden contener otras colecciones (6).

4.2. DSpace

Se trata de una aplicación de repositorio web que permite a investigadores y académicos publicar documentos y datos. Aunque comparte algunas características con los sistemas de gestión documental, este software responde a una necesidad específica como sistema de archivos digitales, centrado en el almacenamiento a largo plazo, el acceso y la preservación digital. Este sistema ha tenido muy buena acogida en el seno de las organizaciones académicas para crear sus proyectos de repositorio digital de acceso abierto. Es gratuito, fácil de instalar y personalizable. Un punto fuerte es su amplia comunidad de desarrolladores en constante crecimiento, comprometida con la ampliación y mejora continuas.

La primera versión pública de DSpace se publicó en noviembre de 2002, en un esfuerzo conjunto de desarrolladores del MIT y HP Labs que terminaron poniendo en marcha de forma conjunta la DSpace Foundation. Actualmente el trabajo de la comunidad y el mantenimiento del software está liderado por LYRASIS, entidad que procede de la antigua fundación (7).

4.3. eprints

Es una plataforma de repositorios digitales de código abierto. Desarrollada en la Universidad de Southampton en el año 2000. Es software gratuito y de código abierto para crear repositorios de acceso abierto que cumplan el protocolo OAI-PMH. Comparte muchas características con los sistemas de gestión de documentos, pero su uso principal es de repositorio institucional y portal de revistas científicas (8).

4.4. InvenioRDM y Zenodo

Éste es un caso particular porque, en esencia, es el mismo software con distintas distribuciones. El repositorio de datos de investigación Zenodo, desarrollado por el CERN dentro del programa OpenAIRE (9), se ejecuta básicamente bajo la versión 3 de Invenio “envuelta” por una pequeña capa extra de código (que también se llama Zenodo). Para simplificar la reutilización de la base de código Zenodo, varias instituciones se unieron en 2019 para distribuir un paquete agnóstico para instituciones bajo el nombre de InvenioRDM, un marco de desarrollo de código

abierto para repositorios digitales a gran escala centrado específicamente en la seguridad y en la conservación a largo plazo (10).

Subgrupo FAIR	*Significado*
F: Localizable	C1 ¿Se puede acceder a un determinado conjunto de datos de investigación en una versión actual a través de un PID/DOI único?
	C2 ¿Está la información de los datos de investigación depositados indexada en catálogos de datos, registros y motores de búsqueda?
	C3 ¿Existe una interfaz de búsqueda con posibilidades de filtrado para datos enlazados estructurados?
A: Accesible	C4 ¿Pueden almacenarse o referenciarse fácilmente nuevos datos de investigación?
	C5 ¿Es la interfaz de entrada de los datos fácil de usar ocultando términos técnicos e identificadores?
	C6 ¿Se puede acceder directamente a los datos y/o metadatos de la investigación vía http(s)?
	C7 ¿Existen parámetros de autenticación y autorización para el acceso de los usuarios?
I: Interoperable	C8 ¿Está disponible la descripción de los metadatos en una serialización RDF?
	C9 ¿Pueden utilizarse determinadas ontologías establecidas para describir el conjunto de datos de investigación de forma general (schema.org/Dataset, DataCite o DCAT/DublinCore)?
	C10 ¿Pueden utilizarse vocabularios específicos de un dominio para describir con más detalle los datos de investigación?
	C11 ¿Puede describirse cada concepto relacionado con el conjunto de datos de investigación con su correspondiente URI?
R: Reutilizable	C12 ¿Puede especificarse una licencia de datos en forma de Linked Data?
	C13 ¿Puede especificarse y actualizarse la procedencia de los datos de forma estructurada?
	C14 ¿Se relacionan los conjuntos de datos basándose en Linked Data y en criterios como tema, comunidad, métodos utilizados o similares?
	C15 ¿Se validan los datos proporcionados o existen controles de conformidad?

Tabla IV. *Criterios de evaluación de las plataformas de gestión de conjuntos de datos basados en los principios FAIR (Langer et al., 2019)*

Langer et al. (2019) analizaron diversas plataformas para implementar repositorios. En su análisis establecieron tres categorías: *genéricas* para una gestión básica de los datos; *aplicaciones* como desarrollo de estas plataformas y *otras herramientas* para la gestión de datos. Las cuatro plataformas que se usan en las universidades españolas pertenecen a la primera categoría. Posteriormente investigaron sus capacidades para la aplicación de una gestión de datos interdisciplinaria sostenible. Para comparar las plataformas establecieron 15 criterios de evaluación a partir de los principios FAIR para la gestión de datos de investigación (Alcalá y Anglada, 2019), que se centran particularmente en el conocimiento y manejo de los enlaces de datos, más en concreto en el conocimiento y la gestión de los datos abiertos enlazados en estas soluciones de publicación de datos. La siguiente tabla recoge los criterios establecidos por Langer et al. (2019) en su análisis:

Langer et al. (2019), en su análisis, definieron seis posibles resultados para la evaluación del cumplimiento de cada criterio.

Indicador	*Significado*
+	El criterio se cumple íntegramente
O	El criterio se cumple parcialmente
-	El criterio no se cumple
%	El criterio no es aplicable
¿?	No fue posible evaluar ese criterio
()	La característica está parcialmente desarrollada en la versión original de la plataforma, pero puede ampliarse desarrollando código

Tabla V. Criterios de evaluación del cumplimiento de los principios FAIR por el software de repositorios digitales (Langer et al., 2019)

Plataforma	*F*	*A*	*I*	*R*
Dataverse	C1+ C2+ C3o	C4+ C5+ C6+ C7o	C8+ C9+ C11o	C12+ C13+ C14o
DSpace	C1+ C2+ C3o	C4o C5o C6+ C7+	C9+	C12o C13+ C14o
eprints	C1+ C2+ C3o	C4+ C6+ C7o		C12+ C13+ C14o
Invenio	C1 + C2 + C3 o	C4+ C6+	C9o C11o	C13o C15+

Tabla VI. Cumplimiento de los principios FAIR por parte de las cuatro plataformas de repositorio digital analizadas (Langer et al., 2019)

En la tabla anterior se recoge un resumen parcial de los resultados de cada plataforma. A efectos de este estudio, únicamente interesan los resultados positivos ("+"), parcialmente positivo ("o") y potencialmente positivo ")("), porque representan el cumplimiento de algún criterio FAIR. Dataverse es la plataforma que más principios FAIR satisface (9 de forma íntegra y 4 parcialmente) y se puede mejorar en los criterios C11 y C14 desarrollando software adicional. El único criterio que no se cumple, C10, tiene que ver con el posible uso de un vocabulario específico de un dominio para la descripción de los conjuntos de datos, criterio que no cumple ninguna de las cuatro plataformas en realidad. En el caso de las otras plataformas no se aprecian diferencias significativas entre ellas.

5. Conclusiones

En cuanto a la implementación de los repositorios de conjuntos de datos de investigación en las universidades públicas españolas, estamos prácticamente al principio de su desarrollo. Algunas prácticamente no han empezado. El escaso número de conjuntos de datos publicados desvirtúa el análisis cuantitativo. Hay universidades donde solo se han publicado dos o tres conjuntos de datos. Los consorcios ayudan a visibilizar estas fuentes de información. Las comunidades autónomas con varias universidades (Andalucía, Castilla y León y la Comunidad Valenciana principalmente), deberían plantearse esta opción como una clara oportunidad de mejora.

El estado embrionario de esta implantación es una oportunidad para unificar criterios que eviten la multiplicidad de esfuerzos similares y la dispersión de los desarrollos e implementaciones, algo que ya ha ocurrido con los repositorios generalistas y que, difícilmente, tiene vuelta atrás.

El software de repositorio digital que mejor cumple con los principios FAIR es Dataverse. Es el que usan los consorcios, con lo cual, un porcentaje importante de universidades satisfacen esos principios de apertura e interoperabilidad. DSpace es el más instalado, aunque el total de conjuntos de datos gestionados por esta plataforma es menor que el gestionado por los consorcios (otra razón más para la búsqueda de la gestión colaborativa).

En casi todos los casos existen guías de orientación y ayuda a los investigadores sobre los planes generales de datos y cómo depositarlos por medio de "biblioguías" dentro de la web de la biblioteca universitaria.

El depósito de los conjuntos de datos se lleva a cabo, de forma mayoritaria, por medio del autoarchivo, revisado (en algunos casos) por el personal técnico de la biblioteca. Quizá esa revisión debería ser obligatoria para mejorar la descripción.

Resulta innegable que una gestión de los conjuntos de datos, acompañada de una campaña de sensibilización a favor de su uso entre la comunidad investigadora, reforzará el papel de las bibliotecas universitarias como servicio de apoyo a la investigación. Un problema típico es la posible confusión que puede tener un investigador que haya recibido financiación europea para su investigación. En principio, sus conjuntos de datos deberían ser depositados en Zenodo, con independencia de publicar una copia en su universidad. Lo cierto es que se debe reflexionar sobre si es conveniente y necesaria esa duplicación de espacio de almacenamiento (algo poco sostenible que puede terminar siendo un problema), si bien puede ser interesante a efectos de la recuperación de información. En España, el marco legal definido por la Ley de la Ciencia (2022) también impone la publicación, así que perfilar políticas claras y sostenibles en el tiempo es algo sobre lo que se debe comenzar a trabajar pronto.

Los repositorios de conjuntos de datos de investigación deberían ofrecer métricas en abierto que permitan establecer si se está cumpliendo el objetivo de la reutilización de los datos. Es una de las líneas de investigación más necesarias porque va a permitir medir si todos los esfuerzos que se están desarrollando en este momento ayudan al desarrollo de nuevas investigaciones (Barrett et al, 2021).

Notas

(1) CC son las siglas de las licencias 'Creative Commons' que tienen varias especificaciones: CC-BY permite copiar y redistribuir el material en cualquier medio o formato, remezclando y transformando el contenido para cualquier propósito, incluso comercialmente. CC0 es bastante parecida, se renuncia a todos los derechos a la obra bajo las leyes de derechos autorales en todo el mundo. Más información sobre estas licencias en https://creativecommons.org

(2) eCienciaDatos es el repositorio de datos de investigación del Consorcio Madroño. Más información en https://edatos.consorciomadrono.es/

(3) CORA.RDR es el repositorio de datos federado y multidisciplinar para la publicación de conjuntos de datos de investigación en modo FAIR, siguiendo las directrices de la European Open Science Cloud (EOSC). Más información en https://www.csuc.cat/es/servicios/cora-repositorio-de-datos-de-investigacion

(4) DSpace es un software de código abierto que provee herramientas para la administración de colecciones digitales, y comúnmente es usada como solución de repositorio bibliográfico institucional. Soporta una gran variedad de

datos, incluyendo libros, tesis, fotografías, filmes, video, datos de investigación y otras formas de contenido. Los datos son organizados como ítems que pertenecen a una colección; cada colección pertenece a una comunidad. Fue liberado bajo una licencia BSD en el 2002, como producto de una alianza de Hewlett Packard y el MIT.

(5) En el estudio se ha analizado el número de conjuntos de datos presentes en los repositorios institucionales. Puede ser que una universidad disponga de una colección en Zenodo o en otro repositorio, como es el caso de A Coruña, si bien solo hemos identificado esta universidad. El total de documentos no ha de variar mucho.

(6) La página del proyecto Dataverse está en la URL: https://Dataverse.org/

(7) Más información sobre DSpace en la URL: https://dspace.lyrasis.org/

(8) Más información sobre este proyecto en https://www.eprints.org/uk/

(9) OpenAIRE es una infraestructura tecnológica y de servicios para apoyar, acelerar y medir la correcta implementación de las políticas europeas de acceso abierto a publicaciones científicas y datos de investigación. Cuenta con una sólida red de agentes nacionales que actúan como puntos de referencia nacional para divulgar y difundir las políticas hacia la Ciencia Abierta auspiciadas por la Comisión Europea entre las instituciones y los investigadores, así como para facilitar la coordinación de las políticas nacionales con las europeas. El agente español es FECYT. Más información está disponible en https://recolecta.fecyt.es/open-aire

(10) La página web del paquete de código abierto InvenioRDM es https://inveniosoftware.org/

Referencias

Abadal, E.; et al. (2023). Ciencia abierta en España 2023: informe de situación y análisis de la percepción. http://hdl.handle.net/2445/200020

Alcalá, M.; Anglada, L. (2019). FAIR x FAIR: Requisitos factibles, alcanzables e implementables para un repositorio de datos de investigación FAIR. https://www.recercat.cat/bitstream/handle/2072/356460/InformeFxF_maquetada_ESP.pdf

Alonso-Arévalo, J. (2019). La gestión de datos de investigación en el horizonte de las bibliotecas universitarias y de investigación. // Cuadernos de Documentación Multimedia. 30, 75-88. https://doi.org/10.5209/CDMU.62806

Angelozzi, S. M. (2020). La gestión de datos de investigación en abierto: introducción al rol emergente para las bibliotecas universitarias y científicas argentinas. // Palabra clave. 9:2, e091. https://doi.org/10.24215/18539912e091

Ashiq, M.; Usmani, M. H.; Naeem, M. (2022). A systematic literature review on research data management practices and services. // Global Knowledge, Memory and Communication. 71:8/9, 649-671. https://doi.org/10.1108/GKMC-07-2020-0103

Ayris, P.; Ignat, T. (2018). Defining the role of libraries in the Open Science landscape: a reflection on current European practice. // Open Information Science. 2:1, 1-22. https://doi.org/10.1515/opis-2018-0001

Barrett; et al. (2021) Metrics for Data Repositories and Knowledge bases: Working Group Report. https://datascience.nih.gov/sites/default/files/Metrics-Report-2021-Sep15-508.pdf

Bethencourt-Aguilar, A.; Castellanos-Nieves, D.; Sosa-Alonso, J. J.; Area-Moreira, M. (2022). Implicaciones técnicas y prácticas de las Redes Adversarias Generativas a la Ciencia Abierta en Educación. // RiiTE Revista Interuniversitaria de Investigación en Tecnología Educativa, 138-156. https://doi.org/10.6018/riite.545881

Borghi, J.A.; Van Gulick, A.E. (2021). Promoting Open Science through research data management. // arXiv preprint https://arxiv.org/abs/2110.00888

De Giusti, M. R. (2021) Calidad en los repositorios digitales: los principios TRUST para repositorios de datos. // Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología. 29, 55-59 https://doi.org/10.24215/18509959.29.e6

España (2022). Ley 17/2022, de 5 de septiembre, por la que se modifica la Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación. // Boletín Oficial del Estado (2022-09-05). https://www.boe.es/eli/es/l/2022/09/05/17/con

España (2023). Estrategia Nacional de Ciencia Abierta (ENCA). https://www.ciencia.gob.es/InfoGeneralPortal/documento/c30b29d7-abac-4b31-9156-809927b5ee49

European Commission (2014). Guía del participante Horizonte 2020. https://www.horizonteeuropa.es/sites/default/files/inline-files/guia-del-participante-h2020.pdf

European Commission (2022). Guía del participante Horizonte Europa. https://www.horizonteeuropa.es/sites/default/files/noticias/Gu%C3%ADa%20del%20participante%20-%20Horizonte%20Europa%20web_0.pdf

European Commission (2023). Políticas de acceso abierto en América Latina, el Caribe y la Unión Europea: avances para un diálogo político. https://data.europa.eu/doi/10.2777/162

FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model. Specification and Guidelines (1.0). https://doi.org/10.15497/rda00050

FECYT (2021) Guía para la evaluación de repositorios institucionales de Investigación. https://www.fecyt.es/es/publicacion/guia-para-la-evaluacion-de-repositorios-institucionales-de-investigacion

Federer, L. M.; Qin, J. (2019). Beyond the data management plan: Expanding roles for librarians in data science and open science. // Proceedings of the Association for Information Sci-ence and Technology. 56:1, 529-531. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.82

Fundación CYD. (2023). La investigación en España: aumenta el gasto y la producción científica, pero con recursos inferiores a países de su entorno. Informe CYD. https://www.fundacioncyd.org/la-investigacion-en-espana-aumenta-el-gasto-y-la-produccion-cientifica-pero-con-recursos-inferiores-a-paises-de-su-entorno/

Johnston, L.; Carlson, J.; Hswe, P.; Hudson-Vitale, C.; Imker, H.; Kozlowski, W.; Olendorf, R. and Stewart, C. (2017). Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services. // Journal of eScience Librarianship. 6:1. e1102. https://doi.org/10.7191/jeslib.2017.1102

Johnston, L.; Carlson, J.; Hswe, P.; Hudson-Vitale, C.; Imker, H.; Kozlowski, W.; Olendorf, R.; Stewart, C.; Blake, M.; Herndon, J.; McGeary, T. and Hull, E. (2018). Data Curation Network: A Cross-Institutional Staffing Model for Curating Research Data. // International Journal of Digital Curation. 13:1. http://www.ijdc.net/article/view/616

Langer, A.; Bilz, E.; Gaedke, M. (2019). Analysis of Current RDM Applications for the Interdisciplinary Publication of Research Data. https://ceur-ws.org/Vol-2447/paper1.pdf

López Carreño, R.; Martínez Méndez, F. J. (2020). Sistemas de recuperación de información implementados a partir de CORD-19: herramientas clave en la gestión de la in-

formación sobre COVID-19. // Revista Española de Documentación Científica. 43:4. https://redc.revistas.csic.es/index.php/redc/article/view/1300

Marín-Arraiza, P.; Puerta-Díaz, M. y Gregorio-Vidotti, S. (2019). Gestión de datos de investigación y bibliotecas: preservando los nuevos bienes científicos. // Hipertext.net. 19, 13-31. https://raco.cat/index.php/Hipertext/article/view/360098

OECD/LEGAL/0347 (2021). Recommendation of the Council concerning Access to Research Data from Public Funding. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347

REBIUN (2017) Plantilla de metadatos para la descripción de datos de investigación. // Estudio de los repositorios de datos, 2017. http://hdl.handle.net/20.500.11967/137

REBIUN (2018) Gestión de datos de investigación en las universidades españolas y en el CSIC: memoria de buenas prácticas de los servicios ofrecidos. https://www.rebiun.org/sites/default/files/Gestion_Datos_Memoria_buenas_Practicas_201811.pdf

Rocca-Serra, P.; et al. (2023). The FAIR Cookbook: the essential resource for and by FAIR doers. // Scientific Data. 10, 292. https://doi.org/10.1038/s41597-023-02166-3

Sheikh, A.; Malik, A.; Adnan, R. (2023). Evolution of research data management in academic libraries: A review of the literature. // Information Development. 0:0. https://doi.org/10.1177/02666669231157405

Redkina, N. S. (2019). Current Trends in Research Data Management. // Scientific & Technical Information Processing. 46:2, 53-58. https://doi.org/10.3103/S0147688219020035

Torres-Salinas, D.; Robinson-Garcia, N.; Castillo-Valdivieso, P. A. (2020). Open Access and Altmetrics in the pandemic age: Forescast analysis on COVID-19 literature. // BioRxiv. 2020-04. https://doi.org/10.1101/2020.04.23.057307

UNESCO (2021) Recomendación de la UNESCO sobre la Ciencia Abierta. https://unesdoc.unesco.org/ark:/48223/pf0000379949_spa (2023-03-02)

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; y Mons, B. (2016). The FAIR Guid-ing Principles for scientific data management and stewardship. // Scientific Data. 3, 160018. https://doi.org/10.1038/sdata.2016.18.

Enviado: 2021-04-01. Segunda versión: 2021-07-25.
Aceptado: 2021-11-02.