ESTUDIOS E INFORMES

LÍNEA 2

2022

DATOS DE INVESTIGACIÓN EN LAS UNIVERSIDADES ESPAÑOLAS

crue Universidades Españolas

Red de Bibliotecas REBIUN

Colección Estudios e Informes, 2022

DATOS DE INVESTIGACIÓN EN LAS UNIVERSIDADES ESPAÑOLAS

Informe elaborado por:

Agnès Ponsati (CSIC)

Remedios Melero (CSIC)

Ainara Cisneros (UAH)

Concha Guijarro (UPNA)

Ángel Delgado (UPO)

Inma Ribes Llopes (UPV)

REBIUN / Plan Estratégico 2020-2023

Línea 2 Transformación digital y conocimiento abierto

Objetivo general 1. Impulsar el papel de las Bibliotecas en

el desarrollo de la Ciencia Abierta.

Objetivo específico 4. Promover la gestión de datos de investigación,

evaluar las infraestructuras y detectar buenas prácticas.

1. INTRODUCCCIÓN

Generar, compartir y reutilizar los datos de investigación es una realidad en la práctica totalidad de las disciplinas científicas. La adopción de las nuevas tecnologías para la creación y tratamiento de la información está modificando la manera de hacer ciencia y en ello juegan un papel muy relevante las nuevas metodologías y prácticas científicas y el uso responsable de los datos de investigación.

El depósito de los datos de investigación en acceso abierto siempre que sea posible, es una de las prioridades que, en el contexto de la ciencia abierta, se han marcado las entidades que financian proyectos de investigación, así como las propias universidades y organismos de investigación, tal y como se recoge en la Ley 17/2022 de la Ciencia, la Tecnología y la Innovación, y en los programas europeos Programa Horizonte 2020, Programa Horizonte Europa.

Con el fin de propiciar que los conjuntos de datos respondan a los principios FAIR (localizables, accesibles, interoperables y reutilizables), contribuir a la reproducibilidad de los resultados de investigación, incrementar la eficiencia y el avance del conocimiento, se han creado infraestructuras y dictado políticas para alcanzar estos objetivos.

El número y proporción de trabajos publicados y depositados en los repositorios institucionales crece cada año, el de los conjuntos de datos depositados parece no seguir el mismo patrón, al menos en el caso de los repositorios institucionales, y a menudo encontramos que la cantidad de conjuntos de datos depositados en abierto no guarda una relación proporcional al de las publicaciones científicas. Ante este escenario cabe preguntarse qué está pasando en relación a los mismos ¿depositan los investigadores los conjuntos de datos?, en caso afirmativo, ¿dónde lo hacen?

Con el objetivo de conocer dónde depositan los datos nuestros investigadores nos propusimos hacer un muestreo mediante la búsqueda de conjuntos de datos de investigación en las instituciones de distinto tamaño y características que participan en este grupo de trabajo: Consejo Superior de Investigaciones Científicas (CSIC), Universidad de Alcalá (UAH), Universidad Pública de Navarra (UPNA), Universidad Pablo de Olavide (UPO) y Universidad Politécnica de Valencia (UPV).

2. METODOLOGÍA DEL ESTUDIO

Para lograr este objetivo se ha establecido una metodología basada en la selección de un conjunto de recolectores de repositorios de datos y de repositorios de datos multidisciplinares. Se han incluido además los repositorios institucionales de las instituciones participantes. No se han tenido en cuenta como fuente de análisis los repositorios temáticos debido a su elevado número, a su dispersión y a la atomización que suponían.

Los recolectores y repositorios seleccionados son:

Recolectores:

- Data Monitor de Elsevier – Mendeley Data (en la parte gratuita) https://data.mendeley.com/
- Data Citation Index
- DataCite https://commons.datacite.org/
- Dimensions https://app.dimensions.ai/discover/publication

Repositorios:

- Figshare https://figshare.com/
- Zenodo https://zenodo.org/
- Dryad https://datadryad.org/stash
- Harvard Dataverse https://dataverse.harvard.edu/
- Repositorios institucionales de datos (Digital CSIC, e-cienciaDatos, Rio, Riunet)

Una de las universidades participantes contaba en el momento de realizar el estudio con la suscripción a la herramienta Data Citation Index lo que nos permitió hacer las búsquedas con una herramienta propietaria y estudiar qué tipo de datos ofrecía. Cada institución facilitó la ecuación de búsqueda de su afiliación institucional para realizar las búsquedas en Data Citation Index.

Para realizar las búsquedas hemos establecido cada institución nuestra estrategia de búsqueda para la afiliación institucional correspondiente con el fin de poder recuperar el máximo de información de estos repositorios.

1. De los resultados obtenidos en las búsquedas únicamente se tuvo en cuenta lo que hemos considerado como conjunto de datos, que son: "datasets, imágenes, videos, datos tabulares y datos de secuencias genéticas".
2. Cada institución recopiló sus datos en un documento Excel, con una hoja por cada recolector / repositorio con la siguiente estructura:

   RECOLECTORES/REPOSITORIOS-INSITUTCIÓN-FUENTE-Nº DE REGISTROS-
   FECHA TOMA DE DATOS-ECUACIÓN DE BÚSQUEDA

3. Se revisaron los datos obtenidos y se agruparon en un único fichero con los datos de todas las instituciones participantes en el estudio.

3. ANÁLISIS DE RESULTADOS

Limitaciones del estudio

Hay que tener en cuenta a la hora de analizar los datos obtenidos que no se ha realizado ningún ejercicio de eliminación de duplicados por lo que un mismo conjunto de datos puede estar contado más de una vez en las distintas fuentes (repositorios o recolectores).

También conviene señalar que, en los resultados obtenidos, no se han tenido en cuenta los resultados del repositorio Global Biodiversity Information Facility (GBIF), porque los resultados de la búsqueda devolvían un número muy elevado de registros que parecía no corresponderse con el número real de conjuntos de datos, sino más bien con el número de ítems disponibles en la fuente con una afiliación particular.

Todos los datos obtenidos se pueden consultar en el documento de Excel denominado "Datsets-globales.xls".

Análisis de datos extraídos en relación a datasets con afiliación CSIC

El CSIC cuenta con fecha 10/10/2022 con 12.470 conjuntos de datos en su repositorio institucional DIGITAL.CSIC.

De las búsquedas en las diferentes fuentes se obtuvieron los siguientes resultados:

1. De las consultas realizadas en los recolectores (DataMonitor-Mendeley, Data Citation Index, DataCite, Dimensions) se obtuvieron resultados dispares siendo las fuentes Data Citation Index y DataCite los que alcanzan un volumen de representación más elevado.
2. Si se analizan los datos en función de las fuentes recolectadas, los resultados procedentes de una misma fuente (Zenodo, Dryad, Figshare...etc) son dispares y no guardan relación numérica. Lo que hace sospechar sobre los mecanismos de recolección en lo que pueden influir muchos factores (curación de los metadatos recolectados, afiliaciones, coberturas, PID utilizados...etc.)

Sería muy interesante chequear los datos que devuelven las distintas fuentes analizadas para ver si los resultados obtenidos:

- ✓ Son descripción de ítems que ya existen en los IR objeto de estudio
- ✓ Son ítems que no están en los IR, pero sí en otros repositorios y pueden ser interoperables con los repositorios institucional, y permitir con ello su importación

El resultado de las consultas realizadas en los repositorios escogidos en la metodología de este estudio nos permite extraer las siguientes posibles conclusiones:

- El IR que da cuenta de más conjuntos de datos con afiliación CSIC es su propio IR DIGITAL.CSIC. Lo cual constata que el **Mandato de Acceso Abierto del CSIC** y la estrategia de carga de este tipo de resultados de investigación puesta en marcha por la Oficina Técnica de DIGITAL.CSIC, aun distando de ser comprensiva, está arrojando ya

resultados positivos, identificando al propio repositorio del CSIC como sitio de referencia para su carga.

- En segundo lugar, el repositorio institucional (IR) que alberga más datasets con potencial afiliación CSIC es Zenodo, lo que es consistente con los mandatos de H2020, Horizonte Europa y los pilotos previos a estos programas. Probablemente un alto porcentaje de estos datasets también estén en DIGITAL.CSIC.
- Los resultados obtenidos en repositorios multidisciplinares son casi residuales, lo que parece reforzar la idea de que los investigadores del CSIC escogen, cuando quieren depositar sus datos a DIGITAL.CSIC como su repositorio de preferencia.

Análisis de datos extraídos en relación a datasets con afiliación de la Universidad de Alcalá (UAH)

De las búsquedas en las diferentes fuentes se obtuvieron los siguientes resultados:

- Zenodo es uno de los repositorios que más datos de la UAH contiene, lo que es lógico por la obligatoriedad que supone la financiación con fondos europeos de los programas Horizonte 2020 y Horizonte Europa.
- El siguiente repositorio donde más datos se depositan es en el repositorio institucional de la UAH e-cienciaDatos.
- Se han depositado pocos datasets en otros repositorios multidisciplinares o temáticos.
- El análisis ha permitido localizar a investigadores de la UAH que no están depositando datasets en nuestro repositorio institucional e-cienciaDatos, sino en otros repositorios, de esta manera se puede contactar con ellos para informarles y recomendarles que los depositen en el repositorio institucional.
- Estos datos nos indican que el hábito de compartir los datos de investigación en repositorios seguros (*trusted*) en abierto no es una práctica generalizada entre los investigadores de la UAH. Este bajo nivel de depósito puede ser debido a numerosas causas, entre ellas la falta de una política institucional que monitorice el depósito de los datos, a la reticencia de los autores a compartir sus datos, y a una falta de formación en cuanto a qué, cómo y dónde poder hacer que sus datos sean FAIR (localizables, accesibles, interoperables y reutilizables).

Análisis de datos extraídos en relación a datasets con afiliación de la Universidad Pública de Navarra (UPNA)

De las búsquedas en las diferentes fuentes se localizaron pocos conjuntos de la UPNA, con las siguientes características:

- Los resultados de las consultas en cada uno de los recolectores de datos, en muchos casos, no guardan relación en cuanto a los repositorios en los que se localizan datos, ni en cuanto al volumen de datos en ellos depositados, en el caso de que coincidan los repositorios fuente.
- Se observa que los repositorios temáticos son los que recogen más datos, siendo IEEE Data Port (54 registros) y Gene Expression Omnibus (61 registros) los más utilizados.
- En cuanto a los datos en los repositorios multidisciplinares son poco utilizados, solo cabe resaltar Zenodo (9 registros) y Hardvard Dataverse (5 registros) .

- La localización de los datos en repositorios y recolectores es muy complicada, seguramente por falta de la afiliación de los autores, la no normalización de la misma, la mala calidad de los metadatos y no menos importante por los mecanismos de recolección de las propias herramientas.
- Esta escasa presencia de conjunto de datos de la UPNA evidencia la falta de una dinámica en el depósito de datos, lo cual pone de manifiesto que es necesario ofrecer a los investigadores más información, formación en los beneficios que esta gestión reporta, así como una mayor difusión de los mandatos de depósito, y de la necesidad de establecer un servicio de apoyo a la gestión de datos.

Análisis de datos extraídos en relación a datasets con afiliación de la Universidad Politécnica de Valencia (UPV)

De las búsquedas en las diferentes fuentes se obtuvieron los siguientes resultados:

Los datasets de los autores UPV son depositados principalmente en Zenodo (173 registros), y en segundo lugar en el repositorio institucional RiuNet (39 registros). El depósito en el resto de repositorios, FigShare, Dryad o Harvard Dataverse es meramente simbólico.

- En cuanto a agregadores/recolectores, la mayor presencia de datasets UPV se encuentra en Data Citation Index (512 registros), seguido del Data monitor Elsevier – Mendeley (366 registros). DataCite y Dimmensions apenas tienen presencia de registros UPV.
- Los registros encontrados evidencian de manera generalizada defectos en la firma institucional que dificultan la localización. Quizás por este motivo el agregador de mayor éxito es Data Citation Index que al realizar tareas de curación en los metadatos corrige parcialmente la falta de normalización en la firma científica, añadiendo afiliaciones.
- El mayor porcentaje de datasets UPV corresponde a autorías de los Institutos mixtos CSIC-UPV. En este sentido y atendiendo a la temática, genética concretamente, la mayor producción de datasets UPV es de autores CSIC-UPV y se encuentra en el repositorio GENE EXPRESSION OMNIBUS (https://www.ncbi.nlm.nih.gov/geo/)
- Se observa una escasa presencia de datasets en el repositorio institucional. La causa puede deberse a la libertad que deja la universidad en la elección del repositorio de destino en esta tipología documental. El control y monitorización de los datasets se está implantando a nivel de CRIS, se están creando infraestructuras y políticas para un apoyo y control interno, pero el depósito del dataset queda a la elección de los grupos de investigación.
- La escasa presencia de datasets de la UPV en repositorios y agregadores evidencian la necesidad de intensificar tanto la política como las infraestructuras y el servicio de apoyo a la gestión de datasets en la universidad.

Análisis de datos extraídos en relación a datasets con afiliación de la Universidad Pablo de Olavide

- El número de conjunto de datos recuperados es bastante escaso si se tiene en cuenta la producción científica de la institución. Por fuentes, es Zenodo, con 69, el repositorio que más conjuntos alberga, seguido de DRYAD. A mucha distancia Harvard Dataverse, el Repositorio Institucional Olavide y de manera casi testimonial Figshare.
- En cuanto a agregadores, es el Data Monitor de Elsevier el que devuelve un mayor número de resultados, 623, cifra muy superior al siguiente de los analizados, el Data Citation Index. Valores mucho más bajos ofrecen Datacite y Dimensions.

Las posibles causas de discrepancia encontradas en las cifras de las diferentes fuentes consultadas para la recolección de conjuntos de datos pueden deberse a los siguientes factores:

1. La propia fuente y los recursos que recolecta o consulta
2. Los metadatos utilizados para la recolección
3. La no normalización de firmas y afiliaciones
4. La cobertura temporal de las fuentes: periodo de recolección (actualización de datos)
5. El uso o no de IDs permanentes
6. Depósito por coautores externos a la institución

4. CONCLUSIONES

- La heterogeneidad de resultados obtenidos es quizás un reflejo de un escenario (el de los conjuntos de datos) que está todavía poco maduro entre la comunidad de investigadores y falto de un mayor nivel de estandarización. La descripción de esta tipología documental en los repositorios es todavía muy incipiente, la curación de los datos es pobre y poco FAIR. La definición de lo que es un "dataset" es ambivalente y muy variable según las disciplinas científicas. Por ello los recolectores están devolviendo resultados que pueden distorsionar la realidad por exceso o por defecto, dando una respuesta que no es veraz ni permite comparaciones entre fuentes, ni entre instituciones.
- El mundo del "dataset" es muy heterogéneo todavía y falta consolidar y asociar esta tipología de resultado de investigación como un elemento clave para garantizar la usabilidad y reproducibilidad de la investigación científica.
- Falta de una cultura en la gestión del dato científico entre la comunidad que dé respuesta a las siguientes cuestiones:
    - ¿Los investigadores no curan sus conjuntos de datos (no los describen y suben a un repositorio)?
    - ¿No lo hacen porque todavía falta de formación, de apoyo técnico?
    - ¿Las instituciones que deberían de ayudarles todavía no prestan la suficiente atención al tema facilitando recursos-infraestructuras-capacitación?
    - ¿No existen políticas claras a nivel institucional que ayuden a clarificar el qué y el cómo entorno a la gestión de los conjuntos de datos? (qué es un conjunto de datos, cómo se describe, estándares disciplinares, workflow de carga y curación de datos, licencias de uso, etc.)

- ¿Las instituciones ofrecen repositorios seguros y fiables (preservación) para la gestión de los datos, asegurando el FAIRness de los conjuntos de datos y su potencial integración en la ESOC?
  - ¿Hay suficiente capacitación para la correcta gestión de los datos de investigación? ¿Los profesionales de los servicios de apoyo a la investigación están preparados para prestar esta ayuda?
  - ¿La gestión de los datos de investigación recibe el reconocimiento adecuado en los sistemas de evaluación, como una buena práctica de ciencia abierta?

- También contribuye a esta foto tan difusa el problema de la limpieza de los metadatos de firma y afiliación institucional en repositorios y recolectores.
- Se percibe una débil vinculación entre las publicaciones (artículos) y los datasets relacionados. En muchos casos se echa en falta la presencia de un metadato que vincule ambos objetos.
- En casi todas las instituciones el repositorio multidisciplinar donde más datasets se depositan es Zenodo, lo que es comprensible por la obligatoriedad de la Comisión Europea de depositar en abierto los datos de los proyectos financiados con fondos europeos de los programas Horizonte 2020 y Horizonte Europa.
- El segundo repositorio con más datasets depositados es el repositorio institucional de las instituciones que lo poseen.

5. RECOMENDACIONES

1. Elaborar políticas institucionales acordes con las directivas/legislación, que incluyan, qué conjuntos de datos son susceptibles de depósito, dónde hacer el depósito, licencias de uso, etc., dirigidas a que los datos cumplan con los principios FAIR . La política debe incluir también un mecanismo de seguimiento o monitorización de su cumplimiento.

2. Dotar de infraestructuras sostenibles y recursos humanos para garantizar la correcta gestión, difusión y preservación de los datos de investigación generados por el personal de la institución.

3. Asesorar a los investigadores en el proceso de depósito de los datos en repositorios seguros (*trusted*), siempre indicando la afiliación correctamente, la correcta descripción de los metadatos, e insistir en la necesidad de relacionar los datos con las publicaciones.

4. Hacer una mayor difusión y formación sobre la gestión de datos de investigación que contribuya a una cultura sobre compartir, acceder y reutilizar datos de investigación y a una mayor concienciación de los investigadores sobre las ventajas de que sus datos sean FAIR (localizables, accesibles, interoperables y utilizables).

5. Implementar mecanismos de incentivos y recompensas por contribuir a la apertura de los datos de investigación como una buena práctica de ciencia abierta.

crue Universidades Españolas
Red de Bibliotecas REBIUN