# FAIR x FAIR

## Requisitos factibles, alcanzables e implementables para un repositorio de datos de investigación FAIR

Consorci de Serveis Universitaris de Catalunya (CSUC)

MARZO DE 2019

## Reconocimientos

Este documento ha sido elaborado juntamente con representantes de la Universitat de Barcelona, la Universitat Autònoma de Barcelona, la Universitat Politècnica de Catalunya, la Universitat Pompeu Fabra, la Universitat de Girona, la Universitat de Lleida, la Universitat Rovira i Virgili, la Universitat Oberta de Catalunya, la Universitat de Vic-Universitat Central de Catalunya y la Universitat Ramon Llull.

Agradecemos a todos los expertos citados en este documento que han dedicado parte de su tiempo en valorar y desarrollar este informe.

Redacción: Mireia Alcalá, Técnica de Recursos de Información del CSUC

Coordinación: Lluís Anglada, Director del Área de Ciencia Oberta del CSUC

# Sumario

# Resumen ejecutivo

En los últimos años, los datos recogidos, generados o utilizados a lo largo de los proyectos de investigación han recibido la atención de la comunidad científica y de los órganos gestores de la investigación. A nivel europeo, el programa marco Horizonte 2020 y su piloto Open Research Data Pilot han promovido el movimiento llamado "gestión de datos de investigación" (Research Data Management, RDM). Este concepto es un paraguas que engloba diferentes actividades relacionadas con la creación, organización, estructuración, almacenamiento, preservación y compartición de los datos.

Es por este motivo que los proyectos financiados por la Comisión Europea (CE) requieren que se elabore un plan de gestión de datos (Data Management Plan, DMP) y se depositen los datos en abierto siguiendo los principios FAIR (Findable, Accesible, Interoperable y Reusable) con el objetivo de aumentar la eficiencia y la transparencia de la investigación a través de una rápida difusión de los resultados y de facilitar su reutilización.

Las universidades de Catalunya iniciaron el servicio de apoyo a la gestión de datos de investigación en septiembre de 2016. Este servicio ha basculado alrededor de tres grandes ejes:

- Favorecer la confección de planes de gestión de datos,
- Informar sobre los repositorios donde se pueden hacer públicos los datos y, en algunos casos, modificar los repositorios institucionales para admitir datos, y
- Preparar materiales que permitan a las instituciones establecer su política de acceso abierto de los datos.

El servicio ofrecido actualmente tiene dos carencias:

- Poca demanda, seguramente debida a varios motivos, pero principalmente derivada de la misma novedad del tema, y
- No se dispone de la infraestructura que la Comisión Europea pide para publicar los datos de investigación en forma FAIR.

Por estas razones los vicerrectores de Investigación que forman la Comisión Funcional del área de Ciencia Abierta del CSUC (ACO) acordaron encargar un informe que determinara los requisitos funcionales razonables que debe tener un repositorio de datos para qué cumpla con los requisitos FAIR. El acuerdo (CF ACO, 02/11/17) dice: 'Iniciar los trabajos para elaborar una propuesta de requisitos funcionales para la creación de lo/s repositorio/s consorciado/s de datos de investigación'. Estos requisitos debían seguir las directrices establecidas en la declaración hecha por el European Open Science Cloud (EOSC, 2017).

Para hacer este informe se ha creado un grupo de trabajo específico, se han consultado expertos tanto en el ámbito del CSUC como de fuera, se han examinado experiencias similares (Bélgica flamenca, Finlandia, Países Bajos, Portugal y Suecia) y se han estudiado los principales documentos que se han publicado recientemente sobre este tema.

Las entrevistas con expertos se realizaron con la intención inicial de hacer una relación de requisitos funcionales. Desde las primeras reuniones y entrevistas se comprobó que estos

expertos consideraban que, adicionalmente a los requisitos técnicos, había dos aspectos fundamentales para definir un repositorio donde publicar datos de investigación: establecer los condicionantes de contexto del repositorio y desarrollar unas buenas prácticas asociadas a la definición de conjuntos de datos FAIR.

Los expertos señalaron reiteradamente que la gestión de datos de investigación y su publicación en forma abierta es un tema de largo recorrido, pero todavía incipiente. Esto conlleva que los requisitos que ahora se puedan fijar probablemente habrá que ampliarlos o modificarlos en los próximos años. Y, a pesar de que no se pueda hablar de requisitos consolidados, los técnicos consultados recomiendan crear de forma inmediata una infraestructura de datos de investigación para generar experiencia y buenas prácticas para su gestión.

Teniendo en cuenta que los repositorios de datos son una realidad aún no consolidada ni homogeneizada pueden encontrarse en Europa diferentes opciones que responden a decisiones de contexto diferentes. Este informe asume que las opciones recomendables para las universidades de Catalunya en relación con el repositorio de datos de investigación son:

- Realizarlo ahora, a pesar de que conlleve readaptarlo, para tener un lugar donde publicar datos y ganar experiencia con su gestión.
- Hacerlo en Catalunya, a pesar de que existe la posibilidad de publicar los datos en repositorios ya en funcionamiento en Europa, dado que los datos se consideran hoy estratégicos y para facilitar el cumplimiento de las medidas legales.
- Ceñirse a datos permanentes o finales, aunque los investigadores tienen también necesidad de gestionar ficheros de datos provisionales.
- Ceñirse a datos de disciplinas que no tienen repositorios de datos consolidados, teniendo en cuenta que, para los que tienen, la mejor opción es publicarlos allí.

Asimismo, los expertos insisten que una infraestructura donde publicar los datos de investigación es, por sí sola, insuficiente, y que la gestión de los datos de investigación requiere desarrollar una serie de buenas prácticas como las siguientes:

- Hacer curación de datos, es decir, documentarlos para que sean comprensibles y reutilizables por otros usuarios que los que los han generado.
- Crear protocolos y criterios de expurga y selección de los conjuntos de datos
- Fomentar el uso de formatos abiertos o no propietarios y, en todo lo que sea posible, migrar los formatos propietarios a formatos abiertos.
- Usar estándares, protocolos y vocabularios controlados ampliamente aceptados.

La parte principal de este informe se centra en describir los requisitos funcionales que se consideran mínimos para garantizar que la infraestructura creada cumpla los requisitos FAIR de la CE. Estos requisitos, a partir de trabajos previos de la comisión ACO, se han clasificado bajo los siguientes grupos:

- Identificadores persistentes
- Alta capacidad de almacenamiento
- Prestaciones medias-altas de preservación
- Interoperabilidad con otros sistemas
- Gestión de características especiales

Vistas las consideraciones anteriores, el documento termina con las recomendaciones finales siguientes:

1. Crear aquí y de forma inmediata un repositorio donde se puedan publicar los datos de investigación de forma FAIR y que permita desarrollar experiencia y buenas prácticas en la gestión de datos de investigación. Este repositorio:
   a) Debe cumplir los requisitos FAIR ya conocidos y los que establezca la CE en un futuro inmediato.
   b) Debe ofrecer prestaciones de valor añadido respecto a las opciones actuales, como, por ejemplo, asignación de DOIs, interoperabilidad con el Portal de la Recerca de Catalunya y de preservación).
   c) Debe hacerse con software ya existente, dado que hay disponibilidad, y de código libre.

2. Promover y facilitar la publicación de los datos de investigación en abierto con acciones en que la Universidad haga público el servicio ya existente de gestión de datos de investigación y que permite la confección de Planes de Gestión de Datos y asesora en la publicación de datos. La acción de difusión debería hacerse entre las diferentes unidades que vehiculan la investigación y con la participación de las oficinas y servicios de investigación.

3. Hacer formación sobre los conceptos de Ciencia Abierta, en general, y, concretamente, sobre gestión de datos de investigación. Siguiendo las directrices de la European Commission Expert Group on FAIR Data, esta formación debe incluir la totalidad de miembros de la comunidad universitaria, y para ser eficaz, debería de distinguir entre usuarios avanzados, investigadores jóvenes y personal de apoyo de las universidades.

# 1. Importancia creciente de los datos de investigación en abierto

La ciencia siempre se ha basado en la utilización de datos, pero hasta hace poco, estos datos de investigación eran difíciles de recoger, conservar, compartir y reutilizar. La capacidad de los ordenadores y las redes de comunicaciones de procesar, conservar y comunicar datos ha cambiado el proceso científico haciéndolo más eficiente, transparente y colaborativo. Tal como se señala en la reciente declaración de la CRUE (2019) sobre Ciencia Abierta, la actividad investigadora actual es intensiva en la generación, consumo y explotación de datos y su preservación y reutilización son clave en la investigación del siglo XXI.

Las bases de este movimiento se asentaron en 2004 en París cuando ministros de Ciencia y Tecnología de los países integrantes de la OCDE, juntamente con China, África del Sur, Israel y Rusia aprobaron la Declaration on Access to Research Data from Public Funding (OCDE, 2004).

La Comisión Europea (CE) es quien ha dado una dimensión global a estos cambios cuando en 2013 inició una consulta pública (European Commission, 2013) sobre el impacto de estos cambios mencionados con el nombre general de 'Ciencia Abierta'. Posteriormente, en 2016 y bajo la presidencia holandesa de la Unión Europea, se celebró en Ámsterdam el congreso "Open Science - From Vision to Action" donde se afirma, en su documento final, que los resultados de la investigación deben ser públicos y reutilizables y para ello es necesario que los conjuntos de datos de los proyectos de investigación tengan planes de gestión de datos y sigan los principios FAIR (ver anexo 1) que establece que los datos de la investigación deben ser 'FAIR: Findable, Accesible, Interoperable and Reusable' (Government of the Netherlands, 2016).

La CE, la gran agencia de financiación en Europa, ha ido preparando el terreno para que los investigadores comiencen a gestionar sus datos. Bajo el programa marco Horizonte 2020 (2014-2020), puso en marcha un piloto llamado EC Research Data Pilot (ORD Pilot). El objetivo era "mejorar y maximizar la reutilización de los datos de investigación generados por los proyectos Horizonte 2020 teniendo en cuenta la necesidad de equilibrar la apertura y la protección de la información científica, la comercialización y los derechos de propiedad intelectual, cuestiones de privacidad y seguridad, así como, cuestiones de gestión y conservación de los datos" (European Commission, 2016). Este piloto sólo afectaba a unas áreas concretas y requería elaborar un DMP durante los 6 primeros meses de la concesión del proyecto y publicar los datos en abierto (siempre que no existieran razones para mantenerlos cerrados).

En 2017, el piloto se extendió a todas las áreas de Horizonte 2020 y se popularizó el principio "as open as posible, as closed as necessary". En el próximo programa marco Horizon Europe (2021-2027) será obligatorio publicar los datos de forma abierta y elaborar un plan de gestión de datos. Estos datos deberán hacerse públicos bajo los principios FAIR y, en principio, deberán estar, por defecto, abiertos, aunque se admitirá, para casos determinados, la posibilidad que estén cerrados.

Diferentes instancias europeas (la misma CE, entidades financiadoras de la investigación, universidades y asociaciones científicas) están dando mucha importancia a estos cambios. Esto queda reflejado con la aprobación de planes nacionales para facilitar el logro de los objetivos de

la Ciencia Abierta como han sido los de Finlandia (2014), Eslovenia (2015), Portugal (2016), Países Bajos (2017), Francia (2018) y Serbia (2018). Todos estos planes tienen entre sus objetivos la apertura de los datos de investigación.

El objetivo de publicar en abierto los datos de investigación también se manifiesta en las declaraciones y hojas de ruta para la Ciencia Abierta que han hecho entidades como la European University Association (2018), la League of European Research Universities (2018), la Young European Research Universities (2018), el Europe's Research Library Network (2018) o la Conferencia de Rectores de las Universidades Españolas (2019).

Entre estas hojas de ruta, destacan las Open Science Policy Platform Recommendations (2018) donde se determina que una de las ocho prioridades de la European Open Science Agenda es poner los datos de investigación en abierto siguiendo los principios FAIR. Estas recomendaciones dicen que las entidades de financiación y de investigación deben dar reconocimiento a los datos FAIR provenientes de la investigación, al igual que se hace con las publicaciones. También que los planes de gestión de datos deben ser obligatorios para todos los proyectos (y explotables por máquina). Y, por último, que los datos financiados con fondos públicos deben ser FAIR y citables y, tan abiertos como sea posible y tan cerrados como sea necesario.

En conclusión, seguir estos principios FAIR es un eje troncal de todas estas declaraciones, hojas de ruta o planes nacionales.

## 2. Actuaciones consorciadas y metodología para la redacción de este informe

Cuando se constituyó el Área de Ciencia Abierta (ACO) del CSUC se aprobó un Plan de trabajo para el período 2017-2019 (Doc. CO17/01). Este plan planteaba dos objetivos en relación con los datos de investigación: asesorar a los investigadores en la confección de planes de gestión de datos y garantizar la existencia de repositorios para datos de investigación.

Mientras, el ya existente Grupo de Trabajo de Apoyo a la Investigación (GTSR) trabajaba para poner en funcionamiento el Servicio de Apoyo a la Gestión de Datos de Investigación (véase más adelante). Este Grupo preparó unas propuestas de trabajo referente a los datos de investigación abiertos (Doc. CO17/11) que fueron aprobadas por la Comisión Funcional de ACO. El documento respondía al objetivo del Plan General de Actuaciones del CSUC de: "analizar las diferentes posibilidades para crear un repositorio de datos consorciado y definir la propuesta".

La Comisión consideró que se tenían que determinar los requisitos funcionales del repositorio de datos que, por economías de escala, debería ser cooperativo, pero no necesariamente centralizado. El repositorio permitiría mejorar los servicios prestados y dar respuesta a necesidades que no quedaban cubiertas con la infraestructura actual. La comisión consideraba que:

- como los datos de investigación son estratégicos, deben ubicarse en una infraestructura propia,
- el repositorio debe ofrecer prestaciones cualitativamente mejores que las ofrecidas por los repositorios institucionales actuales,
- debe hacerse teniendo en cuenta las economías de escala en cuanto a costes y la especialización en los datos a almacenar, y
- debe permitir la interoperabilidad de los datos entre los diferentes elementos (repositorios institucionales, sistemas CRIS y portales de investigación).

Para ello se creó una comisión técnica que debía determinar qué requisitos funcionales razonables debería tener un repositorio de datos de investigación para cumplir con los requisitos FAIR. Los miembros de esta comisión (ver anexo 1) proceden de los diferentes servicios que están implicados en las universidades en el tratamiento de datos de investigación: las bibliotecas, las oficinas de investigación y los servicios TIC. Esta comisión se ha reunido en dos ocasiones.

Paralelamente, se han hecho entrevistas con 32 expertos en gestión de datos de investigación (ver anexo 2), se ha recogido información de la solución dada a este tema en lugares con características similares a las de las universidades de Catalunya como pueden ser: Bélgica flamenca, Finlandia, Países Bajos, Portugal y Suecia (ver anexo 3) y se han recopilado y estudiado diferentes informes técnicos relacionados con la materia (ver anexo 4).

# 3. El servicio de apoyo a la gestión de datos de investigación, hoy

A mediados de 2015, el GTSR se dedicó de manera casi exclusiva a la gestión de datos de investigación para ofrecer apoyo a aquellos proyectos financiados por el ORD Piloto en tres grandes ejes: los planes de gestión de datos, los repositorios de datos y la política de acceso abierto a los datos.

Fruto de las diferentes tareas colaborativas, en septiembre de 2016, las universidades catalanas iniciaron el servicio de apoyo a la gestión de datos de investigación. Dentro del catálogo de servicios encontramos:

- Ayuda para la confección de planes de gestión de datos de investigación
- Recomendaciones para seleccionar un repositorio para el depósito de datos de investigación
- Ampliación de prestaciones de los repositorios institucionales para depositar datos
- Elaboración de unos requisitos para un repositorio de datos consorciado
- Acciones de difusión y formación

En cuanto a los planes de gestión de datos, las universidades ofrecen apoyo mediante la herramienta Research Data Management Plan (https://dmp.csuc.cat) que permite crearlos para proyectos financiados en el marco del programa Horizonte 2020 y el European Research Council del CE. La herramienta permite crear, compartir y exportar FAIR DMPS y es una adaptación del DMPRoadmap, un software de código abierto que se distribuye bajo la licencia MIT y que ha sido desarrollado conjuntamente por Digital Curation Center (DCC) y la University of California Curation Center (UC3). El valor añadido son las diferentes informaciones (CSUC, 2016) de lo que debe contener el plan, con descripciones y ejemplos reales que han sido consensuados por el GTSR y que se mantienen conjuntamente pero que permite a cada institución personalizarlas según le convenga.

Respecto el depósito de los datos, se ofrecen diferentes soluciones. Primeramente, se asesora a los investigadores respecto plataformas donde publicar los datos. Estas recomendaciones (CSUC, 2017) muestran un abanico de opciones disponibles con respecto tanto a los repositorios de datos temáticos, como multidisciplinares. Posteriormente, algunas universidades han adaptado sus repositorios institucionales (CSUC, 2017) y los ofrecen como una posibilidad más dentro de este abanico para algunos casos. Esta opción, si bien es buena para empezar, no proporciona todas las prestaciones de un repositorio FAIR. Actualmente los trabajos se centran de forma importante en la definición de unos requisitos funcionales mínimos y factibles para un repositorio de datos para que sean FAIR. Finalmente, se realizan diferentes actividades de seguimiento (monitorización a través de diferentes indicadores), difusión (con infografías o vídeos cápsula, por ejemplo) y formación (al personal de apoyo a la investigación de las universidades).

En relación con la política de acceso abierto a los datos, los vicerrectores de investigación que forman la Comisión Funcional de ACO aprobaron como recomendaciones un documento donde se definían los elementos que una universidad debe tener en consideración para confeccionar una política en materia de gestión de datos. Los elementos de este documento - "Modelo de política de gestión de datos para una universidad" (CSUC, 2018)- son: la responsabilidad de la gestión de datos, el lugar y el periodo del depósito, la elaboración de un plan de gestión de datos, determinar la preservación y conservación de los datos, entre otros.

Para hacer prospección del servicio se han hecho dos encuestas en las universidades catalanas (CSUC, 2019) a investigadores con el objetivo de conocer sus necesidades en materia de gestión de datos de investigación. De manera clara, el servicio tiene dos puntos claramente débiles: un uso bajo y la no existencia de un repositorio propio donde hacer públicos los datos.

En relación con el primer punto, se observa una gran distancia entre lo que se ofrece y su demanda. La última encuesta revela que el 80% de las personas desconocen el servicio de apoyo a la gestión de datos de investigación que se da en su universidad. Esto seguramente se debe a diversas circunstancias como la misma novedad del tema, que la obligación de publicar los datos aún no afecta a los proyectos de investigación en curso, a que las acciones de difusión del servicio no han conseguido llegar a los investigadores que potencialmente la usarían, entre otros.

La segunda gran carencia es la falta de una infraestructura propia donde hacer públicas en forma abierta los datos de un proyecto de investigación. Para datos de investigación de algunas disciplinas científicas se pueden encontrar repositorios consolidados como Protein Data Bank para las estructuras 3D de las proteínas o el National Oceanographic Data Centre para datos oceanográficos, que son el lugar más adecuado donde publicar los datos. Para la mayoría de estos, sin embargo, la alternativa es depositar los datos en un repositorio generalista externo a su institución, en el de una institución participante en el proyecto de investigación que tenga un repositorio de datos o en el repositorio institucional propio. Con respecto a este último caso, los repositorios institucionales no tienen las características solicitadas (de identificadores, de preservación o de capacidad de almacenamiento, por ejemplo) para cumplir los requisitos que pide la CE para que los datos sean FAIR. La inexistencia de un repositorio de datos FAIR es, pues, un inhibidor a la publicación de datos.

# 4. Determinantes de contexto

Los expertos consultados consideran que los repositorios de datos son una realidad aún no consolidada ni homogeneizada. Esto contrasta con la importancia estratégica que se da a la publicación en abierto de los datos de investigación (ver apartado 1). La suma de estas dos consideraciones hace que hoy haya casi más documentos e informes sobre el tema que no realidades en funcionamiento. Estas, por otra parte, son muy diversas y a menudo incipientes. Todo esto hace que -a diferencia de en otros casos- no sea efectiva la opción de seleccionar las mejores prácticas y, sencillamente, copiarlas.

De las reuniones y entrevistas con expertos se han destacado una serie de decisiones que deberían tomarse de forma previa al establecimiento del repositorio de datos, ya que la elección de una opción u otra condiciona los requisitos que debe tener el repositorio.

Las disyuntivas que habría que considerar y entre las que hay que decidir son las siguientes:

### Hacer el repositorio ahora o esperar

- Los expertos coinciden en considerar que la publicación de datos de investigación en abierto es un tema aún incipiente y que todavía hay muchos aspectos a definir. Si fuera sólo por eso, habría que esperar estas concreciones para hacer un repositorio de datos, pero hay consenso también en que la publicación de datos de investigación en abierto requiere desarrollar protocolos y buenas prácticas y que esto sólo se puede hacer teniendo un repositorio en funcionamiento.
- Se asume, pues, que hay que hacer el repositorio ahora.

### Considerarlo una infraestructura en evolución o definitiva

- Siguiendo el argumento anterior, se asume que el repositorio deberá readaptarse a las exigencias y prácticas que se vayan consolidando a nivel internacional. Si bien todos los programas están en evolución, los que permiten hacer públicos datos de investigación lo están en un mayor grado, pero sólo se puede adquirir la experiencia necesaria teniendo uno.
- El repositorio que ahora se haga debe considerarse una infraestructura en evolución.

### Ceñirse a datos finales o incluir también datos provisionales

- El uso de datos en proyectos de investigación conlleva usar diferentes conjuntos de datos en diferentes momentos. Muchos de estos datos son provisionales o instrumentales y, si bien hay que usarlos para poder hacer la investigación, tienen sólo un valor temporal. Los investigadores tienen, pues, la necesidad de gestionar los datos de sus proyectos de investigación, además, de la de publicarlos; sin embargo, las necesidades derivadas de la gestión de datos difieren de los que ahora mismo se exigen que son de publicarlos.
- De cara a reducir las exigencias de los requisitos del repositorio y su coste se propone que el repositorio tenga, de momento, sólo función de publicación de datos finales.

### Tener el repositorio en Catalunya o usar los externos

- En este estadio aún incipiente de publicación de los datos de investigación en abierto, y de cara a dar una respuesta a las necesidades creadas por el programa Horizonte 2020, se han creado algunos repositorios que permiten publicar datos a cualquier investigador. Las prestaciones de estos repositorios son buenas y pueden evidentemente usarse para publicar los datos, pero existe la opinión generalizada que los datos de la investigación son un recurso estratégico que conviene tener en el propio sistema de investigación. Adicionalmente, debido a temas relacionados con el reglamento de protección de datos personales, legalmente la publicación de datos se ve facilitada si se hace dentro del territorio nacional.
- Por motivos estratégicos y legales el repositorio debería estar en Catalunya, a pesar que existe la posibilidad de publicar en repositorios ya en funcionamiento en Europa.

### Centrarse en disciplinas sin repositorios de datos consolidados o admitir de todas

- Cuando se habla de datos de investigación, los ejemplos que se dan a menudo son de disciplinas científicas que en los últimos años han experimentado grandes avances justamente por el uso de datos y por haberlos compartido. Un ejemplo claro es la Genómica. Hoy, algunas disciplinas han consolidado buenas prácticas e infraestructuras para publicar en abierto los datos de su investigación. En estos casos, lo mejor es consolidar los repositorios disciplinarios existentes. Asimismo, hay una 'cola larga' de especialidades que no están en estas circunstancias y que sí necesitan que se les cree un repositorio que puedan usar.
- Se considera que la mejor opción para el repositorio de datos es ceñirse a las disciplinas que no tienen uno de consolidado, teniendo en cuenta que, para las que lo tienen, la mejor opción es publicarlos allí.

En función de las consideraciones anteriores, y teniendo en cuenta que esto afecta el establecimiento de requisitos, **este informe recomienda hacer el repositorio ahora, si bien habrá que considerarlo una infraestructura en evolución, tenerlo aquí y destinarlo a la publicación de datos de investigación finales y centrar su uso en datos de disciplinas que no tienen un repositorio de datos temático consolidado**.

# 5. Requisitos funcionales mínimos y factibles

Para definir los requisitos funcionales (RF) mínimos y factibles que debe tener un repositorio de datos se han llevado a cabo diferentes entrevistas a 32 expertos en diferentes ámbitos que han permitido recoger aquellos requisitos que consideraban esenciales. A la vez, se ha tenido en cuenta diferentes artículos que tratan esta materia y de los cuales se destacan los siguientes:

- Amorim, RC (2017). A comparison of research data management platforms: architecture, flexible metadata and interoperability. Universal Access in the Information Society, 16 (4), 851-862. DOI: 10.1007 / s10209-016-0475-y
- Kim, S. (2018). Functional Requirements for Research Data Repositories. International Journal of Knowledge Content Development & Technology, 8 (1), 2-36. DOI: 10.5864 / IJKCT.2018.8.1.025.

Los diferentes requisitos recogidos anteriormente, se han clasificado según las categorías identificadas al documento de propuestas de trabajo referente a los datos de investigación abiertas "Repositorio(s) de Datos de Investigación Consorciado (RDC)" (CSUC, 2017) y que son las siguientes:

- identificadores persistentes,
- capacidad de almacenamiento de dimensiones elevadas y diferentes formatos,
- preservación de altas prestaciones,
- interoperabilidad entre los diferentes elementos y
- gestión de características especiales.

## 5.1 Identificadores persistentes

Para mejorar el acceso a los datos de investigación, es necesario dotarlos de identificadores persistentes (*Persistent Identifiers,* PID). Se trata de un "identificador construido e implementado de manera que el recurso identificado continúe siendo el mismo independientemente de la ubicación de su representación y también del hecho que diferentes copias puedan encontrarse en diferentes ubicaciones" (IASA-TC04, 2009).

### $RF_1$: Asignar el DOI como identificador

Existen diferentes tipos de PIDs pero, últimamente, el DOI ha aparecido como un estándar internacional emergente para los datos de investigación. El DOI (Digital Object Identifier) es una cadena alfanumérica única que identifica materiales publicados en línea. El uso del DOI aporta beneficios extras, según la ANDS (2018): garantiza la calidad y precisión de los datos, aumenta las citaciones, garantiza su persistencia de acceso a largo plazo, provee de un enlace fácil y equipara los conjuntos de datos al nivel de las publicaciones.

Actualmente, la organización sin ánimo de lucro DataCite es la principal entidad que proporciona DOIs para datos de investigación. Aun así, DataCite no asigna DOIs directamente, sino que esta actividad es asumida por los diferentes miembros de DataCite que actúan como agentes (ver anexo 5).

### $RF_2$: Soportar ORCID

Las universidades de Catalunya acordaron (CSUC, 2016) utilizar el sistema único y unívoco de identificador ORCID para los autores de la producción científica que se genera en sus instituciones. Por este motivo, es requisito indispensable que el repositorio de datos soporte ORCID como PID y que permita así asociar un investigador con los conjuntos de datos que genera.

## 5.2 Capacidad de almacenamiento alto

### RF3: Admitir archivos hasta 10GB por defecto

En las recomendaciones para seleccionar un repositorio para el depósito de datos de investigación (CSUC, 2017) se comparan los cinco repositorios de datos multidisciplinares más destacados que pueden usarse gratuitamente. Dos de ellos permiten el depósito de archivos de 10GB o más y, en cambio, los otros tres sólo aceptan archivos hasta un máximo de 5 GB.

Además, en dos ocasiones[1] se ha preguntado a los investigadores catalanes sobre el tamaño de los archivos que usan y se ha constatado que la mayoría de conjuntos de datos son menores a 10GB. De todas formas, habrá que prever mecanismos para almacenar aquellos ficheros de tamaños superiores que, para cumplir con los requisitos de las agencias de financiación, también deben ser públicos.

### $RF_4$: Espacio elástico para que pueda crecer

Aunque la publicación de datos de investigación es todavía un tema incipiente, se considera que esta actividad irá incrementando exponencialmente. Por este motivo, es necesario que el espacio para almacenar los datos sea elástico, para que pueda ir adaptándose a la demanda de los usuarios.

[1] En las dos encuestas (2016 y 2018) se ha preguntado por la medida del fichero. Ambos informes se pueden consultar en RECERCAT.

## 5.3 Prestaciones medias-altas de preservación

Todos los archivos digitales que se están creando hoy en día tienen el riesgo de convertirse en obsoletos; y los datos de investigación no son una excepción (Johnston, 2017). Los datos digitales son más frágiles que los registrados en papel ya que según el tipo de medio en el que se almacena (magnético, óptico, etc.), con el tiempo, quedan expuestos a diferentes daños o a la descomposición. Tener una estrategia de preservación y comprometerse a mantener el acceso a la información a largo plazo ayuda a mitigar estos riesgos.

### $RF_5$: Disponer del archivo para 10 años como mínimo

El proyecto Leaders Activating Research Networks (LEARN) recogió diferentes políticas europeas sobre gestión de datos de investigación (LEARN, 2017) y constató que, generalmente, se establecía un periodo de 10 años como plazo para conservar datos. Por este motivo, se considera que el repositorio debe permitir disponer del archivo para 10 años como mínimo.

Este plazo, sin embargo, no debe vincularse a la fecha de publicación de los ficheros, sino que, con el fin de preservar aquellos datos más relevantes y utilizados, se podría considerar que el plazo de 10 años contabilice desde la última consulta o descarga.

### $RF_6$: Disponer de 2 copias geográficamente distribuidas como mínimo

La National Digital Stewardship Alliance (NDSA) ha definido el número de copias como uno de los elementos a tener en cuenta dentro de sus niveles de preservación digital (Phillips, 2013). Según indican, el primer requisito para asegurar el acceso al archivo en el futuro es tener, como mínimo, una segunda copia del archivo. Además, estas copias deben estar geográficamente distribuidas para evitar amenazas regionales, como desastres naturales o humanos.

Por este motivo, es necesario que el repositorio dé acceso a una de las copias en tiempo real y que tenga, como mínimo, otra copia oscura separada geográficamente.

### $RF_7$: Comprobar la integridad de los datos periódicamente

A fin de proteger la integridad de los datos, será obligatorio verificar periódicamente que los datos almacenados no han sido corrompidos o modificados accidental o deliberadamente. Esto debe hacerse mediante una suma de verificación (o checksum) que consiste en sumar cada uno de los componentes básicos de un sistema (normalmente cada byte) y almacenar el valor del resultado para compararlo posteriormente. En caso de que el valor sea el mismo, se considera que ese archivo no ha sido alterado.

El repositorio deberá realizar periódicamente estas verificaciones. Además, será necesario mantener un registro con la información de la integridad de los archivos para detectar aquellos que sean dañados.

### $RF_8$: Seguir el modelo de preservación OAIS

El modelo Open Archival Information System (OAIS) se usa actualmente para preservar a largo plazo la información científica en formato digital (Hirtle, 2001). Además de capturar los metadatos disponibles durante el proceso de ingesta, los depósitos de datos a menudo distribuyen esta información a otras instancias, mejorando la visibilidad de las publicaciones a través de motores de búsqueda especializados en investigación o indexadores de repositorios.

Se considera, pues, que el repositorio deberá seguir este modelo, ya que es la base para conseguir el certificado CoreTrustSeal[2].

## 5.4 Interoperabilidad con otros sistemas

### $RF_9$: Comunicar con diferentes repositorios

Los datos, generalmente, necesitan integrarse con otros, así como interoperar con aplicaciones o flujos de trabajo para su análisis, almacenamiento o procesamiento (GO FAIR, 2018). También, para evitar la duplicación tareas.

Por este motivo, el repositorio debe permitir intercambiar sus datos con otros sistemas como: con otros repositorios disciplinares para involucrar al resto de la comunidad, con los repositorios institucionales de las universidades, con repositorios de software (por ejemplo, GitHub) o con las herramientas de gestión de la investigación de las universidades (CRIS).

### $RF_{10}$: Comunicar con herramientas de almacenamiento en la nube

Durante el proyecto de investigación, los investigadores almacenan, gestionan y comparten sus datos. Aunque estos procesos se suelen realizar en unidades físicas (discos duros, CDs, etc.) de las universidades cada vez más, se usan herramientas de almacenamiento en la nube (como pueden ser Dropbox, Google Drive, Amazon, UNIDISCAT, entre otros).

Para permitir una carga amigable de los conjuntos de datos desde estas herramientas de almacenamiento en la nube es necesario que el repositorio se comunique vía API con ellas, facilitando así la tarea a los investigadores. En esta comunicación con la nube, será necesario establecer mecanismos para garantizar un control de calidad de los datos.

[2] CoreTrustSeal ofrece una certificación a los repositorios de datos basada en el catalogo y en los processos .

### $RF_{11}$: Exportar los metadatos a diferentes herramientas de descubrimiento

Uno de los objetivos de publicar y compartir los datos de investigación es su reutilización. Por este motivo, la exposición de los contenidos del repositorio en otras plataformas de investigación mejora su visibilidad y hace que su alcance incremente (Amorim, 2017). Es esencial, pues, que el repositorio exporte sus metadatos a diferentes herramientas de descubrimiento.

En una primera instancia, el Portal de la Recerca de Catalunya debe servir como punto para visibilizar la investigación realizada en las universidades catalanas. Por otra parte, también será necesario que esta misma información se haga visible en el European Open Science Cloud (EOSC)[3], así como, también tener en cuenta otros organismos o productos que puedan usar los investigadores: OpenAIRE o Google DataSearch, por ejemplo.

Además, es necesario que el repositorio sea aceptado dentro del R3Data, el registro mundial de los diferentes repositorios de datos de investigación, como una herramienta más de descubrimiento.

### $RF_{12}$: Usar protocolos de comunicación estándar

La información contenida por los repositorios de datos también es relevante para otros sistemas de información (por ejemplo, los de las agencias de financiación) y, por lo tanto, es necesario que se comunique con ellos. Por este motivo, es esencial que el repositorio permita consumir y distribuir metadatos mediante el Open Archive Initiative-Protocolo for Metadata Harvesting (OAI-PMH). Se trata de un protocolo de interoperabilidad que sirve como medio para recoger metadatos de otros repositorios y permitir la distribución y reutilización de metadatos de los depósitos (Devarakonda, 2011).

Para facilitar otras conexiones, es necesario que el repositorio permita la inclusión de diferentes APIs.

### $RF_{13}$: Usar formatos estándares de datos

Inicialmente, cada conjunto específico de datos y sus aplicaciones asociadas se codificaban con sus propios formatos, pero esto no fomentaba la interoperabilidad ni la extensibilidad de los datos, como tampoco de los datos de investigación. Por este motivo, permitir que los datos de una máquina se pudieran almacenar o procesar en otra máquina era el objetivo de la creación de formatos estándares de datos.

De entre los diferentes formatos estándares es necesario que el repositorio soporte dos de los más comunes, como son el eXtensible Markup Language (XML) y el JavaScript Object Notation (JSON) (DeYoung, 2015).

[3] Actualmente, se estan definiendo los requisitos para poder participar y visibilitzar los datos de investigación en el EOSC. Por este motivo, el repositorio deberá adaptar sus características a la vez que se definen los requisitos.

## 5.5 Gestión de características especiales

### $RF_{14}$: Permitir diferentes versiones de un mismo dataset

Los conjuntos de datos digitales son mucho más flexibles y se actualizan periódicamente a medida que se recogen nuevos datos (lo que se conoce como versioning). Este hecho, sin embargo, conlleva problemas de reproducibilidad a la hora de citarlos (FREYA, sf).

Actualmente, el proyecto europeo FREYA[4] está estudiando la mejor solución para asignar DOIs a las versiones. Siguiendo sus recomendaciones, y para ayudar al usuario a navegar entre las diferentes versiones, el repositorio debe asignar un DOI para cada conjunto de datos y, añadir la versión al final. El DOI siempre llevará a la página de destino con la última versión y contendrá un registro de los diferentes cambios.

### $RF_{15}$: Gestionar diferentes esquemas de metadatos

Los metadatos son la columna vertebral de la curación de datos (Higgins, 2007), ya que informan descriptiva o contextualmente un objeto o un recurso. Se componen por una serie de elementos: metadatos descriptivos, técnicos, administrativos, de gestión y de preservación. Los esquemas de metadatos incluirán todos o algunos de los elementos citados y se usan para facilitar a la comunidad a usar los datos de otros investigadores.

El repositorio debe permitir gestionar diferentes esquemas de metadatos, desde un esquema general y básico para todos (como podría ser el Dublin Core) hasta otros más específicos para disciplinas. Como los requisitos de los metadatos pueden variar según las disciplinas, el repositorio debe ser flexible y adaptarse a cada contenido.

Independientemente del esquema de metadatos que se elija, es necesario que el repositorio pueda incluir metadatos de entidades financiadoras, proyectos, etc. para vincular un conjunto de datos con el proyecto asociado. Además, vincular el conjunto de datos con otros resultados relacionados, ya sean artículos u otros conjuntos de datos.

Por otro lado, y teniendo presente que algunas de los metadatos relacionados con un conjunto de datos se suelen repetir (entidad financiadora, investigadores relacionados, etc.), el repositorio debe permitir copiar o replicar estos datos a fin de facilitar y agilizar las tareas en el depósito.

### $RF_{16}$: Tipo de acceso

Siguiendo las premisas de la CE (European Commission, 2016), los datos deben ser tanto abiertos como sea posible y tan cerrados como sean necesario. Por tanto, el repositorio debe

[4] El proyecto FREYA està financiado por la Comisión Europea y tiene como objectivo ampliar la infraestructura de identificadores persistentes (PIDs) como componente básico de la investigación en abierto, tanto en la Unión Europea como a nivel mundial.

permitir publicar los datos en abierto por defecto o, si es necesario, con embargo de un periodo razonable.

Además, en los casos que por razones legales o éticas que no se permita compartir los datos en abierto, el repositorio deberá permitir incluir datos sensibles o confidenciales de manera cerrada, pero informando de que se encuentran depositados aquí. Los controles de acceso a este tipo de datos deben ser siempre proporcionales a la tipología de datos y el nivel de confidencialidad de que se trate. El repositorio debe permitir acceder a algunos de estos datos mediante controles de acceso, ya sea por IP, por invitación, etc.

Por otra parte, el repositorio ha de permitir la autenticación única (*single sign-on*), ya que es un método que permite a los usuarios sólo tener que proceder a una autenticación para acceder a diversas aplicaciones informáticas o sitios web. Por ejemplo, en el contexto de las universidades del CSUC, con UNIFICAT; pero también con otros sistemas.

### $RF_{17}$: Aceptar cualquier tipo de formato

El formato y el software con el que se crean los datos de investigación suelen depender de cómo los investigadores recogen y analizan los datos, que, a su vez, viene determinado por normas y costumbres específicos de cada disciplina (UK Data Service, 2014). Por este motivo, el repositorio debe aceptar cualquier tipo de formato surgido de las diferentes disciplinas.

Algunos formatos privativos (como pueden ser Microsoft Excel y SPSS) son utilizados ampliamente y susceptibles de ser accesibles por un tiempo determinado. A pesar de esto, la opción más segura para garantizar el acceso a los datos a largo plazo es usar formatos estándares abiertos[5].

### $RF_{18}$: Permitir diferentes tipos de ingesta

El repositorio debe permitir la transferencia de datos mediante diferentes tipos de ingesta como puede ser: centralizada, delegada (por medio del personal de apoyo a la investigación de las universidades), auto archivo (los propios investigadores son los que cuelgan los ficheros) y por lotes.

### $RF_{19}$: Ofrecer la citación recomendada

La citación es una de las vías básicas sobre las que se basa la publicación científica y académica. La cita de datos, así como la cita de otras fuentes y pruebas, es una buena práctica y forma parte del ecosistema académico que promueve la reutilización de los datos (Data Citation Synthesis Group, 2014).

[5] UK Data Service ha definido los formatos recomendados y aceptado por cada tipología de datos. Por ejemplo, para datos tabulados, aunque recomiendan los ficheros .csv también aceptan el .xls.

El repositorio debe ofrecer la citación recomendada siguiendo los estándares que se están definiendo actualmente en comunidades de trabajo como el Data Citation WG[6] de la Research Data Alliance (RDA). Además, el hecho de citar los datos permitirá contabilizar el uso de éstos (Data Level Metrics, DLM) tal y como ya se hace con las altmetricas (Article Level Metrics, ALM)

### $RF_{20}$: Permitir la difusión de los datasets a través de plugin de compartición

Cada vez más investigadores quieren difundir sus datos de investigación por las redes sociales para favorecer la reproducibilidad y poder comparar los resultados de investigación (Weller & Kinder-Kurland, 2016). Además, permiten comunicar en tiempo real y saber qué opinan otros investigadores.

Por este motivo, el repositorio debe permitir el uso de las APIs de las redes sociales más utilizadas, como Twitter, Facebook, Linkedin, etc.

### $RF_{21}$: Gestionar diferentes tipos de licencias

Cuando se publican datos de investigación, es obligatorio otorgarle una licencia que puede ir del espectro más abierto hasta la más restrictivo. La CE recoge que "en la medida de lo posible, los proyectos deben tomar medidas que permitan a terceros acceder, hacer minería, explotar, reproducir y difundir los datos de investigación. Una forma sencilla y eficaz es otorgar una licencia de Creative Commons (CC-BY o CC0)".

Dado que hay un número elevado de licencias posibles (Ball, 2014), el repositorio debe permitir que se puedan gestionar y otorgar diferentes tipos de licencias (como, por ejemplo, las GNU). Así como, también debe jugar un papel autoexplicativo sobre qué representa otorgar una licencia u otra.

### $RF_{22}$: Ofrecer datos analíticos de uso de la plataforma

En cualquier plataforma en línea, el módulo de estadísticas es esencial para comprobar el uso que hacen los usuarios. El repositorio debe ofrecer estadísticas de uso a diferentes niveles: institucional, por departamento, por investigador y por conjunto de datos.

Por otra parte, también es necesario que el repositorio ofrezca el número de veces que se ha descargado y citado el conjunto de datos.

[6] El grupo de trabajo del RDA sobre citación de los datos, tiene el objetivo de reunir expertos para discutir problemas, requisitos, ventajas y deficiencias de las iniciativas para citar eficientemente los conjuntos de datos.

### $RF_{23}$: Suministrar los metadatos para su reutilización

El repositorio de datos debe poder suministrar vía API los metadatos para su reutilización, así como otorgar una licencia CC0 para la reutilización de los metadatos y comunicarlo en la web (tanto la política de acceso como la de reutilización).

### $RF_{24}$: Ser fácilmente usable

La usabilidad del repositorio de un repositorio de datos es importante para garantizar que los usuarios puedan acceder a los datos, permitiendo cargar, descargar y citar fácilmente el conjunto de datos. Así pues, el repositorio debe disponer de una interfaz web de interacciones sencillas e intuitivas siguiendo los parámetros del diseño adaptativo para que se adapte a cualquier tipo de pantalla.

Además, es necesario que esta usabilidad también se aplique al apartado de administración, ya que el personal de apoyo a la investigación de las universidades debe poder gestionar fácilmente sus colecciones tanto desde el nivel institucional hasta el nivel de investigador.

El repositorio también debe previsualizar los conjuntos de datos en la plataforma para que no sea necesario descargar el archivo para darle un vistazo. La previsualización debe ser para todos aquellos formatos de archivos de código abierto, y en la medida de lo posible, para el resto de los formatos. Del mismo modo, el repositorio debe tener mecanismos para poder ver qué archivos hay dentro de los ficheros deshidratados o comprimidos.

Finalmente, el repositorio debe usar una terminología consensuada y comprensible para la comunidad de investigadores.

### $RF_{25}$: Cumplir con la legislación vigente

Uno de los motivos por los que es necesario disponer y almacenar los datos es para asegurar el cumplimiento de la legislación vigente.

# 6. Buenas prácticas

Las entrevistas con expertos que se hicieron para elaborar este informe tenían la intención de hacer una relación de requisitos funcionales para el repositorio. Pero desde las primeras reuniones y entrevistas se comprobó que estos expertos consideraban que, adicionalmente a los requisitos técnicos, había un aspecto fundamental para la gestión de datos de investigación y que era desarrollar buenas prácticas para la gestión de datos.

El acceso y reutilización de los datos de investigación no depende tan sólo de las prestaciones del repositorio donde estén publicados. En muchos casos la utilidad potencial de los datos está asociada a la propia disposición de estos o a los elementos que los describen (metadatos). Mientras que hace mucho tiempo que se publican artículos científicos, de hecho, hace muy poco que se publican y comparten datos. Si a esto le sumamos que la diversidad de los datos (en tipología, en dimensiones, en forma de presentación...) es muy alta, nos encontramos con que la falta de prácticas estandarizadas o consolidadas para disponer los datos es uno de los principales obstáculos para su gestión.

La poca experiencia en gestión de datos hace que estas buenas prácticas no estén establecidas y que sea necesario adquirirlas. Y para ello es necesario experimentar con un repositorio de datos. Si bien una primera función del repositorio es publicar datos, la segunda, pero no menos importante, es permitir generar experiencia y buenas prácticas a partir del momento en que este sea operativo.

Las 'buenas prácticas' que los expertos recomiendan adquirir y consolidar se pueden agrupar en las siguientes categorías:

- Hacer curación de los datos
- Seleccionar los conjuntos de datos
- Fomentar el uso de formatos abiertos
- Usar estándares, protocolos y vocabularios controlados ampliamente aceptados

## 6.1 Hacer curación de los datos

Curar los datos de investigación es el proceso de gestionarlos durante todo su ciclo de vida para que puedan estar disponibles y reutilizables a largo plazo. Entre las diferentes actividades relacionadas con la curación observamos: el descubrimiento, la identificación, selección, obtención, verificación, análisis, gestión, almacenamiento, publicación y citación.

Todas estas actividades requieren que se documenten correctamente los datos para garantizar la transparencia y la reproducibilidad de la investigación en el futuro, no sólo para los mismos investigadores sino para otros (Radboud Univeristy, 2018).

Por ello es necesario que los datos se acompañen de la documentación correcta, incluyendo ficheros que expliquen: el contexto (cómo se ha realizado la investigación y suele incluir, registros de versiones, cuadernos de laboratorio, metodologías o protocolos estandarizados, software, etc.), la estructura (a menudo, archivos readme.txt que contienen una descripción general de las diversas carpetas y ficheros) y el contenido (que describe los conceptos, su significados, los valores numéricos y los códigos y esquemas de clasificación utilizados).

## 6.2 Seleccionar los conjuntos de datos

Una vez los datos se han recolectado o generado, los archivos resultantes pueden actualizarse varias veces hasta llegar a la versión final que se publicará. Es necesario, pues, establecer protocolos y criterios que permitan elegir qué datos finales hay que preservar a largo plazo, así como las razones para descartar algunos otros.

Entre los diferentes criterios a tener en cuenta está la singularidad o repetibilidad, así como el valor, la calidad, los costes de reproducción, el riesgo de pérdida y las indicaciones para la reutilización en publicaciones o por otros usuarios (Tjalsma & Rombouts, 2011).

Las universidades catalanas deberán ofrecer información y pautas para ayudar a los investigadores a decidir qué datos deben conservar, suprimir y publicar en el repositorio.

## 6.3 Fomentar el uso de formatos abiertos

Para asegurar que los datos sean utilizables y recuperables a largo plazo, la elección del formato de archivo es esencial y siempre estará asociado a un software y un hardware. Los formatos abiertos, o no propietarios, son aquellos donde el código del software está disponible para cualquier persona, de manera gratuita, para que cualquiera pueda usarlas sin ninguna limitación en la reutilización (Biblioteca de la CEPAL, 2019).

Por ello, será necesario fomentar el uso de formatos no propietarios, estándares abiertos y documentados, que sean comúnmente utilizados dentro de la comunidad de investigación, que se transmitan mediante formas de representación estándar (ASCII, Unicode), que no estén encriptados ni comprimidos. Para facilitar la tarea a los investigadores, sería conveniente elaborar materiales informativos sobre las mejores prácticas en formatos.

En los casos en que sea posible, será necesario migrar los formatos propietarios a formatos abiertos.

## 6.4 Usar estándares, protocolos y vocabularios controlados ampliamente aceptados

Cada disciplina dispone de esquemas de metadatos, protocolos y vocabularios especializados. Para asegurar la comprensión y la reutilización de estos datos, será necesario que estén descritos con estándares de metadatos apropiados para la disciplina.

Del mismo modo, habrá que apoyar a los investigadores para que usen vocabularios controlados en los metadatos para fomentar la interoperabilidad entre los distintos sistemas.

# 7. Recomendaciones finales

En función de las informaciones recogidas y consideraciones anteriores, se hacen las siguientes recomendaciones:

1. Crear de forma inmediata y aquí un repositorio donde se puedan publicar los datos de investigación de forma FAIR y que permita desarrollar experiencia y buenas prácticas en la gestión de datos de investigación. Este repositorio:
    a. Debe cumplir los requisitos FAIR ya conocidos o los que establezca la CE en un futuro inmediato.
    b. Debe de ofrecer prestaciones de valor añadido respecto a las opciones actuales, como, por ejemplo, asignación de DOIs, interoperabilidad con el Portal de la Investigación de Catalunya y de preservación).

2. El repositorio que se cree:
    a. Debe cumplir los requisitos que se han citado en el apartado 5 de este documento y que se consideran al mismo tiempo mínimos y razonables
    b. Debe hacerse con software ya existente, dado que hay disponibilidad, y de código libre.

3. Paralelamente a la puesta en funcionamiento del repositorio,
    a. Es necesario promover la existencia del servicio e informar de la importancia de publicar los datos de investigación en abierto y de elaborar planes de gestión de datos.
    b. Es necesario que las acciones de difusión se hagan de forma coordinada y que participen todas las unidades que vehiculan la investigación en la universidad o centro determinado y que las oficinas y servicios de investigación participen en las mismas.
    c. Es necesario hacer formación sobre los conceptos de Ciencia Abierta en general y, concretamente, sobre la gestión de datos de investigación. Siguiendo las directrices de la European Commission Expert Group on FAIR Data, esta formación debe incluir a la totalidad de miembros de la comunidad universitaria, pero para ser eficaz, debería distinguir entre usuarios avanzados, investigadores jóvenes y personal de apoyo de las universidades.

# Referencias


Amorim, R. C. (2017). A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society, 16*(4), 851-862.

Ayris, P., Bernal, I., Cavalli, V., & al, e. (2018). *LIBER Open Science Roadmap.* doi:10.5281/zenodo.1303002

Ball, A. (2014). *How to licence Research Data.* Digital Curation Center. Recollit de http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_To_License_Research_Data.pdf

Biblioteca de la CEPAL. (2019). *Formatos abiertos y cerrados*. Recollit de Gestión de datos de investigación: https://biblioguias.cepal.org/gestion-de-datos-de-investigacion/formatos

Couto Corrêa, F. (2016). *Gestión de datos de investigación.* Barcelona: UOC.

CRUE. (2019). *Compromiso de las universidades ante la Open Science.* Recollit de http://www.crue.org/Documentos%20compartidos/Informes%20y%20Posicionamientos/2019.02.20-Compromisos%20CRUE_OPENSCIENCE%20VF.pdf

CRUE. (2019). *Compromisos de las universidades ante la Open Science.* Recollit de http://www.crue.org/Documentos%20compartidos/Informes%20y%20Posicionamientos/2019.02.20-Compromisos%20CRUE_OPENSCIENCE%20VF.pdf

CSUC. (2016). *Data management plans: Version 2, December 2016*. RECERCAT. Recollit de http://hdl.handle.net/2072/270395

CSUC. (2016). *Plans de gestió de dades: Versió 2, Desembre 2016*. RECERCAT. Recollit de http://hdl.handle.net/2072/273194.

CSUC. (2016). *Proposta per establir una política d'accés obert a les dades de recerca a les Universitats de Catalunya (Doc.16/30).*

CSUC. (2016). *Universitats catalanes acorden l'ús de l'identificador ORCID als seus investigadors*. Recollit de https://www.csuc.cat/ca/novetat/universitats-catalanes-acorden-l-us-de-l-identificador-orcid-per-als-seus-investigadors

CSUC. (2017). *Ampliació de prestacions dels repositoris institucionals per dipositar dades (Doc.CO17/03).*

CSUC. (2017). *Propostes de treball referent a les dades de recerca obertes (Doc.CO17/11).*

CSUC. (2017). *Recomanacions per seleccionar un repositori per al dipòsit de dades de recerca: Versió 3, Maig 2017*. RECERCAT. Recollit de http://hdl.handle.net/2072/284974

CSUC. (2018). *Model de política de gestió de dades de recerca per a una universitat (Doc.CO18/03).*

CSUC. (2019). *Gestió de dades de recerca: resultats de l'enquesta de 2018 (Doc.CO18/19).*

Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles.* San Diego CA: FORCE11. doi:10.25490/a97f-egyk

Devarakonda, R. P. (2011). Data sharing and retrieval using OAI-PMH. *Earth Sci Inform, 4*(1). doi:10.1007/s12145-010-0073-0

DeYoung, L. (2015). *An Analysis of XML and JSON*. Recollit de COMP 150-IDS: 2015 Spring Term Final Papers: https://www.cs.tufts.edu/comp/150IDS/final_papers/lizzied.3/FinalReport.html

Dutch Ministry of Education, Culture and Science. (2017). *National Plan Open Science*. doi:10.4233/uuid:9e9fa82e-06c1-4d0d-9e20-5620259a6c65

EOSC (2017). *European Open Science Cloud Declaration*. https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf

European Commission. (2013). *Digital science in Horizon 2020*. Recollit de https://ec.europa.eu/digital-single-market/en/news/digital-science-horizon-2020

European Commission. (2016). *Guidelines on FAIR Data Management 2020*. Recollit de http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

European Commission. (2016). *Open innovation, open science, open to the world: A vision for Europe*. doi:10.2777/061652

European Commission. (2016). *Open Research Data in Horizon 2020*. Recollit de http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf

European Commission. (2018). *Open Science Policy Platform Recommendations*. doi:10.2777/958647

European University Association. (2018). *EUA Roadmap on Research Assessment in the Transition to Open Science*. Recollit de https://eua.eu/downloads/publications/eua-roadmap-on-research-assessment-in-the-transition-to-open-science.pdf

FREYA. (sense data). *Case study: Versioning with identifiers*. Recollit de https://project-freya.readme.io/docs/creating-dois-and-doi-metadata

GO FAIR. (2018). *FAIR Principles*. Recollit de https://www.go-fair.org/fair-principles/

Government of the Netherlands. (2016). *Amsterdam Call for Action on Open Science*. Recollit de https://www.government.nl/binaries/government/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science/amsterdam-call-for-action-on-open-science.pdf

Government of the Republic of Slovenia. (2015). *National strategy of Open Access to scientific publications and research data in Slovenia (2015-2020)*. Recollit de http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/Znanost/doc/Zakonodaja/Strategije/National_strategy_for_open_access.pdf

Higgins, S. (2007). *What are Metadata Standards*. Recollit de Digital Curation Center: http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards

Hirtle, P. (2001). OAI and OAIS: what's in a name? *D-lib magazine, 7*(4).

Johnston, L. (2017). *Curating reserach data: A handbook of current practice*. Chicago: Association of College and Research Libraries.

Kim, S. (2018). Functional Requirements for Research Data Repositories. *International Journal of Knowledge Content Development & Technology, 8*(1), 2-36. doi:10.5864/IJKCT.2018.8.1.025

LEARN. (2017). *LEARN Toolkit of Best Practice for Research Data Management.* doi:10.14324/000.learn.00

Lee DJ, S. B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PLoS ONE, 12*(3). doi:10.1371/journal.pone.0173987

LERU. (2018). *Open Science and its role in universities: a roadmap for cultural change.* Recollit de https://www.leru.org/files/LERU-AP24-Open-Science-full-paper.pdf

Ministère de l'enseignement supérieur, de la recherche et de l'innovation. (2018). *National plan for Open Science.* Recollit de https://libereurope.eu/wp-content/uploads/2018/07/SO_A4_2018_05-EN_print.pdf

Ministry of Education, Science and Technological Development. (2018). *Open Science Platform.* Recollit de https://www.openaire.eu/blogs/serbia-has-adopted-a-national-science-policy

OCDE. (2004). *Declaration on Access to Research Data from Public Funding.* Paris. Recollit de http://goo.gl/Iovbt7

Phillips, M. B. (2013). The NDSA levels of digital preservation: Explanation and uses. *Archiving Conference Society for Imaging Science and Technology., 1*, 216-222.

Presidência do Conselho de Ministros. (2016). *Resolução do Conselho de Ministros n.º21/2016 para a implementação de una Política Nacional de Ciência Aberta.* Recollit de https://dre.pt/pesquisa/-/search/74094659/details/maximized

Radboud Univeristy. (2018). *Documenting data.* Recollit de Research data management: https://www.ru.nl/rdm/processing-data/documenting-data/

The Ministry of Education and Culture's Open Science and Research Initiative. (2014). *The Open Science and Research Roadmap.* Recollit de http://www.avointiede.fi/

Tjalsma, H., & Rombouts, J. (2011). *Selection of Research Data: Guidelines for appraising and selecting research data.* Data Archiving and Networked Services (DANS). Recollit de https://dans.knaw.nl/en/about/organisation-and-policy/legal-information/DANSselectionofresearchdata.pdf

UK Data Service. (2014). *Formatting and organising research data*. Recollit de https://www.ukdataservice.ac.uk/media/440281/formattingorganising.pdf

Weller, K., & Kinder-Kurlanda, K. (2016). A manifesto for data sharing in social media research. *Proceedings of the 8th ACM Conference on Web Science* (p. 16-172). Hannover: ACM. doi:10.1145/2908131.2908172

YERUN. (2018). *YERUN Statement on Open Science.* Recollit de https://www.yerun.eu/wp-content/uploads/2018/05/YERUN_OpenScience_Statement-3.pdf

# Anexos

## Anexo 1 – Principios FAIR

En 2016 se publicaron los "'FAIR Guiding Principles for scientific data management and stewardship" en Scientific Data. Los autores pretendían proporcionar directrices para mejorar la investigación, la accesibilidad, la interoperabilidad y la reutilización de los recursos digitales. Estos principios tienen en cuenta el potencial de las máquinas, ya que los humanos confían cada vez más en su apoyo para hacer frente al aumento del volumen, la complejidad y la velocidad de creación de los datos.

Los principios son los siguientes:

- Encontrables:
  - E1. Asignar un identificador único y persistente a los datos y los metadatos
  - E2. Describir los datos con metadatos enriquecidos
  - E3. Registrar/Indexar los datos y los metadatos en un recurso de búsqueda
  - E4. En los metadatos se debe especificar el identificador de los datos que se describen.
- Accesibles:
  - A1 Los datos y los metadatos pueden ser recuperados por sus identificadores mediante protocolos estandarizados de comunicación
    - A1.1 Los protocolos tienen que ser abiertos, gratuitos e implementados universalmente
    - A1.2 El protocolo debe de permitir procedimientos para la autentificación y la autorización (cuando sea necesario).
    - A2 Los metadatos deben de estar accesibles incluso cuando los datos ya no estén disponibles.
- Interoperables:
  - I1. Los datos y los metadatos deben de usar un lenguaje formal, accesible, compartible y ampliamente aplicable para representar el conocimiento
  - I2. Los datos y los metadatos usan vocabularios que sigan los principios FAIR
  - I3. Los datos y los metadatos incluyen referencias cualificadas a otros datos o metadatos
- Reutilizables:
  - R1. Los datos y los metadatos contienen una multitud de atributos precisos y relevantes
    - R1.1. Los datos y los metadatos se publican con una licencia clara y accesible sobre su uso y reutilización
    - R1.2. Los datos y los metadatos se asocian con información sobre su procedencia
  - R1.3. Los datos y los metadatos siguen los estándares relevantes que usa la comunidad del dominio concreto.

## Anexo 2 – Expertos

### 1. Miembros de la Comisión

| | |
|---|---|
| **Ignasi Labastida i Juan** | Universitat de Barcelona. Responsable de la Unidad de Investigación e Innovación del CRAI |
| **Jordi Hernández Sánchez** | Universitat Autònoma de Barcelona. Comisionado de la rectora para las Tecnologías de la Información y la Comunicación |
| **Anna Rovira Fernández** | Universitat Politècnica de Catalunya. Responsable de la Unidad de Recursos para la Investigación del Servicio de Bibliotecas, Publicaciones y Archivos |
| **Antoni Borràs i Escorihuela** | Universitat Pompeu Fabra. Gestor de proyectos del Servicio de Informática |
| **Brigit Nonó Rius** | Universitat de Girona. Responsable de la Unidad de desarrollo de proyectos de la Biblioteca |
| **Leticia Carro de Diego** | Universitat de Lleida. Directora del Área de Investigación y Transferencia |
| **José Luis González** | Universitat Rovira i Virgili. Responsable de la Sección de Gestión de Producción Científica del CRAI |
| **Rosa Padrós Cuxart** | Universitat Oberta de Catalunya. Técnica de Biblioteca para la Investigación |
| **Mireia Salgot** | Universitat de Vic-Universitat Central de Catalunya. Directora de la Biblioteca |
| **Anna Caellas Camprubí** | Universitat Ramon Llull. Técnica de la Oficina de Investigación e Innovación |

## 2. Expertos del estado español

| | |
|---|---|
| **Ernest Abadal i Falgueras** | Universitat de Barcelona. Director del Centro de Investigación en Información, Comunicación y Cultura |
| **Lluís Alfons Ariño Martín** | Universitat Rovira i Virgili. Director del Servicio de Recursos Informáticos y TIC |
| **Mercè Cabo Rigol** | Universitat Pompeu Fabra. Vicegerente del Área de Servicios, Tecnología y Recursos de Información |
| **Anna Maria Casaldàliga Riera** | Universitat Pompeu Fabra. Subdirectora de Biblioteca |
| **Eva Estupinyà Pinyol** | Universitat de Lleida. Responsable de Servicios a los Usuarios de la Biblioteca |
| **Jorge García Pérez** | Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria. Jefe de sección en el Servicio de Informática |
| **Alexandre López-Borrull** | Universitat Oberta de Catalunya. Director del Grado de Información y Documentación |
| **Manuel Lozano Nebro** | Universitat Pompeu Fabra. Responsable del Servicio de Informática |
| **Teresa Malo de Molina Martín-Montalvo** | Universidad Carlos III de Madrid. Directora de Biblioteca |
| **Eva María Méndez Rodríguez** | Universidad Carlos III de Madrid. Vicerrectora Adjunta de Política Científica. Open Science |
| **María Fernanda Peset Mancebo** | Universitat Politècnica de València. Investigadora y bibliotecaria |
| **Antonio Juan Prieto Jiménez** | Universitat Politècnica de Catalunya. Técnico informático del Servicio de Biblioteca |

| | |
|---|---|
| **Marta Renato** | Barcelona Supercomputing Center. Responsable de la RES y coordinador de proyectos |
| **Sandra Reoyo Tudó** | Consorci de Serveis Universitaris de Catalunya. Coordinadora de Recursos de Información para la Investigación |
| **Pilar Rico Castro** | Fundación Española para la Ciencia y la Tecnología (FECYT). Jefe de la Unidad de Acceso Abierto, Repositorio y Revistas |
| **Laia Ros i Blanco** | Universitat Ramon Llull. Responsable de la Oficina de Investigación e Innovación |
| **Antonio Sánchez-Padial** | Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria. Jefe del Servicio de Biometría |
| **Jordi Sorribas Cervantes** | Centre Mediterrani d'Investigacions Marines i Ambientals (CMIMA), CSIC. Director |
| **Miquel Térmens Graells** | Universitat de Barcelona. Decano de la Facultad de Biblioteconomía y Documentación |
| **Nadia Tonello** | Barcelona Supercomputing Center. Responsable de la gestión de datos |
| **David Vicente Dorca** | Barcelona Supercomputing Center. Gestor de apoyo al usuario |
| **Ricard de la Vega Sivera** | Consorci de Serveis Universitaris de Catalunya. Responsable de Cálculo y Aplicaciones |

## 3. Expertos internacionales

| | |
|---|---|
| **Lucy Amez** | Vrije Universiteit Brussel. Asesora de las políticas en publicaciones científicas, bibliometría y ciencia abierta, Departamento de Investigación y gestión de datos |
| **Jessica Parlant von-Essen** | CSC-IT Center for Science in Finland. Coordinador senior |
| **Joakim Phillipson** | University of Stockholm. Analista de datos de investigación |
| **Jan van Mansum** | DANS. Coordinador de desarrollo de software |
| **Linda Reijnhoudt** | DANS. Desarrollador de software |
| **Eloy Rodrigues** | Universidade do Minho. Director del Servicio de Documentation |
| **Jäarno Saarti** | University of Eastern Finland. Director de Biblioteca |

# Anexo 3 – Países

### Bélgica flamenca

El Vlaamse Interuniversitaire Raad (VLIR) es un órgano consultivo que representa las cinco universidades flamencas (Katholieke Universiteit Leuven, Universiteit Antwerpen, Universiteit Gent, Universiteit Hasselt y Vrije Universiteit Brussel) para facilitar la cooperación interuniversitaria y la interacción con el gobierno flamenco.

El VLIR ha elaborado a través de un grupo de trabajo una encuesta para determinar las necesidades y los requisitos que debería tener el servicio sobre gestión de datos de investigación. Han puesto en marcha una herramienta en línea para elaborar planes de gestión de datos y han elaborado el informe Research Data Management en de Vlaamse Universiteiten: White Paper con diferentes recomendaciones sobre las inversiones que debería hacer el gobierno (infraestructuras, educación, legislación, incentivos).

En cuanto a la infraestructura recomiendan proveer de una infraestructura sostenible conjunta que sea fácilmente usable y sirva de almacenamiento y conservación de los datos.

### Finlandia

Fairdata Services es un conjunto de herramientas interoperables para todo el ciclo de vida de los datos: almacenamiento y preservación (IDA), descripción de los datos y publicación (Qvain) y descubrimiento (ETSIN). Todos estos servicios dependen del Ministry of Education and Culture y han sido desarrollados por el CSC-IT Center for Science Ltd.

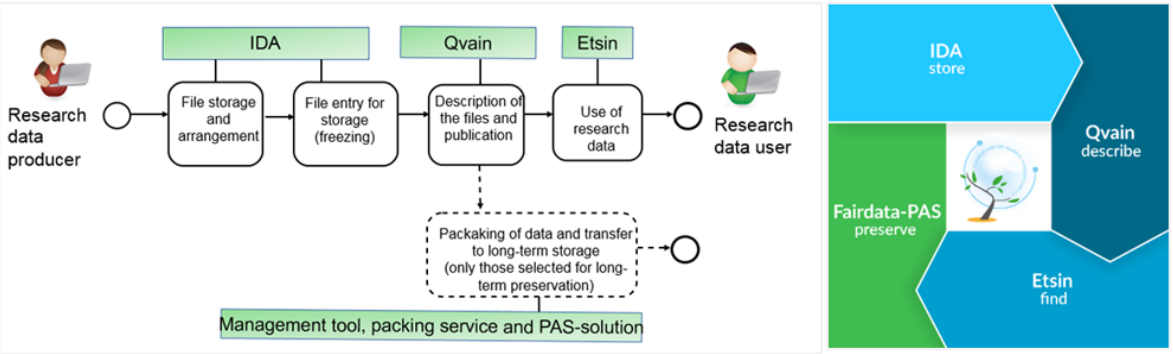


Figura 1. Los diferentes servicios de Fairdata Services de Finlàndia

IDA sirve como almacenamiento fiable para los datos de investigación durante su fase activa. Los datos que se seleccionan se congelan pasando a un estado inmutable válido para la publicación. Se generan metadatos técnicos de los archivos y se traspasan a Qvain.

Qvain es una herramienta de creación de metadatos para que los datos puedan ser publicados. Los conjuntos de datos obtienen un identificador persistente y una página de destino. Los registros de metadatos finalizados se publican en ETSIN.

Etsin es la herramienta de descubrimiento de los conjuntos de datos que indexa fuentes como Qvain y otros recursos externos

**Países Bajos**

El Data Archiving and Networked Services (DANS) es un instituto de la Dutch Academy KNAW financiado por la Netherlands Organisation for Scientific Research (NWO). El DANS tiene la misión de promover y proveer acceso permanente a los recursos de investigación digital y anima a los investigadores a hacer que sus publicaciones y datos de investigación sean encontrables, accesibles, interoperables y reutilizables.

El DANS coordina los diferentes servicios en el "back office" y las universidades ofrecen el apoyo a los usuarios al "front office". Los servicios se organizan según el ciclo de vida de los datos: DataverseNL mientras dure el proyecto, EASY una vez finalice el proyecto para preservar y NARCIS para descubrir.


Figura 2. Los diferentes servicios ofrecidos por el DANS

DataverseNL se inició en 2014 y participan 12 instituciones (Eindhoven University of Technology, Leiden University, Delft University of Technology, Universidad de Maastricht, NIOO-KNAW, Protestantse Theologische Universiteit, Tilburg University, University of Groningen, University of Twente, Utrecht University, University of Applied Sciences Utrecht y Vrije Universiteit Amsterdam). Este repositorio ofrece hasta 10 años después de la finalización de un proyecto: almacenar datos en línea, compartir los con otros investigadores y asignarles un 'handle' como identificador. Las instituciones pagan una cuota para participar.

EASY se puso en marcha en 2005 y es un repositorio de preservación en línea certificado con los sellos Fecha Seal of Approval (DSA), el World Data System (WDS) y Nestor Seal que aseguran el cumplimiento de un conjunto de criterios transparentes sobre calidad, sostenibilidad y accesibilidad. A diferencia del DataverseNL, este repositorio asigna Dois como identificadores persistentes.

Finalmente, NARCIS es el portal nacional para buscar información científica, incluyendo los datos de investigación.

**Portugal**

La Fundação para a Ciência e a Tecnologia (FCT) es la agencia pública nacional para apoyar la investigación en ciencia, tecnología e innovación en todos los ámbitos del conocimiento y depende del Ministério da Ciência, Tecnologia e Ensino Superior.

Esta institución elaboró una propuesta de repositorio de datos consorciado que está a la espera de obtener financiación. Por este motivo, universidades como la Universidade do Minho está implantando su propio repositorio institucional de datos.

**Suècia**

El SND Consortium está formado por 7 universidades (Göteborg Universitet, Karolinska Institutet, Lunds Universitet, Sveriges Iantbruksuniversitet, Stockholm Universitet, Umea Universitet, Uppsala Universitet) que trabaja bajo los mandatos del Swedish Research Council. Las universidades del consorcio han desarrollado buenas prácticas y conocimientos en diferentes disciplinas a través de especialistas que han servido de vínculo entre los investigadores y las oficinas de apoyo en cada universidad.

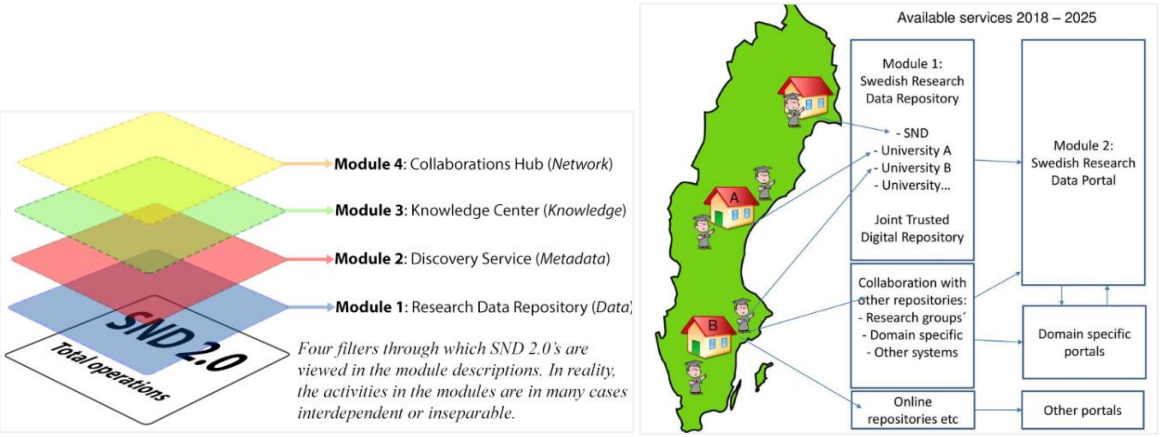


Figura 3. Mòduls que conformen l'SND 2.0

El Swedish National Data Service 2.0 (SND 2.0) se divide en cuatro módulos: Research Data Repository (para los datos), Discovery Service (para los metadatos), Knowledge Center (parapor el conocimiento) y Collaborations Hub (para generar red).

El Swedish Research Data Repository (SRDR) ofrece los servicios de ingesta, curación, acceso a los datos y garantía de calidad. En este estadio se ofrecen buenas prácticas sobre las versiones de los datasets y asignan DOIs como identificadores persistentes, ya que la citación de los datos a través del DOI es mucho más fácil y permite que se indexen como cualquier otro tipo de publicación.

Este repositorio se vehicula juntamente con el Swedish Research Data Discovery Service que quiere ser la herramienta de descubrimiento única de los datos de investigación suecas. La intención era recolectar tanto los datasets del repositorio nacional, como otros repositorios disciplinares.

Hoy por hoy, EUDAT cumple la función de almacenamiento y preservación (suficiente para los próximos años y gratuito). Aun así, se ha formado un piloto con los grandes centros productor de datos a Suecia para desarrollar una solución común.

**Referencias**


CSC-IT (2019). *CSC-IT Center for Science Ltd.* https://www.csc.fi/

DANS (2019). *Data Archiving and Network Services.* https://dans.knaw.nl/en

DANS (2019). *DataverseNL.* https://dataverse.nl/

DANS (2019). *EASY.* https://easy.dans.knaw.nl/ui/home

DANS (2019). *NARCIS.* https://www.narcis.nl/

FAIRDATA.FI (2019). *Etsin. Research dataset finder.* https://www.fairdata.fi/en/etsin/

FAIRDATA.FI (2019). *IDA. Research data storage service.* https://www.fairdata.fi/en/ida/

FAIRDATA.FI (2019). *Qvain. Research dataset metadata tool.* https://www.fairdata.fi/en/qvain/

FCT (2019). *Fundação para a Ciência e a Tecnologia.* https://www.fct.pt/

SND (2018). *Descripcion of the infraestructure and its activities.* SND 2.0)

SND (2019). *Swedish National Data Service.* https://snd.gu.se/en

UNIFI (2018). *Open Science and Data. Action programme for the Finnish scholarly community.* http://urn.fi/URN:NBN:fi-fe2018111648265

VLIR Werkgroep Research Data Management & Open Science (2018). *Research Data Management en de Vlaamse Universiteiten: White Paper.* http://www.vlir.be/media/docs/Onderzoeksbeleid/20180525%20White%20Paper_RDM%20en%20de%20Vlaamse%20Universiteiten_addendum.pdf

VLIR (2019). *Vlaamse Interuniversitaire Raad.* http://www.vlir.be/

## Anexo 4 – Documentación adicional

Además de la documentación disponible en el apartado de referencias de este informe, también se ha consultado la siguiente documentación relacionada con los datos FAIR:


Allen, R. & Hartland, D. (2018). *FAIR in practice*. JISC report on the Findable Accessible Interoperable and Reusable Data Principles. DOI: 10.5281/zenodo.1245568

Austin et al (2015). Reserach Data Repositories: review of corrent features, gap analysis, and recommendations for minimum requirements. *IASSIST Quarterly 39(4)*. DOI: 10.29173/iq904

COPDESS (2018). *Enabling FAIR data commitment State ment in the Earth, Space, and Environ Mental Sciences*. http://www.copdess.org/enabling-fair-data-project/commitment-to-enabling-fair-data-in-the-earth-space-and-environmental-sciences/

CORE TRUST SEAL (2018). *Core Trustworthy Data Repositories Extended Guidance*. https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf

European Commission (2018). *Commission staff working document. Implementation Roadmap for the European Open Science Cloud*. Brussels. SWD(2018) 83 final

Hodson et al. (2018). *FAIR Data Action Plan*. Interim recommendations and actions from the European Commission Expert Group on FAIR data. DOI: 10.5281/zenodo.1285290

Hodson et al. (2018b). *Turning FAIR data into reality*. Interim report of the European Commission Expert Group on FAIR data. DOI: 10.5281/zenodo.1285272

Principe & Rodrigues (2018). *Data RepositóriUM: Projeto de implementaçaõ do repositório de dados para a Universidade do Minho*. 4ª Fórum de Gestaõ de Investigaçaõ

Rodrigues, E. (2019). *Definición e implementación de estrategias y servicios institucionales para la gestión de datos de investigación*.

SPARC (2018). *FAIR and Open Data: a briefing for policymakers and senior managers*. https://sparceurope.org/new-briefing-paper-explores-fair-and-open-data/

## Anexo 5 – Asignar DOIs

Con el fin de crear nuevos Dois y asignar los a los conjuntos de datos es necesario formar parte de la comunidad de DataCite o colaborar con alguno de sus miembros.

Pueden convertirse en miembros, cualquier tipo de organización -incluyendo los centros de datos, editores, bibliotecas, etc.- que muestre su apoyo para compartir los datos de investigación de la siguiente manera:

- Demostrando un elevado nivel de compromiso con los datos de investigación y la ciencia abierta
- Formen parte de una comunidad global de difusión de los datos, aprendiendo, colaborando y defendiendo una red de expertos de datos de investigación.
- Apoyen y participen en la creación y gestión de identificadores persistentes (DOE) para resultados de investigación
- Jueguen un papel crítico para el avance de la misión de compartir los datos

Para unirse a la comunidad, hay que presentar[7] una candidatura que será valorada por los directores de DataCite y pagar una suscripción.

[7] https://datacite.org/assets/datacite_application.pdf

# Anexo 6 – Hardware y software para un repositorio de datos

Para disponer de un repositorio de datos capaz de dar cumplimiento a los requisitos funcionales del quinto apartado de este documento se necesita una combinación de hardware de almacenamiento y un software.

### 6.1 Hardware

Los requisitos piden disponer de espacio elástico para poder crecer y disponer de al menos dos copias alejadas geográficamente. Las actuales cabinas de discos facilitan esta elasticidad, y si se dispone de al menos dos de separadas, se podría dar cobertura a ambos requisitos. La infraestructura de almacenamiento puede ser local a un centro de procesamiento de datos (CPD) que disponga de las suficientes medidas de seguridad y disponibilidad o en la nube.

Los costes de almacenamiento serán recurrentes y crecientes a medida que vamos introduciendo datos al repositorio para su preservación.

### 6.2 Software

Existen varias opciones que se han analizado en diversos estudios (Amorim, 2017; CSUCA, 2017; Rodrigues, 2019) comparando requisitos similares a los detallados en el quinto apartado de este documento. Hay dos tipos de softwares:

- softwares de repositorios institucionales o de portales de transparencia que se han adaptado para introducir también los datos de investigación (por ejemplo, DSpace, ePrints, CKAN) y
- softwares hechos específicamente para datos de investigación

Entre los citados en último lugar, en función de las recomendaciones de este informe, nos hemos centrado sólo en los que pueden almacenar los datos en servidores locales. Mencionemos las características principales de tres softwares, los cuales cumplen, en mayor o menor medida, y de manera directa o mediante plugins, todos los requisitos detallados en el apartado quinto de este documento. Estos son:

- Fighshare como opción comercial
- Dataverse de código libre
- Invenio de código libre (usando la adaptación que ha hecho EUDAT)

Para poder disponer de un mayor cumplimiento de los requisitos de preservación, seguir el modelo OAIS y gestionar que los datos se almacenen en al menos dos cabinas separadas geográficamente, es necesario o recomendable complementar los softwares detallados con otros software especializados para estas funciones.

## Anexo 7 – Glosario

| Ítem | Descripción |
|---|---|
| **European Open Science Cloud** | También conocido como EOSC. Portal común que ofrece acceso consolidado a las diferentes infraestructuras existentes en Europa. |
| **FAIR** | FAIR es el acrónimo de *Findable*, *Accessible*, *Interoperable* y *Reusable* que no sólo se aplica a los datos de investigación sino a todos los outputs de la investigación. Los datos FAIR tienen el objetivo de facilitar la descubierta, la integración y el análisis de los datos de investigación relevantes y de los algoritmos y flujos de trabajo asociados a los mismos. |
| **Gestión de datos de investigación** | También conocido como Research data management (RDM). Se refiere al desarrollo, ejecución y supervisión de los planes, políticas, programas y prácticas que controlan, protegen y mejoran el valor de los datos de investigación |
| **Interoperabilidad** | Proceso que permite compartir datos entre diferentes organizaciones. El objetivo es crear una comprensión compartida de los datos. |
| **Licencia** | Describe los términos en que un material puede ser reusado, almacenado, redistribuido, etc. |
| **Metadatos** | Describen las características básicas de los datos. Suelen incluir la autoría, el título, la fecha, el resumen, palabras clave y la información de licencia. |
| **Plan de Gestión de Datos** | También conocido como Data Management Plan (DMP), es un documento formal que describe como deben gestionarse los datos durante todo el ciclo de vida. |
| **Repositorio** | Un repositorio es una infraestructura que permite el almacenaje persistente, eficiente y sostenible de objetos digitales. |


Consorci de
Serveis Universitaris
de Catalunya