Open Data en Centros de Investigación, los casos del CERN y ALBA

Jose Carcelén Manzanilla
jcarcelen@cells.es

Universitat Oberta de Catalunya / CELLS-ALBA Synchrotron Computing

Elena Planas Hortal y Javier Luis Cánovas Izquierdo

Systems, Software and Models Research Lab, Universitat Oberta de Catalunya

12 de junio de 2017

Resumen

El fenómeno de los datos abiertos se ha convertido en una tendencia mundial, publicándose cada día más y más datos provenientes tanto de empresas privadas como de instituciones públicas. En el mundo de la investigación científica, los datos abiertos están relacionados con el modelo Open Science, que por una parte aboga por el libre acceso a las publicaciones académicas (Open Access) y por otra al acceso a los datos en los que se basan estas publicaciones, de manera que los datos puedan ser usados libremente, reutilizados, y redistribuidos, con el único requisito del de atribuir y compartir por igual (Open Data). Este artículo revisa el estado del arte de la publicación de datos en abierto en el campo de la investigación, y analiza los casos del CERN, a partir de la literatura publicada, y del Sincrotrón ALBA, gracias a la entrevista al responsable de Computing and Control Division, David Fernández Carreiras. El artículo finaliza con una serie de recomendaciones basadas en el análisis personal para investigadores y centros de investigación que deseen publicar sus datos en abierto.

1. Introducción

El acceso por parte de la ciudadanía a los datos generados por las administraciones públicas se ha consolidado durante estos últimos años. En el ámbito de la investigación pública, la apertura de los resultados y los datos científicos ha encontrado reticencias, tanto por la comunidad científica como por la industria editorial. En cambio, organizaciones como la *European Commission* y la *Organisation for Economic Co-operation and Development* defienden y promueven su implantación. En este artículo se revisa la procedencia del término Open Data, sus supuestos beneficios y motivaciones, la normativa actual que lo ampara, las distintas fases del ciclo de publicación de datos y las barreras a las que se enfrenta. Por último, se describe el caso del CERN[1], donde la publicación en abierto es un hecho, y el caso del Sincrotrón ALBA[2], que se encuentra en una fase preliminar de su implantación.

2. Estado del arte

En la declaración de Budapest de 2002 se define el término *Open Access* o acceso abierto a la literatura científica como:

"*su disponibilidad gratuita en Internet público, permitiendo a cualquier usuario leer, descargar, copiar, distribuir, imprimir, buscar o usarlos con cualquier propósito legal, sin ninguna barrera financiera, legal o técnica, fuera de las que son inseparables de las que implica acceder a Internet mismo. La única limitación en cuanto a reproducción y distribución y el único rol del copyright en este dominio, deberá*

[1] European Organization for Nuclear Research https://home.cern

[2] Sincrotrón ALBA https://www.cells.es

ser dar a los autores el control sobre la integridad de sus trabajos y el derecho de ser adecuadamente reconocidos y citados"[3].

Un año después el término *Open Science* fue acuñado por el economista Paul David en un intento de describir las propiedades de los bienes científicos generados por el sector público, considerando el conocimiento científico creado por la investigación pública como un bien público[1].

En el año 2004, los ministros de ciencia y tecnología de los países de la OECD se reúnen en París para discutir las directrices sobre el acceso a los datos de la investigación científica, instando a los países miembros a "*desarrollar un conjunto de directrices basadas en principios comúnmente acordados para facilitar un acceso óptimo y rentable a los datos de investigación digital procedentes de la financiación pública*" [4].

De esta manera se sientan las bases para promover el acceso a la producción científica financiada por el sector público, abarcando los dos pilares que conforman el conocimiento científico, la literatura resultante de la actividad científica y los datos que sirven como soporte al razonamiento, discusión o al cálculo. Estos datos son tanto los obtenidos en observaciones, estudios y mediciones de instrumentos, como los resultados de experimentos, pasando por los generados en simulaciones o análisis computacionales, todos ellos usados como base de la investigación científica y la posterior validación de sus resultados[2].

2.1. Beneficios y motivaciones

La publicación de los datos en abierto conlleva una serie de beneficios políticos, sociales, económicos, operacionales y técnicos[3], como serían la preservación de los datos, su reutilización por otros investigadores, el aprovechamiento de su potencial científico, la mayor visibilidad del proveedor, además de ser un estímulo para la innovación, evitar la adquisición repetida de los mismos datos eludiendo su duplicación y los costes asociados, optimizar los procesos administrativos, permitir la creación de nuevos datos basándose en la combinación de los existentes, facilitar la validación de los datos por parte de agentes externos y ofrecer sostenibilidad a los datos evitando su pérdida.

El informe final para la Comisión Europea del *High level Expert Group on Scientific Data*[4], que pretende ser una referencia para las futuras inversiones en investigación en la UE, analiza los beneficios según grupos de interés, incluyendo a la ciudadanía, los financiadores y responsables de las políticas, los investigadores y la empresa e industria.

Por otro lado, Birgit et al. en su encuesta en el *Belmont Forum's Open Data*[5], con 1330 participantes, identifica por orden de importancia las motivaciones de los asistentes para publicar los datos científicos en abierto:

1. La mejora en la aceleración de la investigación científica y sus aplicaciones.
2. El interés en la diseminación y reconocimiento de los resultados de la investigación.
3. El compromiso personal de abrir los datos.
4. Las peticiones de los usuarios de datos.
5. El cumplimiento de las políticas del financiador de la investigación, la sociedad científica, la institución o los editores.

2.2. Normativa

España está considerada como uno de los países pioneros y más avanzados en cuanto a políticas de publicación de datos por parte de las administraciones públicas, llegando a situarse como líder en la UE en cuanto a madurez y preparación[6], y en la posición decimotercera a nivel mundial[7].

Estos buenos resultados no se repiten en el campo de la investigación. Una razón podría ser que la promoción de la apertura de los datos en las administraciones públicas, no se traslada a las instituciones dedicadas a la investigación. La Ley 37/2007 sobre reutilización de la información del sector público, que dispone un marco general mínimo para las condiciones de reutilización de los documentos del sector público, en su artículo 3 excluye de su aplicación a los datos producidos en centros de investigación:

"*Los documentos producidos o conservados por instituciones educativas y de investigación (incluidas las organizaciones para la transferencia de los*

[3] Budapest Open Access Initiative http://www.budapestopenaccessinitiative.org/read

[4] Declaration on access to research data from public funding http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157

resultados de la investigación, centros escolares y universidades, exceptuando las bibliotecas universitarias) así como los museos y archivos estatales como agentes de ejecución del Sistema Español de Ciencia, Tecnología e Innovación siempre que sean resultado de una investigación" .

En cambio la UE en su programa Horizonte 2020, la mayor dotación económica para la investigación e innovación en la historia de la UE con casi 75.000 millones de euros durante el período 2014 a 2020, obliga a la publicación en abierto de las publicaciones de los proyectos de investigación financiadas por este programa, y a partir del año 2017 el acceso en abierto a los datos de la investigación se define como la opción por defecto. A pesar de que permite a los beneficiarios de los fondos liberarse de esta obligación, alienta el libre acceso y la reutilización de los datos en la investigación[8].

Esto provoca que a nivel nacional exista una ciencia de dos velocidades en lo referente a la publicación de los datos en abierto, por un lado los proyectos de investigación con financiación europea donde la tendencia es la obligación de publicar los datos en abierto, y por otro lado los proyectos nacionales, donde la publicación de los datos en abierto es a voluntad del investigador.

2.3. Ciclo de publicación de los datos

En un escenario de publicación de datos abiertos Zuiderwijk[9] identifica varios elementos clave. En una primera fase, una organización pública crea y reúne los datos, almacenándolos internamente. Si la organización decide publicar los datos, es necesario que los compruebe, seleccione y organice, eliminado la información sensible de carácter personal. Después la organización publicará estos datos en Internet utilizando un sistema de publicación de datos sobre un portal de datos abiertos. En una segunda fase, con los datos disponibles para los usuarios finales, debe ser posible localizar estos datos, a través de los portales o catálogos de datos abiertos, para su visualización o descarga, y la aplicación de herramientas que permitan el filtrado, el análisis, su enriquecimiento y combinación con otros conjuntos de datos.

La primera fase es la que abarca los procesos que componen el ciclo de publicación de datos, pero antes de definir este ciclo, las organizaciones deben crear una estructura de gestión de datos que defina y gobierne los objetivos de la publicación de los datos en abierto, alineándolos con los de la organización. Este proceso formal debería incluir[10]:

- La identificación de los datos maestros susceptibles de ser publicados como datos abiertos.
- La identificación de los sistemas fuente donde los datos y los metadatos son producidos.
- La recopilación de los metadatos subyacentes en los datos maestros.
- El nombramiento de administradores de datos con experiencia tanto en los sistemas fuente actuales como en Open Data.
- La elaboración de un programa de gestión de datos que defina cómo, dónde y con qué definición se establecen los datos maestros.
- El nombramiento de un consejo de gestión de datos que decida qué procedimiento de normalización se utilizará en los datos maestros.
- El desarrollo de un modelo de datos tanto lógico como físico.
- El diseño de una infraestructura de apoyo que permita la publicación automática de los datos en abierto, implementando procesos de extracción, transformación y publicación de los datos (ETL).
- La generación y prueba de los datos maestros, que deberán ser verificados en cuanto a su calidad y consistencia con inspecciones manuales o automáticas.
- La implementación de procesos de mantenimiento, que mantengan la calidad de los metadatos y de las funcionalidades ETL.

Los conceptos técnicos que comprenden el ciclo de publicación de datos en abierto se dividen en una serie de subprocesos[11] descritos a continuación.

Recopilación de los datos

El ciclo de publicación empieza creando una visión general de los datos que están disponibles en la organización, priorizando aquellos susceptibles de ser publicados. Para cada conjunto de datos es deseable identificar su potencial para ser publicado en abierto, su información organizacional, legal

y técnica, junto con una evaluación sobre su valor, diferenciando los conjuntos de datos que proactivamente se publicaran y aquellos que lo harán bajo demanda.

Preparación de los datos

Los datos en bruto normalmente no son publicables y es necesario realizar un conjunto de acciones para su preparación.

En lo referente al contenido, se valoran aspectos como la exhaustividad de los datos y su información, las versiones, su origen y estado, también la limpieza de los datos en sí, evitando valores vacíos o erróneos, y su exactitud. En cuanto a la temporalidad, si los datos son cambiantes, se valora que los datos se actualicen periódicamente. Respecto la consistencia, es indispensable el uso de estándares y la coherencia en la publicación de conjuntos de datos de igual calidad y propósito.

Existen aspectos técnicos que mejoran la interoperabilidad y el descubrimiento de los conjuntos de datos, como *Linked Data*[12], que es el uso de la Web para crear enlaces entre datos de diferentes fuentes, por ejemplo enriqueciendo un conjunto de datos enlazando sus metadatos. Para este propósito se utilizan técnicas como las ontologías de vocabularios, Resource Description Framework (RDF) y Unique Resource Identifier (URI).

Los metadatos proporcionan información adicional que ayuda a los consumidores de datos a comprender mejor su significado, su estructura y aclarar otros temas, como los derechos y licencias, la organización que los generó, su calidad y los métodos de acceso[13]. Las prácticas recomendadas en este sentido por el World Wide Web Consortium (W3C)[5] son:

- Definir metadatos tanto para usuarios humanos como para aplicaciones informáticas. En el caso de los humanos se pueden incluir como parte de la página HTML o en un fichero de texto separado. Para que los metadatos pueden ser interpretados por máquinas, deben proveerse en un formato serializado como Turtle o JSON, o embebido en el código de la página HTML usando RDF o JSON-LD[6]. Es esencial el uso de términos y vocabularios estándares como el Dublin Core Metadata (DCMI)[7] y el Data Catalog Vocabulary (DCAT)[8], o vocabularios específicos adaptados a la naturaleza de los datos.

- Incluir metadatos que describan las características generales de los conjuntos de datos, como podría ser el título, la descripción, las palabras claves, la fecha de publicación, la organización responsable, la cobertura espacial, el período temporal, la fecha de la última modificación, los temas y categorías cubiertos. Se recomienda el uso de tesauros multilingües para asegurar la interoperabilidad internacional, como por ejemplo EuroVoc[9].

- Los metadatos que describen el esquema y la estructura interna del conjunto de datos son esenciales para su exploración y consulta, ayudando a la compresión del significado de los datos. En formato humano normalmente se describen las propiedades o columnas del esquema del conjunto de datos. En formato máquina, se puede optar por proveer esta información en documentos separados o embebido en el mismo documento, mediante diferentes tecnologías como Tabular data[10], JSON-LD, XML[11] o Multi-dimensional data[12].

Publicación de los datos

Una vez finalizada la preparación de los datos, estos están listos para su publicación en Internet. Esta fase es específica para cada organización y depende tanto de las características de los datos como de los recursos disponibles. Cuando son pocos los conjuntos de datos, es posible publicar los datos como ficheros en una web, pero al aumentar el número de datos, la opción más habitual es la utilización de un portal o repositorio de datos. El uso de una API que permita automatizar la consulta y el acceso a los datos dependerá del tipo de software usado, siendo necesario publicar también su especificación.

[5] W3C https://www.w3.org/
[6] JSON-LD 1.0 https://www.w3.org/TR/json-ld/
[7] DCMI Metadata Terms http://dublincore.org/documents/dcmi-terms/
[8] W3C Data Catalog Vocabulary https://www.w3.org/TR/vocab-dcat/
[9] EuroVoc http://eurovoc.europa.eu/
[10] Tabular Data Models https://www.w3.org/TR/tabular-data-model/
[11] XML Schema https://www.w3.org/XML/Schema
[12] RDF Data Cube Vocabulary https://www.w3.org/TR/vocab-data-cube/

Cuando los conjuntos de datos ocupan grandes cantidades de espacio, del orden de Tera Bytes, su descarga y análisis conlleva un uso importante de recursos difícil de asimilar por ciertas organizaciones o particulares. Para facilitar la reutilización de estos datos, una tendencia es permitir el acceso a servidores llamados Cloud Data Servers donde los datos están disponibles, ofreciendo herramientas para su consulta y tratamiento sin la necesidad de su descarga.

Otra tendencia para paliar la problemática de los grandes conjuntos de datos es el uso de servidores de datos, como por ejemplo HDF Server[13], que permiten el acceso a porciones de los datos, en lugar de obligar a la descarga completa de todo el conjunto de datos.

Mantenimiento de los datos

Los datos históricos permanecen estables, mientras que los datos recientes pueden cambiar frecuentemente, por tanto es recomendable un proceso que mantenga actualizados los datos y los metadatos regularmente, revisando los URI, las URL y la respuesta de los usuarios, en un proceso de mejora continua.

Asegurar y supervisar el uso de los datos

Para asegurar el éxito de la iniciativa de publicación de datos, es necesario involucrar a los potenciales usuarios y monitorizar varios factores. Tim Davies[14] identifica 5 aspectos clave para promover el uso de los datos:

1. Basar la publicación de los datos en la demanda y necesidades de la comunidad, proporcionando canales para recibir estas peticiones.

2. Poner los datos en contexto, proporcionando información sobre la descripción de los datos, la frecuencia de las actualizaciones, el formato, la calidad, como se crearon, manuales de uso o herramientas, y vinculando información sobre análisis de los datos previamente realizados.

3. Apoyar el debate entorno a los datos, creando conversaciones estructuradas con los usuarios sobre los datos facilitando el contacto con los publicadores.

4. Desarrollar capacidades, habilidades y redes, proporcionado o enlazando herramientas que permitan trabajar sobre los datos, creando guías de como usar los datos e impartiendo o patrocinando sesiones de formación sobre el uso de los datos.

5. Colaborar en los datos como en un recurso común, facilitando canales para que los usuarios puedan ayudar en la mejora de los datos, colaborando con la comunidad para crear nuevos conjuntos de datos derivados, apoyando el desarrollo o mantenimiento de herramientas y servicios útiles para los datos, y apoyándose en otras organizaciones para conectar las fuentes de datos.

En cuanto a la supervisión del uso de los datos, se debe considerar la utilización de métricas que permitan evaluar varios indicadores como el rendimiento de los datos contabilizando el número de descargas, el rendimiento del sistema comprobando que el sistema pueda soportar la carga de peticiones o si existen periodos de falta de servicio, y el rendimiento de la recopilación y preparación de los datos evaluando la respuesta de los usuarios.

2.4. La publicación de datos en el ámbito científico

La publicación en abierto de los datos provenientes de la investigación posee un factor diferencial con el resto de ámbitos, y es la gran variedad de estándares entre las diferentes disciplinas. El grupo de trabajo Publishing Data Workflows[15] de la Research Data Alliance (RDA)[16] y de la World Data System (WDS)[17] propuso estudiar el panorama actual del ciclo de publicación de datos entre disciplinas e instituciones, examinando un conjunto diverso de flujos de trabajo para identificar componentes y prácticas comunes.

Los resultados de este examen se han utilizado para definir un modelo de referencia de publicación de datos científicos que comprende componentes genéricos[14]. En este modelo se identifican varios flujos de trabajo:

[13] HDF Server https://support.hdfgroup.org/projects/hdfserver/

[14] Five Stars Open Data Engagement http://www.opendataimpacts.net/engagement/

[15] Publishing Data Workflows Working Group https://www.rd-alliance.org/groups/rdawds-publishing-data-workflows-wg.html

[16] Research Data Alliance https://www.rd-alliance.org/

[17] ICSU World Data System http://www.icsu.org/what-we-do/interdisciplinary-bodies/wds/

- La publicación tradicional de un artículo, donde los datos y metadatos asociados se solicitan al investigador o se descargan de un sitio web con el soporte del investigador.

- La publicación de una investigación reproducible, estos procesos apoyan una mayor reproducibilidad en la investigación e incluyen alguna forma de publicación de datos.

- Los dos flujos de trabajo de publicación de datos emergentes predominantes: la presentación de un conjunto de datos en un repositorio, y la presentación de un artículo de datos a una revista de datos. Ambos flujos de trabajo requieren que los conjuntos de datos se envíen a un repositorio de datos.

Este modelo define una serie de componentes claves, elementos que son necesarios para constituir una publicación de datos científicos como: la entrada de repositorio con un identificador persistente, la generación y revisión de los metadatos, la distribución y el descubrimiento de los datos, y los procesos de curación para la creación controlada de los datos, su mantenimiento y gestión.

Además, el modelo define componentes opcionales como servicios y funciones. Por una lado se encuentran los que mejoran el contexto, por ejemplo con una documentación enriquecida o con enlaces al artículo con los resultados, al artículo de datos, y a programas, código o simulaciones utilizados. También cuentan entre estos servicios los procesos para garantizar y controlar la calidad de los datos, el proceso de su edición, la curación de los datos mejorada por expertos, su evaluación por parte de colegas en la materia, y el apoyo a la presentación de los datos. Por último, cabe enumerar los servicios que permiten mejorar la visibilidad y accesibilidad a los datos: la indexación, la legibilidad por parte de programas informáticos, el soporte en el acceso y en servicios de valor añadido, y las plataformas de gestión de la evaluación.

2.5. Licencias

Si los datos no están legalmente abiertos, no existe el derecho para que puedan ser reutilizados. La apertura legal es uno de los principios básicos de Open Data, y cada conjunto de datos debe acompañarse de una licencia.

Si alguien desea usar el trabajo de otra persona, es necesario disponer del permiso del poseedor del trabajo. Las licencias son la vía para otorgar explícitamente a alguien el permiso para usar ese trabajo. En Europa existen dos tipos de derechos sobre la propiedad intelectual que se reconocen automáticamente en una creación[18]:

- El *copyright* sobre el contenido original creado, como el texto y las fotografías.

- El *database right* sobre colecciones de datos que han supuesto un esfuerzo sustancial en su obtención, verificación o presentación.

Estos derechos se mantienen aunque el trabajo se publique en Internet, debiéndose obtener el permiso para su reutilización.

La directiva 96/9/EC que recoge estos derechos no es de obligado cumplimiento para los países miembros, y su aplicación puede variar de unos países a otros. Para evitar estas ambigüedades, lo recomendable es que los datos vayan acompañados por una licencia otorgada por el propietario y que establezca los derechos de quien desee reutilizar estos datos. Existen licencias específicas de obligado uso, o recomendadas en algunas instituciones, o bajo algunos tipos de financiamientos. También pueden confeccionarse licencias a medida, pero están disponibles una serie de licencias estándar utilizables en la publicación de datos de investigación en abierto[15], agrupadas bajo dos instituciones. **Creative Commons**[19] propone 6 tipos de licencias para los trabajos creativos, permitiendo un control ajustado sobre su uso.

- *CC BY 4.0* permite copiar, redistribuir y adaptar el material en cualquier formato o medio, para cualquier finalidad, siempre que se reconozca la autoría.

- *CC BY-SA 4.0* permite copiar, redistribuir y adaptar el material en cualquier formato o medio, para cualquier finalidad, siempre que se reconozca la autoría y el nuevo material se difunda con la misma licencia que el original.

- *CC BY-ND 4.0* permite copiar y redistribuir el material en cualquier formato o medio, para

[18] Publisher's Guide to Open Data Licensing http://theodi.org/guides/publishers-guide-open-data-licensing

[19] Creative Commons http://creativecommons.org/

cualquier finalidad, siempre que se reconozca la autoría y el material original no sea modificado.

- *CC BY-NC 4.0* permite copiar, redistribuir y adaptar el material en cualquier formato o medio, siempre que se reconozca la autoría y no se utilice con finalidad comercial.
- *CC BY-NC-ND 4.0* permite copiar, redistribuir y adaptar el material en cualquier formato o medio, siempre que se reconozca la autoría, no se utilice con finalidad comercial y el nuevo material se difunda con la misma licencia que el original.
- *CC BY-NC-SA 4.0* permite copiar y redistribuir el material en cualquier formato o medio, siempre que se reconozca la autoría, no se utilice con finalidad comercial y el material original no sea modificado.

Otras licencias disponibles son las definidas por el proyecto **Open Data Commons**[20], actualmente de la Open Knowledge International[21], estas licencias siguen el modelo de Creative Commons pero son específicas para bases y conjuntos de datos.

- *ODC-By* permite copiar, distribuir y usar los datos para crear nuevos materiales o su modificación y su transformación para cualquier propósito. Si se produce un nuevo material, se debe acompañar de una explicación sobre los datos utilizados para su creación. Si los datos forman parte sustancial de una nueva base de datos o colección de datos, la referencia a esta licencia o su texto deben distribuirse con los nuevos datos.
- *ODC-ODbL* igual que la licencia ODC-By pero con algunas condiciones adicionales. Si se crea una nueva base de datos o colección derivada de la licenciada, debe incluir una licencia igual o similar a la ODC-ODbL. Solamente se pueden incluir restricciones tecnológicas, como *Digital Rights Management*, si existe una versión disponible de los datos sin esta restricción.

Licenciar un trabajo bajo **Dominio Público** es la forma más permisiva de distribuir una creación, permitiendo un uso de los datos tan libre como sea posible. Existen dos licencias de este tipo:

- *CC0 1.0*[22] el propietario renuncia a los derechos sobre la obra. Permite copiar, modificar, distribuir e interpretar la obra, incluso con fines comerciales y sin solicitar permiso.
- *PDDL*[23] renuncia a los derechos igual que en la CC0, pero esta licencia es específica para datos, colecciones o bases de datos.

La licencia escogida debe acompañar a los datos en forma de declaración, y facilitar un mecanismo para el acceso al texto completo de la licencia, de forma que cualquier persona que acceda a los datos vea de forma inequívoca su licenciamiento.

Uno de las premisas de Open Data es la interoperabilidad, permitiendo que la lectura de los datos abiertos pueda automatizarse programáticamente. En este sentido, cuando la licencia usada puede identificarse mediante una URL estándar como las anteriores, es posible utilizar Resource Description Framework (RDF)[24] para insertar en el código la licencia, por ejemplo en XML o HTML, y que esta sea reconocida por las herramientas de automatización.

2.6. Portales

La opción más utilizada para la publicación de los datos en abierto es el uso de portales Web. Existen una serie de funciones comunes en este tipo de repositorios[16] como la de ayudar a los usuarios a encontrar los datos abiertos que necesitan, la de garantizar que los datos accedidos siguen siendo relevantes, útiles y utilizables; la de supervisar y mejorar la calidad y la entrega de los datos, y la de mantener el ritmo de las tecnologías y servicios de datos junto con las necesidades de los usuarios a medida que evolucionan.

Una organización que desee publicar sus datos en abierto puede optar por un software que le permita gestionar su propia plataforma de portal de datos, ya sea con infraestructura propia o en la nube, o puede optar por utilizar una plataforma gestionada por terceros. En el primer caso, para gestionar un portal de datos propio, es posible elegir entre

[20] Open Data Commons https://opendatacommons.org/

[21] Open Knowledge International https://okfn.org/about/

[22] CC0 https://creativecommons.org/publicdomain/zero/1.0/

[23] PDDL https://opendatacommons.org/licenses/pddl/

[24] RDF 1.1 Primer https://www.w3.org/TR/rdf11-primer/

alternativas de software libre como CKAN[25] o Invenio[26], o alternativas de software propietario como OpenDataSoft[27] o Socrata[28]. En lo referente a portales gestionados por terceros, los cuales son compartidos con otros usuarios e instituciones, se pueden encontrar iniciativas privadas como Junar[29] y Figshare[30], o iniciativas financiadas por organismos públicos, como Zenodo[31], EUDAT[32] y el European Union Open Data Portal[33].

El Registry of Research Data Repositories (re3data)[34] es un catálogo de repositorios de datos de investigación recomendado por la Comisión Europea[2], donde es posible localizar más de un millar de repositorios por temática, tipos de contenido, países, además de otros parámetros. A fecha de 15 de abril de 2017, el catálogo muestra que España participa en 21 repositorios de investigación. Kindling et al.[17] realizan un análisis descriptivo y estadístico de la información de los metadatos de 1.381 repositorios de investigación basándose en los datos de re3data. Los resultados dan la distribución de repositorios de investigación por países mostrada en la figura 1.

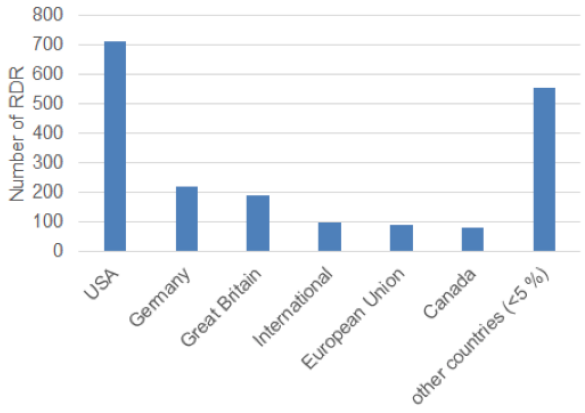

Figura 1: Repositorios de investigación indexados en re3data por países (n=1.381)[17].

El catálogo re3data utiliza la clasificación temática de la German Research Foundation[35]. El diagrama de Venn de la figura 2 muestra la distribución de los repositorios de investigación de acuerdo a las cuatro categorías principales "Humanidades y Ciencias Sociales (SSH)", "Ciencias de la Vida", "Ciencias Naturales" e "Ingenierías", teniendo en cuenta que un repositorio puede ofrecer datos de más de una temática.

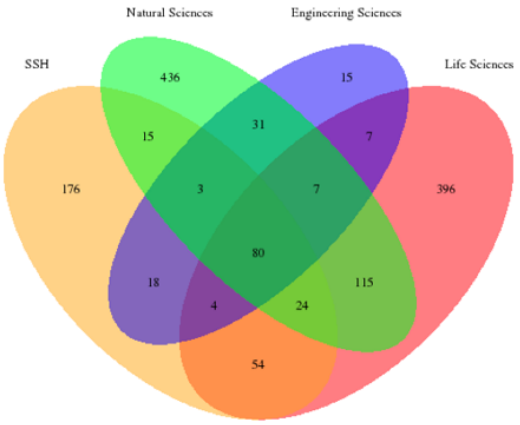

Figura 2: Repositorios de investigación indexados en re3data por temática (n=1.381)[17].

En lo referente a las licencias encontradas en los repositorios indexados en re3data.org, la tabla 1 muestra su distribución, teniendo en cuenta que son posibles múltiples valores en los repositorios.

Research data license	*Count (n=1895)*	*Percentage (n=1381)*
Other	790	57,2
Copyrights	553	38,6
CC	301	21,8
Public Domain	196	14,2
ODC	27	2,0
CC0	26	1,9
OGL	12	0,9
BSD	6	0,4
Apache License 2.0	2	0,1
OGL C	1	0,1
RL	1	0,1

Tabla 1: Tipos de licencia en los repositorios de investigación indexados en re3data (n=1.895)[17].

En cuanto a los tipos de contenido de los repositorios indexados en re3data.org mostrados en la tabla 2, las mayores apariciones son las de formatos científicos y estadísticos, seguidos por formatos

[25] CKAN https://ckan.org/
[26] Invenio http://invenio-software.org/
[27] OpenDataSoft https://www.opendatasoft.com/open-data-solutions/
[28] Socrata https://socrata.com/
[29] Junar http://junar.com
[30] Figshare https://figshare.com/
[31] Zenodo https://zenodo.org/
[32] EUDAT https://eudat.eu/
[33] European Union Open Data Portal http://data.europa.eu/euodp/en/data
[34] Registry of Research Data Repositories http://www.re3data.org/
[35] Deutsche Forschungsgemeinschaft http://www.dfg.de/en/

ofimáticos, texto plano, imágenes, datos en bruto, gráficos y texto estructurados.

Content type	*Count (n=6340)*	*Percentage (n=1381)*
Scientific and statistical formats	881	63,8
Standard office documents	786	56,9
Plain text	690	50,0
Images	686	49,7
Raw data	586	42,4
Structured graphics	541	39,2
Structured text	490	35,5
Other	446	32,3
Archived data	339	25,5
Software applications	258	18,7
Audiovisual data	253	18,7
Databases	204	14,8
Networkbased data	104	7,5
Source code	49	3,5
Configuration data6	27	2,0

Tabla 2: Tipos de contenido de los repositorios de investigación indexados en re3data[17].

2.7. Barreras

Son varias las barreras que impiden la adopción y un amplio despliegue de la publicación de los datos en abierto. En el año 2012 Janssen et al.[3] llegaban a identificar más de 50 motivos, categorizándolos en institucionales, por complejidades, de uso y participación, legislativos, por la calidad de la información, o técnicos. El informe *Open Data Maturity in Europe 2016*[6] muestra que se ha progresado en este sentido respecto a años anteriores, aunque todavía identifica ciertas barreras.

Políticas

Un tercio de los países de la Unión Europea se enfrentan a barreras de este tipo. Los políticos no son completamente conscientes de los beneficios de la publicación de datos en abierto y por tanto no lo consideran una prioridad, si no como una funcionalidad complementaria. El patrocinio de los datos abiertos al más alto nivel Europeo podría conducir a un aumento de la publicación de datos en todos los niveles de gobierno, creando así presión en las administraciones políticas. La directiva europea PSI[36] va en este sentido, haciendo obligatoria la publicación de la información del sector público para su reutilización.

Legales

Es necesario un marco legal para publicar los datos en abierto que vaya más allá de la directiva PSI, y que incluya definiciones, roles y responsabilidades. Una segunda barrera legal es la concerniente a las licencias de los datos, en algunos países estas licencias no están reguladas o existen licencias solamente de ámbito nacional. Por último, algunos países poseen leyes de privacidad que impiden la apertura de los datos, la *General Data Protection Regulation* (GDPR)[37] que entrará en vigor el 25 de Mayo de 2018 armonizará los derechos de privacidad de todos los países miembro adaptándolos a la era digital.

Técnicas

Se publican datos de poca calidad, desestructurados o no automatizables, que dificultan su reutilización, denotando una falta de estandarización en la recolección de los datos, en su publicación y en su formato, así como en los metadatos asociados.

Financieras

Algunas organizaciones reciben ingresos por la venta de sus datos, y al publicarlos en abierto perderán estos ingresos. Los beneficios de esta apertura no están claramente documentados dificultando la justificación de la pérdida de ingresos. También es un inconveniente el gasto considerable en personal y equipamientos dedicados a esta iniciativa.

Comunicativas

Los ciudadanos, empresas y demás usuarios potenciales desconocen los conjuntos de datos disponibles, ni los posibles beneficios que conlleva su reutilización.

Ámbito científico

En investigación existen barreras adicionales, Sá y Grieco[18] recopilan impedimentos de varios autores, entre los que se hayan la protección de los datos por parte de investigadores para asegurarse el control de los hallazgos científicos y conclusiones derivadas de estos datos. También destacan la resistencia al cambio ante la implantación de nuevos sistemas, así como desafíos operacionales, ya que la comunicación de datos científicos o estadísticos de

[36] Directive 2013/37/EU http://data.europa.eu/eli/dir/2003/98/2013-07-17

[37] General Data Protection Regulation http://www.eugdpr.org/

forma comprensible para el público en general no es una tarea sencilla, requiriendo a veces de intermediarios que ayuden a interpretar esta información y a diseñar modelos interpretables por una audiencia general.

El informe sobre la European Cloud Initiative[19], que destaca el potencial de los datos como un factor clave de Open Science y de la cuarta revolución industrial, reconoce 5 razones por las que todavía no se está aprovechando plenamente el potencial de los datos científicos:

1. Los datos procedentes de la investigación financiada no siempre se publican en abierto, uno de los motivos es la falta de una estructura clara de incentivos y recompensas para la compartición de los datos.

2. La falta de interoperabilidad evita abordar los grandes desafíos sociales que requieren un intercambio de datos eficiente y un enfoque multidisciplinario y multiactivo. El intercambio de datos de investigación también se ve obstaculizado por el tamaño de los conjuntos de datos, sus formatos variados, la complejidad del software necesario para su análisis y los muros entre las disciplinas.

3. La fragmentación dificulta la ciencia basada en datos. Las infraestructuras de datos están divididas por dominios científicos y económicos, por países y por modelos de gobierno. Las políticas de acceso para la creación de redes, el almacenamiento de datos y la informática difieren, y las infraestructuras informáticas y de datos desconectadas y lentas obstaculizan el descubrimiento científico, crean silos y frenan la circulación del conocimiento.

4. Existe una creciente demanda en Europa de una infraestructura de alto rendimiento (HPC) de nivel mundial para procesar datos en ciencia e ingeniería que necesitan capacidades de computación exascale. La industria europea provee el 5 % de los recursos HPC del mundo, mientras que consume un tercio de ellos. Ningún estado miembro, por sí solo, dispone de los recursos financieros para desarrollar el ecosistema de HPC necesario para competir con las actuales potencias mundiales.

5. Los productores y usuarios de datos científicos deben ser capaces de reutilizar datos y usar técnicas avanzadas de análisis en un entorno que sea al menos tan fiable como sus propias instalaciones. Toda utilización y reutilización de datos científicos debe garantizar que los datos personales estén adecuadamente protegidos de conformidad con las normas de la UE en materia de protección de datos.

3. El caso del CERN

En el laboratorio CERN, Organización Europea para la Investigación Nuclear[38], se investiga la estructura fundamental del universo utilizando los instrumentos científicos más grandes y complejos del mundo para estudiar los componentes básicos de la materia: las partículas fundamentales. Las partículas se hacen colisionar a una velocidad cercana a la de la luz. El proceso da a los físicos pistas sobre cómo interactúan las partículas, y proporciona información sobre las leyes fundamentales de la naturaleza. Fundado en 1954, el laboratorio del CERN se ubica en la frontera franco-suiza cerca de Ginebra y fue una de las primeras empresas conjuntas de Europa.

El Large Hadron Collider (LHC) es el acelerador de partículas más grande y potente del mundo. Se inició por primera vez el 10 de septiembre de 2008, y sigue siendo la última incorporación al complejo acelerador del CERN. El LHC consiste en un anillo de 27 kilómetros de imanes superconductores con una serie de estructuras de aceleración para aumentar la energía de las partículas. Dentro del acelerador, dos haces de partículas de alta energía viajan cerca de la velocidad de la luz antes de que se hagan colisionar en cuatro lugares alrededor del anillo del acelerador, que corresponden a las posiciones de cuatro detectores de partículas: ATLAS, CMS, ALICE y LHCb.

En su centro de datos se procesan aproximadamente un petabyte de datos todos los días y alberga 11.000 servidores con 100.000 núcleos de procesador. Este centro de datos, junto con el del *Wigner Research Centre for Physics*[39] de Budapest, forman el *Tier 0*, el primero de cuatro niveles donde se procesan, almacenan y analizan todos los datos del LHC, representando el 20 % de la capacidad total de cómputo.

[38] CERN https://home.cern/

[39] Wigner Research Centre for Physics http://wigner.mta.hu/en/

El CERN aborda la publicación de datos en abierto desde dos perspectivas, una ofreciendo Zenodo[40] como un portal público multidisciplinar, y la otra mediante el portal privado CERN Open Data[41] específico para sus experimentos en el LHC, que obtienen unos conjuntos de datos de gran tamaño. Esta segunda aproximación le permite ofrecer en este portal específico herramientas de valor añadido para el análisis y tratamiento de los datos.

3.1. Repositorio Zenodo

El proyecto OpenAIRE[42], a la vanguardia de los movimientos Open Access y Open Data en Europa, fue encargado por la Comunidad Europea para apoyar su naciente política de datos abiertos, ofreciendo un repositorio global para la investigación financiada con sus fondos públicos. El CERN, socio de OpenAIRE proporciona Zenodo, lanzándolo en mayo de 2013 y abriéndolo a la comunidad. De esta manera, se posibilita que cualquier investigador, independientemente de su nación, financiación o disciplina, acceda a las herramientas y recursos necesarios para compartir y publicar datos o software en abierto.

Zenodo no impone ningún requisito de formato, restricciones de acceso o licencia. Los datos, el software y otros artefactos en apoyo de las publicaciones son su núcleo, pero igualmente son permitidos los materiales asociados con las conferencias, los proyectos o las mismas instituciones, todos necesarios para comprender el proceso académico. Ofrece la posibilidad de alojar contenido cerrado y restringido, para que los datos puedan ser capturados y almacenados de forma segura mientras la investigación está en curso, de tal manera que no falte nada cuando más tarde se comparten abiertamente en el flujo de trabajo de investigación.

Para ayudar en el proceso de publicación, los materiales de investigación se pueden subir con seguridad a Zenodo en registros restringidos, posteriormente los enlaces protegidos pueden ser compartidos con los revisores de las publicaciones. El contenido también puede ser embargado y abierto automáticamente cuando se publica el artículo asociado.

[40] Zenodo https://zenodo.org/
[41] CERN Open Data http://opendata.cern.ch/
[42] OpenAIRE https://www.openaire.eu/

Para soportar todos estos casos de uso, se proporciona una interfaz web complementada con una API que permite a herramientas y servicios de terceros utilizar Zenodo como *backend* en su flujo de trabajo. Zenodo está basado en el framework de librería digital Invenio[43], y se ejecuta completamente sobre productos de software libre.

Contenido

Zenodo acepta cualquiera disciplina y cualquier artefacto de apoyo a la investigación que no viole la privacidad o los derechos de autor, y cualquier estado de los datos de la investigación durante su ciclo de vida. Cualquier usuario puede depositar contenido, y su subida al repositorio no altera su propiedad, debiendo especificar el tipo de licencia de cada fichero. Cualquier formato de dato está permitido, y el tamaño máximo por registro es de 50 GB, aunque tamaños superiores pueden solicitarse. Los metadatos se guardan en formato JSON y se pueden exportar como MARCXML, Dublin Core, y DataCite Metadata Schema.

Acceso y reutilización

El acceso a los archivos puede definirse como abierto, embargado, restringido o cerrado, su uso y reutilización está sujeto a la licencia definida al depositar los objetos. En el estado de embargo, se proporciona una fecha a partir de la cual el contenido se publicará automáticamente. El acceso a archivos restringidos solo es posible mediante la aprobación del depositante del archivo original. El acceso a los metadatos se realiza mediante protocolos estándar como HTTP y OAI-PMH. Estos están autorizados bajo licencia CC0, excepto para las direcciones de correo electrónico.

Eliminación

El contenido que se considera que no es del ámbito del repositorio se elimina y las DOIs asociadas emitidas por Zenodo se revocan, idealmente antes de 24 horas. La retirada del objeto de investigación se considera una acción excepcional que debe ser solicitada y justificada por el remitente original. El DOI y la URL del objeto original se conservan, y en su lugar se sirve una página con el motivo de la retirada.

[43] Invenio http://inveniosoftware.org/

Longevidad

Los archivos de datos se pueden versionar, quedando el contenido original inmutable, en cambio los registros no se versionan. Los archivos de datos y metadatos se almacenan en los centros de datos del CERN y se guardan múltiples réplicas de los archivos de datos en un sistema de archivos distribuido, comprobándose regularmente contra la suma MD5 del original, para asegurar que su contenido permanece inalterado. Los datos se conservarán durante toda la vida útil del repositorio, que es el tiempo de vida del laboratorio CERN, actualmente tiene un programa experimental definido mínimo durante los próximos 20 años.

Estadísticas de uso

A fecha de 7 de mayo de 2017, Zenodo presenta 197.636 registros de objetos de investigación. Estos registros se clasifican como 179.975 de acceso abierto, 17.169 de acceso cerrado, 270 restringidos y 222 embargados. Según el tipo de documento, los registros se distribuyen como muestra la tabla 3.

Tipo de documento	Número
Imagen	110068
Publicación	63826
Software	15307
Conjunto de datos	4836
Presentación	2164
Póster	947
Vídeo	293
Lección	195

Tabla 3: Número de registros en Zenodo según el tipo de documento.

3.2. Repositorio CERN Open Data

El portal CERN Open Data es el punto de acceso a los datos producidos a través de la investigación realizada en el CERN, difundiendo y preservando la producción de diversas actividades de investigación, incluyendo el software y la documentación necesarios para comprender y analizar los datos publicados. El portal utiliza los estándares mundiales establecidos en la preservación de los datos y en Open Science, compartiendo los datos bajo licencias abiertas y otorgando un identificador único de objeto digital (DOI) para su referencia.

El repositorio utiliza diversas tecnologías, entre ellas el framework de librería digital Invenio[44], CernVM[45] que proporciona un entorno de usuario para desarrollar y ejecutar análisis de datos LHC en local o en la nube independientemente del sistema operativo, y EOS[46] que ofrece una infraestructura de almacenamiento de baja latencia y altamente escalable.

Los datos producidos por los experimentos del LHC se clasifican generalmente en cuatro niveles diferentes definidos por la DPHEP[20]. Este repositorio se centra en los datos de nivel 2 y 3:

- Los datos de nivel 1 comprenden datos que están directamente relacionados con publicaciones que proporcionan documentación para los resultados publicados.
- Los datos de nivel 2 incluyen formatos de datos simplificados para análisis en ejercicios de extensión y formación.
- Los datos de nivel 3 incluyen datos reconstruidos y simulaciones, así como el software necesario para un análisis científico completo.
- Los datos de nivel 4 cubren los datos básicos en bruto, si no están cubiertos en el nivel 3, y su software asociado, permitiendo el acceso al potencial completo de los datos.

Contenido

Los cuatro experimentos del LHC: ATLAS, CMS, ALICE y LHCb; han aprobado políticas de preservación y acceso a los datos que permiten la publicación en abierto de sus datos, exceptuando los datos del nivel 4. Los nuevos datos entrarán en el portal una vez terminados sus respectivos períodos de embargo. En apoyo a estas políticas de datos, el portal publica y conserva los datos de los niveles 2 y 3, así como formatos simplificados y eventos totalmente reconstruidos, junto con el software asociado y la documentación necesaria para acceder y utilizar los datos.

Actualmente, el único experimento que ha publicado sus datos en abierto es el Compact Muon Solenoid (CMS), el resto están en periodo de embargo. El CMS es uno de los dos experimentos de propósito general en el LHC. Desde 2010 ha recogido alrededor de 28 fb^{-1} de datos de colisión de protones-protones a energías de centro de masa de hasta 8

[44] Invenio http://inveniosoftware.org/

[45] CernVM http://cernvm.cern.ch/

[46] EOS https://eos.web.cern.ch/

TeV, así como datos de colisiones de protón-plomo y plomo-plomo. El análisis de estos datos ha producido casi 400 artículos publicados que describen búsquedas de nuevos fenómenos físicos, medidas de procesos conocidos, así como el descubrimiento del bosón de Higgs[21].

Los conjuntos de datos que proporciona este portal son conjuntos de datos primarios, que son datos de colisión completamente reconstruidos, datos de simulaciones, y ejemplos de conjuntos de datos simplificados derivados de los primarios.

Este portal también ofrece herramientas que dan valor añadido a los datos, como una imagen de Máquina Virtual (VM) con el entorno de software CMS para acceder a los conjuntos de datos primarios, un ejemplo de proceso de análisis, aplicaciones en línea para mostrar eventos e histogramas y el código fuente de los diversos ejemplos y aplicaciones.

Acceso y reutilización

Los datos abiertos se publican bajo licencia Creative Commons CC0, todos los conjuntos de datos tienen un DOI único que se debe citar en cualquier publicación y ocupan aproximadamente 27TB de espacio.

El portal se basa en los éxitos previos de la publicación de datos para la educación y la divulgación, pero va más allá al incluir la posibilidad de realizar análisis más profundos y complejos con los datos de alto nivel ahora publicados. El portal se divide en las secciones de educación e investigación, a las que se asigna el material en función de su uso potencial y grado de dificultad[21].

Longevidad

El repositorio se abrió al público en noviembre de 2014 y uno de los principales objetivos de este proyecto es la preservación de estos datos, permitiendo su acceso más allá de la vida útil de los experimentos, ya que una vez desmantelado el LHC difícilmente se puedan volver a repetir estos experimentos.

Estadísticas de uso

Un mes después de su apertura, el portal había sido visitado por 82000 usuarios diferentes[21] con los tipos de acceso descritos en la tabla 4.

Tipo de acceso	Número
Descarga de ficheros por HTTP	600
Visita a colecciones	21000
Uso del visor de eventos	16000
Uso del histograma	3000

Tabla 4: Estadísticas de uso de CERN Open Data durante el primer mes

La descarga de conjuntos de datos se estimó en 1000 accesos, a través del protocolo XRootD de las máquinas virtuales, y unas 200 descargas directas desde el portal en los 3 meses posteriores.

4. El caso del Sincrotrón ALBA

La información sobre este caso ha sido proporcionada por David Fernández Carreiras, responsable de *Computing and Control Division* del Sincrotrón ALBA, a través de una entrevista realizada el día 12/05/2017 siguiendo el esquema definido por Carlson[22].

El ALBA[47] es una instalación de luz sincrotrón de tercera generación, se trata de un complejo de aceleradores de electrones para producir luz de sincrotrón, que permite visualizar la estructura atómica y molecular de los materiales y estudiar sus propiedades.

El Sincrotrón ALBA está gestionado por el Consorcio para la Construcción, Equipamiento y Explotación del Laboratorio de Luz Sincrotrón (CELLS) y cofinanciado a partes iguales por la Administración española y catalana. Pertenece a la red de Infraestructuras Científicas y Técnicas Singulares (ICTS), grandes instalaciones, recursos, equipamientos y servicios, únicas en su género, que están dedicadas a la investigación y desarrollo tecnológico de vanguardia y de máxima calidad, así como a fomentar la transmisión, intercambio y preservación del conocimiento, la transferencia de tecnología y la innovación.

Actualmente, el ALBA dispone de ocho líneas de luz operativas, que comprenden tanto los rayos X blandos como los rayos X duros, y que se destinan principalmente a las biociencias, a la materia condensada (nanociencia y propiedades magnéticas y

[47] ALBA https://www.cells.es/

electrónicas) y a la ciencia de los materiales. Se encuentran en construcción dos líneas de luz más que se destinarán a la fotoemisión de baja energía y alta resolución angular para materiales complejos, y microfocus para cristalografía de proteínas.

A día de hoy no hay ningún dato de ALBA que se publique en abierto, pero está en proceso de aprobación la nueva política de datos que implica la publicación de los datos después del período de embargo. Esta nueva política de datos ya ha sido aprobada por el Comité Científico Asesor de ALBA, y resta pendiente de la aprobación de su Consejo Rector.

Contenido

La intención de la nueva política de datos es publicar los datos de todos los niveles de la DPHEP[20] producidos por experimentos financiados con fondos públicos. En ciertas disciplinas, los datos brutos de nivel 4 ocupan una gran cantidad de espacio, en estos casos dependerá de los recursos asignados mantener estos datos abiertos, o reducir la publicación a los datos procesados o simplificados, y los directamente relacionados con las publicaciones.

El volumen de los datos generados depende de la técnica de adquisición y del detector utilizado, que van desde una adquisición continua de imágenes a 300MB/s, a la adquisición de valores escalares de 4 bytes. Los experimentos pueden ocupar desde Mega Bytes de datos a Tera Bytes.

El formato que utilizan todas las líneas es el HDF5[48] que permite representar objetos de datos complejos y una amplia variedad de metadatos, además no limita el número o el tamaño de los objetos de datos. También se utiliza el formato de texto ASCII para datos escalares, y formatos específicos como por ejemplo ESRF Data Format[49] o definidos por fabricantes de instrumentación como el XRM.

A partir de los datos obtenidos en bruto se realizan varios tratamientos de análisis y refinamiento en las instalaciones de ALBA, bien en estaciones de trabajo o bien en el clúster HPC, a partir de herramientas desarrolladas en MATLAB[50] o Python[51], o programas de terceros, obteniendo nuevos datos en diferentes formatos. El software a veces es tan específico que el propio investigador llega a desarrollar programas propios para su experimento.

Un aspecto importante a destacar es la gran importancia dada a los metadatos para la correcta definición y categorización de los datos según el experimento, la línea de luz implicada, el tipo y la preparación de las muestras, las condiciones del haz de luz, los investigadores implicados, los instrumentos utilizados, la frecuencia de adquisición, etc., que permitirán a posteriori una correcta localización y correlación de los datos mediante diferentes parámetros.

Acceso y reutilización

En la actualidad los datos no se comparten, después de que un experimento ha pasado por las fases de propuesta, selección, aprobación, asignación de tiempo y ejecución, los investigadores se llevan una copia de los datos obtenidos, bien mediante medios portátiles, o bien descargando en remoto los datos. A su vez ALBA guarda una copia de todos los datos, que pasado un tiempo se mueven de un almacenamiento basado en discos rotacionales a otro basado en cintas magnéticas.

La intención con la nueva política de datos es que después del período de embargo de 3 años, los datos estén disponibles públicamente y puedan ser reutilizados, reconociendo que los datos se han adquirido en ALBA citando la correspondiente estación experimental y sus responsables. El acceso a los datos no será inmediato y existirá una latencia debido al medio de respaldo en el que se encontrarán almacenados.

ALBA plantea la publicación de datos en abierto como un proyecto transversal donde, a parte de los investigadores de la división de *Experiments*, será necesaria la participación de los grupos de *Systems*, *Controls* y *Management Information Systems*, con la finalidad de una integración en repositorios de datos y metadatos federados con otros sincrotrones como ICAT[52] e ISPyB[23].

[48] HDF5 https://support.hdfgroup.org/HDF5/

[49] EDF https://datatypes.net/open-edf-files

[50] MATLAB https://es.mathworks.com/products/matlab.html

[51] Python https://www.python.org/

[52] ICAT Project https://icatproject.org/

Longevidad

La intención es preservar los datos durante 5 años después del período de embargo, pero dependiendo de la disponibilidad de recursos financieros se podría alargar este periodo, ya que los datos que justifican una publicación son siempre útiles, al menos hasta que esta quede obsoleta y deje de citarse. Puede que algunos datos pierdan utilidad porque se haya mejorado la disciplina técnica con la que se obtuvieron, aunque es algo complejo de predecir.

Estadísticas de uso

Se considera prioritario la medición del impacto, y la capacidad de examinar las estadísticas de uso del portal de datos, para justificar la inversión, junto con la elaboración de encuestas de satisfacción y patrones demográficos de los usuarios.

Los usuarios objetivos de los datos obtenidos en un sincrotrón son limitados, es necesario un conocimiento profundo de la disciplina y del experimento, siendo imprescindible guardar unos metadatos muy completos y consistentes sobre el entorno y como se produjeron los datos, además de como se fabricaron las muestras. Esto facilitará que cualquier investigador del campo que haya realizado experimentos de sincrotrón pueda estar interesado en procesar los datos.

5. Recomendaciones

Los investigadores disponen de herramientas para la creación de un plan de gestión de datos[53]y guías para la publicación de datos en abierto[11], de forma individual pueden utilizar los portales públicos para compartir los datos producidos en sus investigaciones. Si bien esta es una manera de cumplir con los requerimientos impuestos por las fuentes de financiación, o por una convicción de la utilidad de la publicación de datos en abierto, la apuesta por esta filosofía debe venir de los centros de investigación.

Son estos los que pueden aprovechar todo el potencial de los metadatos que acompañan a los datos. Dejando que cada investigador publique de una forma individualizada los datos, se pierde la oportunidad de definir una ontología común para los datos producidos en los centros de investigación que permita realizar análisis, correlaciones y aplicar técnicas de inteligencia de negocio sobre las investigaciones que se llevan a cabo en los centros. A partir de diferentes aspectos y variables contenidos en los metadatos es posible obtener información de apoyo para la toma de decisiones y la mejora de la competitividad.

Los centros de investigación que deseen implantar estos procesos de publicación de datos también disponen de guías[16] y herramientas para la creación de portales web propios, como las descritas en la sección 2.6 de este artículo.

En cuanto a las barreras a las que se enfrentan los centros de investigación para publicar los datos en abierto, en el estudio de los casos del CERN y ALBA se observan condiciones desiguales. El portal CERN Open Data es una solución madura que ha superado todas las barreras descritas, tanto políticas, legales, técnicas, financieras y comunicativas, debido principalmente al apoyo de la UE, a los recursos financieros disponibles, y al compromiso de este centro con la vanguardia en la investigación.

El caso de ALBA es distinto, en su estado preliminar ya ha tenido que superar barreras políticas, como la reticencia de investigadores a la publicación de datos en abierto, por miedo a un detrimento competitivo si la publicación de datos no se implanta en todos los laboratorios de luz de sincrotrón, y barreras legales, superadas por el amparo de las normativas europeas.

Pero ALBA todavía debe enfrentarse a otras barreras, la principal, y de la que dependen el resto, es la financiera, ya que sin una asignación de recursos adecuada la parte técnica del proyecto se ve comprometida. El soporte elegido para los datos abiertos, la cinta magnética, no es el más adecuado para un acceso y recuperación rápidos de los datos, llegando incluso a requerir la intervención humana en el proceso. En este sentido, sería conveniente la utilización de un almacenamiento basado en disco. No es necesario que este almacenamiento disponga de las prestaciones de rendimiento requeridas para la adquisición de datos en las líneas experimentales, en su lugar puede utilizarse un almacenamiento *Software Defined Storage* como son las tecnologías

[53]DMP Plan https://www.openaire.eu/opendatapilot-dmp

basadas en software libre BeeGFS[54], CEPH[55], Gluster[56] o Lustre[57], que permiten abstraerse del hardware reduciendo los costes, a la vez que ofrecen la posibilidad de un crecimiento escalado.

Para reducir el coste inicial del proyecto también es recomendable que la apertura de datos se lleve a cabo de una forma progresiva, empezando por un conjunto de datos reducido que permita la verificación y corrección de los procesos definidos, así como una inversión en infraestructura menor, y que vaya creciendo a medida que se incorporan otros conjuntos de datos en sucesivas fases del proyecto.

En cuanto a las barreras comunicativas, ALBA deberá llevar a cabo iniciativas que promocionen los datos abiertos entre los potenciales usuarios, así como la implantación de herramientas que permitan analizar el impacto de la iniciativa y aplicar medidas correctivas, en caso de que sean necesarias, para asegurar el retorno de la inversión.

6. Conclusiones

El análisis del estado actual de la publicación de datos en abierto muestra que esta es una realidad en las administraciones gubernamentales, en cambio en el ámbito científico no es tan evidente, aunque el apoyo de las autoridades europeas, en forma de subvenciones, incentivos y ofreciendo herramientas, sin duda impulsará en los próximos años la publicación en abierto de los datos provenientes de la investigación.

Una muestra de esta situación son los casos estudiados del CERN y el ALBA. El primero ha logrado implantar soluciones maduras como CERN Open Data y Zenodo, este último gracias a iniciativas europeas como OpenAire y H2020. En cambio ALBA está iniciando este proyecto y debe enfrentarse a las barreras habituales que impiden una adopción más amplia de la publicación de datos en abierto.

La tabla 5 muestra un resumen comparativo de los aspectos analizados en los casos estudiados en este artículo.

[54] BeeGFS https://www.beegfs.io
[55] CEPH http://ceph.com/
[56] Gluster https://www.gluster.org/
[57] Lustre http://lustre.org/

Característica	Zenodo	CERN Open Data	ALBA
Tipo	multi-disciplinar	específico LHC	específico Sincrotrón
Contenido	cualquiera <50GB	niveles 2 y 3 DPHEP	todos los niveles DPHEP
Acceso	abierto, embargado y cerrado	abierto y embargado	abierto y embargado
Reutilización	según licencia	dominio público	autoría
Longevidad	20 años	No definida	Mín. 5 años
Uso (05/2017)	197.636 registros	27TB	Por determinar

Tabla 5: Resumen de los casos estudiados.

Referencias

[1] Organisation for Economic Co-Operation and Development. Making Open Science a Reality, 2015. http://dx.doi.org/10.1787/5jrs2f963zs1-en.

[2] Directorate-General for Research & Innovation EUROPEAN COMMISSION. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. https://erc.europa.eu/sites/default/files/document/file/ERC%20Open%20Access%20guidelines-Version%201.1._10.04.2017.pdf.

[3] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4):258 – 268, 2012.

[4] Final report of the High Level Expert Group on Scientific Data. Riding the wave - How Europe can gain from the rising tide of scientific data. 2010. http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204.

[5] Birgit Schmidt, Birgit Gemeinholzer, and Andrew Treloar. Open data in global environmental research: The belmont forum's open data survey. *PLOS ONE*, 11(1):1–29, 01 2016.

[6] Wendy Carrara, Margriet Nieuwenhuis, and Heleen Vollers. Open Data Maturity in Europe 2016. European Commission, Directorate-General of Communications Networks,

Content & Technology., 2016. https://www.europeandataportal.eu/en/highlights/open-data-maturity-europe.

[7] World Wide Web Foundation. Open Data Barometer Global Report 3rd Edition, 2015. http://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-GlobalReport.pdf.

[8] Guidelines on the Implementation of Open Access to Scientific Publications and Research Data in Projects supported by the European Research Council under Horizon 2020, Noviembre 2016. https://erc.europa.eu/sites/default/files/document/file/ERC_Guidelines_Implementation_Open_Access.pdf.

[9] Anneke Zuiderwijk, Marijn Janssen, and Chris Davis. Innovation with open data: Essential elements of open data ecosystems. *Information Polity: The International Journal of Government & Democracy in the Information Age*, 19(1/2):17 – 33, 2014.

[10] Noël Van Herreweghe. Open Data Manual: Practice-oriented manual for the publication and management of Open Data using the Flemish Open Data Platform. 2015. Government of Flanders in Belgium https://www.w3.org/2013/share-psi/wiki/images/b/bb/Open_Data_Handbook_12022015_EN.pdf.

[11] Wendy Carrara, Fréderique Oudkerk, Eva van Steenbergen, and Dinand Tinholt. Open Data Goldbook for Data Manager and Data Holders. 2016. European Data Portal. European Commission Directorate General for Communications Networks, Content and Technology, https://www.europeandataportal.eu/sites/default/files/goldbook.pdf.

[12] Christian Bizer, Tom Health, and Tim Berners-Lee. Chapter 8, Linked Data: The Story So Far from the book Semantic Services, Interoperability and Web Applications: Emerging Concepts. 2011. IGI Global.

[13] Bernadette Farias Lóscio, Caroline Burle, Newton Calegari, Annette Greiner, Antoine Isaac, Carlos Iglesias, Carlos Laufer, Christophe Guéret, Deirdre Lee, Doug Schepers, Eric G. Stephan, Eric Kauz, Ghislain A. Atemezing, Hadley Beeman, Ig Ibert Bittencourt, João Paulo Almeida, Makx Dekkers, Peter Winstanley, Phil Archer, Riccardo Albertoni, Sumit Purohit, and Yasodara Córdova. Data on the Web Best Practices, January 2017. W3C Recommendation, https://www.w3.org/TR/dwbp/.

[14] Claire C. Austin, Theodora Bloom, Sünje Dallmeier-Tiessen, Varsha K. Khodiyar, Fiona Murphy, Amy Nurnberger, Lisa Raymond, Martina Stockhause, Jonathan Tedds, Mary Vardigan, and Angus Whyte. Key components of data publishing: using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries*, pages 1–16, 2016. http://dx.doi.org/10.1007/s00799-016-0178-2.

[15] Alex Ball. How to License Research Data. *DCC How-to Guides*, 2014. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides/license-research-data.

[16] Tom Sasse, Amanda Smith, Ellen Broad, Jeni Tennison, Peter Wells, and Ulrich Atz. Recommendations for Open Data Portals: from setup to sustainability, 2017. European Commission, https://www.europeandataportal.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf.

[17] Maxi Kindling, Heinz Pampel, Stephanie van de Sandt, Jessika Rücknagel, Paul Vierkant, Gabriele Kloska, Michael Witt, Peter Schirmbacher, Roland Bertelmann, and Frank Scholze. The Landscape of Research Data Repositories in 2015: A re3data Analysis. *D-Lib Magazine*, 23(3/4), March/April 2017. https://doi.org/10.1045/march2017-kindling.

[18] Creso Sá and Julieta Grieco. Open Data for Science, Policy, and the Public Good. *Review of Policy Research*, 33(5):526–543, 9 2016.

[19] EUROPEAN COMMISSION. European Cloud Initiative - Building a competitive data and knowledge economy in Europe. 2016. http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15266.

[20] Study Group for Data Preservation and Long Time Analysis in High-Energy Physics. Data Preservation in High-Energy Physics. *DPHEP-2009-001*, 2009.

[21] A. Calderon, D. Colling, A. Huffman, K. Lassila-Perin, T. McCauley, A. Rao, A. Rodriguez-Marrero, and E Sexton-Kennedy. Open access to high-level data and analysis tools in the CMS experiment at the LHC. volume 664 of *Conference Series*. Journal of Physics, 2015.

[22] Jake Carlson. The data curation profiles toolkit: User guide. *Purdue University Libraries*, 2010.

[23] Solange Delagenière, Patrice Brenchereau, Ludovic Launer, Alun W. Ashton, Ricardo Leal, Stéphanie Veyrier, José Gabadinho, Elspeth J. Gordon, Samuel D. Jones, Karl Erik Levik, Seán M. McSweeney, Stéphanie Monaco, Max Nanao, Darren Spruce, Olof Svensson, Martin A. Walsh, and Gordon A. Leonard. ISPyB: an information management system for synchrotron macromolecular crystallography. *Bioinformatics*, 27(22):3186, 2011.