DATOS ABIERTOS

1. Proyecto piloto de datos abiertos en H2020

La Comisión Europea ha puesto en marcha un Proyecto Piloto para el acceso y la reutilización de los datos de investigación generados por los proyectos financiados por H2020 (ORD pilot). En 2017 este piloto abarcará a todos los proyectos de H2020.

2. Qué se entiende por datos abiertos

El término Open Data hace referencia a dos tipos de apertura:

- los datos están legalmente abiertos, de forma que una licencia de uso permite su reutilización siempre que se cite la autoría de los datos
- los datos están técnicamente abiertos, registrados en formatos no propietarios que pueden ser leídos por las máquinas.

3. A qué tipo de datos se aplica el ORD pilot

Principalmente a los datos necesarios para validar los resultados que se presentan en las publicaciones científicas, pero también pueden ser puestos en acceso abierto datos puros no asociados a publicaciones científicas. Se consideran datos: estadísticas, resultados de experimentos, medidas, observaciones resultado del trabajo sobre el terreno, resultados, entrevistas e imágenes.

Estos datos deben ponerse en abierto tan pronto como sea posible.

4. Excepciones

H2020 permite no participar en el piloto de datos abiertos (opción opt out) siempre que se justifique esta opción. Están contemplados como posibles motivos los siguientes:

- la participación es incompatible con la protección de datos que pueden ser explotados comercial o industrialmente
- los datos son confidenciales
- compartir los datos vulnera regulaciones de protección como la Ley de Protección de datos de carácter personal
- la participación en el piloto impediría alcanzar el objetivo del proyecto
- el proyecto no genera o recoge ningún tipo de dato
- otros motivos legítimos.

H2020 además excluye del piloto:

- "co-fund" and "prizes" instruments
- "ERC proof of concept" grants
- "ERA-Nets" that do not produce data

- SME instrument, phase 1

5. ¿Hay penalizaciones por no participar en el piloto?

La participación en el piloto ORD no influye en la evaluación de las propuestas, por lo que no se penaliza la opción de permanecer al margen a la hora de evaluar las propuestas de proyectos.

6. Qué debo presentar al solicitar un proyecto H2020

A la hora de solicitar un proyecto hay que presentar un documento que indique:
- qué estándares se van a aplicar para describir y gestionar los datos
- cómo serán explotados los datos o hechos accesibles para verificación y reutilización
- si los datos no pueden ser hechos públicos, por qué
- cómo serán tratados y conservados los datos.

Este documento debería además:
- indicar los acuerdos tomados por el consorcio sobre la gestión de los datos
- ser compatible con explotación y derechos de propiedad intelectual.

Se debería hacer también una planificación presupuestaria de los gastos derivados de la gestión de los datos, ya que son gastos reintegrables.

7. Qué es el plan de gestión de datos

El plan de gestión de datos es un elemento clave para gestionar correctamente los datos ya que describe el proceso de recogida, procesamiento y/o generación de los datos.

El plan de gestión de datos debe incluir información sobre:
- el manejo de los datos de investigación durante el proyecto y una vez finalizado
- qué datos se van a recoger, procesar y/o crear
- qué metodología y qué estándares se van a aplicar
- si los datos serán compartidos en acceso abierto
- de qué manera van a ser preservados incluso una vez acabado el proyecto.

8. Cuándo debo presentar el plan de gestión de datos (DMP)

Una vez aprobada la financiación del proyecto, se debe presentar una primera versión del plan de gestión de datos dentro de los 6 primeros meses de realización del mismo. La Comisión ofrece una plantilla que recomienda utilizar y que se puede encontrar como anexo en este documento:
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

9. Ejemplos de planes de gestión

El Digital Curation Centre ha elaborado una guía para crear un plan de gestión de datos y en su web se pueden encontrar varios planes presentados por diversos proyectos.

Acceso a la guía: http://www.dcc.ac.uk/resources/how-guides/develop-data-plan

Ejemplos de planes de gestión: http://www.dcc.ac.uk/resources/data-management-plans/guidance-examples

10. Herramientas para crear planes de gestión

A la hora de crear un plan de gestión de datos pueden ser útiles dos herramientas:

- Data Management Planing Tool: https://dmptool.org/
- Data Curation Profiles: http://datacurationprofiles.org/

11. Cómo calculo el coste de gestión de los datos

El UK Data Service ha preparado un listado de cuestiones que se tienen que tener en cuenta para calcular el coste de hacer accesibles los datos de investigación. Está orientado a las ciencias sociales, pero puede servir se ayuda: http://www.data-archive.ac.uk/media/247429/costingtool.pdf

12. Quién paga el coste

Los gastos asociados al acceso abierto a los datos de investigación pueden ser solicitados como reintegrables mientras dure el proyecto en las condiciones establecidas en el artículo 6 del H2020 Grant Agreement

Artículo 6: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf#page=36

13. Cómo hago accesibles los datos

Para que los datos sean accesibles se deben cumplir dos requisitos:

- en 1º lugar, hay que depositarlos en un repositorio institucional o temático de los que se incluyen el Registry of Research Data Repositories
- en 2º lugar, en la medida de lo posible se deben adoptar medidas para permitir el acceso a terceros, la minería de datos, la difusión y la reproducción gratuitas por medio de licencias Creative Commons (CC BY o CC BY SA) u OpenData Commons

14. Qué repositorios puedo utilizar para depositar datos

El depósito de los datos debe realizarse en un repositorio de los que aparecen recogidos en el Registry of Research Data Repositories (http://www.re3data.org/). El repositorio institucional del CSIC está en este directorio por lo que puede utilizarse para depositar los datos de investigación.

15. Qué tipo de datos puedo depositar en Digital.CSIC

El repositorio institucional del CSIC acepta el depósito de las siguientes categorías de datos: cuadernos de laboratorio, cuadernos de campo, datos de investigación primaria (incluidos los datos en papel o en soporte informático), cuestionarios, cintas de audio, vídeos, desarrollo de modelos, fotografías, películas, y comprobaciones y respuestas de la prueba.

Digital.CSIC asigna DOIs automáticamente a todos los datasets que se suben en acceso abierto y estos entran a formar parte del portal de datos DataCite.

Hay que tener en cuenta que el repositorio no es una plataforma para albergar big data, sino que se centra en la llamada "long tail of science".

16. Qué información debo aportar cuando deposito los datos

Los datos que se depositan en los repositorios deben ir acompañados de metadatos que incluyan información sobre los programas, algoritmos o protocolos necesarios para validarlos. Además es recomendable aportar una referencia bibliográfica completa e información sobre el contenido del conjunto de datos, el contexto y la fuente, información sobre su metodología, instrumentos y técnicas empleadas en la creación o recolección de datos, así como referencias a publicaciones y/o sitios web relativos.

17. Esquemas de metadatos para datos de investigación

Son muchos los esquemas de metadatos que se pueden utilizar para describir datos de investigación. Os pueden servir de ayuda los siguientes recursos:

- el RDA Metadata Standards Directory Working Group mantiene una página web con información sobre los diferentes estándares que existen para cada disciplina y con herramientas para extraer y convertir metadatos: http://rd-alliance.github.io/metadata-directory/
- el DataCite Metadata Working Group ha publicado un esquema de metadatos para la publicación y cita de datos de investigación: https://schema.labs.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf

18. Plantilla para la descripción de datos de Digital.CSIC

El repositorio institucional del CSIC dispone de una plantilla de descripción de datasets (http://digital.csic.es/bitstream/10261/81323/4/Datasets_DC_plantilla.pdf) que permite crear una descripción detallada de la información contenida en los mismos para facilitar su reutilización.

19. Licencias para datasets

A la hora de sujetar un conjunto de datos a una licencia de uso los autores de los datos deben considerar:

• la identificación del material que debe cubrir la licencia,

• la identificación de material que haya sido usado como fuente en la elaboración de los datos,

• la identificación de cualquier tipo de restricción de uso que pudiera existir en el material original a partir del cual se han originado los datos.

Los tipos de licencias más frecuentes para datos de investigación son las Creative Commons y las OpenData Commons. Para datasets las licencias más utilizadas son:

- CC-BY: esta licencia permite la reutilización del conjunto de datos, sin necesidad de pedir permiso expreso a los autores, para estos usos: reproducción, distribución, difusión, y transformación (creación de obras derivadas) siempre y cuando se reconozca la autoría y se cite el conjunto de datos tal y como se indica en la licencia

- CC BY SA: permite una amplia reutilización pero obliga a licenciar bajo las mismas condiciones las obras derivadas

- Open Data Commons Open Database License (ODbL): permite a cualquier usuario de Internet reproducir, distribuir y usar el conjunto de datos, y adaptarlo y transformarlo siempre y cuando: se haga un reconocimiento explícito a la autoría del conjunto de datos originales y a los términos de uso expresados en la licencia; si se realizan obras derivadas, se deben ofrecer bajo la misma licencia de uso (ODbL); si se realizan versiones o adaptaciones con restricciones de acceso, se debe seguir garantizando la disponibilidad de una copia en acceso abierto. Esta licencia cubre el dataset en su conjunto y los contenidos individuales (bajo una licencia Database Content License, que puede ser sustituida por otra preferida por el autor).

- Open Data Commons Attribution License (ODC-BY): permite a cualquier usuario de Internet reproducir, distribuir y usar el conjunto de datos, adaptarlo y transformarlo siempre y cuando: se haga un reconocimiento explícito a la autoría del conjunto de datos originales y a sus términos de uso. Esta licencia gobierna los usos del dataset en su conjunto, no de sus contenidos individuales

En caso de optar por una licencia Creative Commons, es conveniente utilizar la versión 4.0 de carácter internacional.

20. Otras cuestiones que hay que tener en cuenta para depositar los datos: formatos recomendados, nombre de los ficheros, autoría múltiple, fichero Readme

- Formatos recomendados para los datos

Para garantizar el acceso a largo plazo de los datos de investigación es necesario que estén codificados en formatos estándar que puedan ser interpretados por la mayoría de los software y que permitan el intercambio y la transformación de datos. Se recomienda, por ejemplo:

- PDF/A mejor que Word o PowerPoint
- ASCII o CSV mejor que Excel
- MPEG-4 mejor que QuickTime
- TIFF o JPEG2000, mejor que GIF o JPG
- XML o RDF, mejor que RDBMS

- Nombre de los ficheros

A la hora de nombrar los ficheros de datasets para depositarlos en un repositorio, es conveniente buscar un nombre consistente que refleje el contenido del fichero, pero se debe evitar usar el nombre que se le haya dado a un artículo científico basado en ese dataset.

- Autoría múltiple

Si los datos contenidos en el dataset son responsabilidad de un gran número de autores es recomendable poner el nombre del grupo como responsable del mismo para facilitar la citación del dataset y especificar el nombre de los componentes del grupo en un fichero Readme que acompañe al dataset.

- Fichero Readme

El fichero Readme proporciona información sobre el dataset para permitir su correcta interpretación por personas y máquinas. Este fichero debe ir asociado al dataset correspondiente, codificado en texto plano como txt y contener la siguiente información: breve descripción del dataset, contacto del investigador principal, fecha de recogida de datos y de creación del dataset, información geográfica de los datos, metodología empleada, enlace a publicaciones y otra documentación; unidades de medida, protocolos, abreviaciones, códigos y símbolos asociados a los datos; licencia de uso y forma de cita recomendada.

Podéis ver aquí un esquema desarrollado:

https://drive.google.com/file/d/0B5Dm3XFQloc4RDY4VEM4OFJobUk/view?pref=2&pli=1

21. Qué pasa con la información sensible

A la hora de poner en acceso abierto un dataset se debe tener la precaución de eliminar información sensible como datos de carácter personal o datos de localización de especies en peligro. En estos casos se pueden llevar a cabo diferentes acciones:

- eliminación de identificadores directos como nombre, dirección, fechas personales
- eliminación de datos indirectos como sexo, edad, nivel de estudios, etc
- modificación de datos para limitar la identificación, eliminando líneas de información o redondeando resultados

El UK Data Service ha creado una herramienta para anonimizar datos:

- https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/identifiers

22. Herramientas para gestionar datasets

DataONE dispone en su página web de un catálogo de herramientas de software para trabajar con datasets: https://www.dataone.org/software_tools_catalog

Tres herramientas útiles:

- para gestionar datos: OpenRefine
- para visualizar datos: Infogr.am, Carto

Aquí podéis encontrar una explicación de su funcionamiento:

https://digital.csic.es/handle/10261/138091

23. Buenas prácticas para la gestión de datasets

DataONE ha elaborado una guía con recomendaciones para trabajar eficazmente con los datos durante todo su ciclo de vida: https://www.dataone.org/best-practices

24. Cómo puedo encontrar datos generados por otros proyectos

Para encontrar datos generados por otros grupos de investigación se pueden utilizar buscadores específicos de datos o directorios de repositorios de datos.

Algunos buscadores de datos que se pueden usar son:

- B2Find de EUDAT http://b2find.eudat.eu/
- Datacite Metadata Search busca datasets de las instituciones miembro de DataCite http://search.datacite.org/ui
- OpenAire indexa datasets de repositorios y algunas revistas de acceso abierto https://www.openaire.eu/search/browse/publications?type=0021
- ONEMercury es un buscador de datasets generados por instituciones miembros de DataONE https://cn.dataone.org/onemercury/
- DataONE Search https://search.dataone.org/#data/page/0

Repositorios de datos:

- Re3data http://www.re3data.org/

- Open Access Directory http://oad.simmons.edu/oadwiki/Data_repositories

- Biosharing https://biosharing.org/databases/

- Research Discovery Service http://ckan.data.alpha.jisc.ac.uk/dataset

25. Cómo citar datasets

Para citar datasets se recomienda usar el siguiente modelo:

Autores, fecha de publicación "Título del dataset [Dataset]", Repositorio, enlace permanente (DOI o Handle), versión (si procede)

Ejemplo:
Olabarria, Celia; Gestoso, Ignacio; Lima, Fernando P.; Vázquez, Elsa; Comeau, Luc A.; Gomes, Filipa; Seabra, Rui; Babarro, José M. F. "Response of Two Mytilids to a Heatwave: the Complex Interplay of Physiology, Behaviour and Ecological Interaction [Dataset]", DIGITAL.CSIC, http://hdl.handle.net/10261/137446