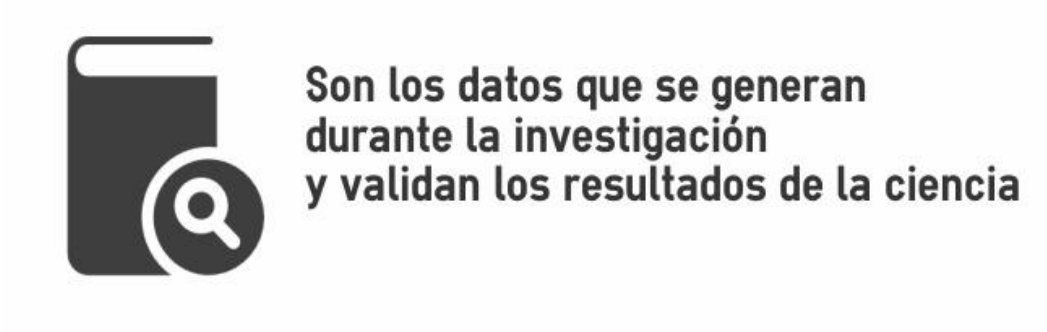
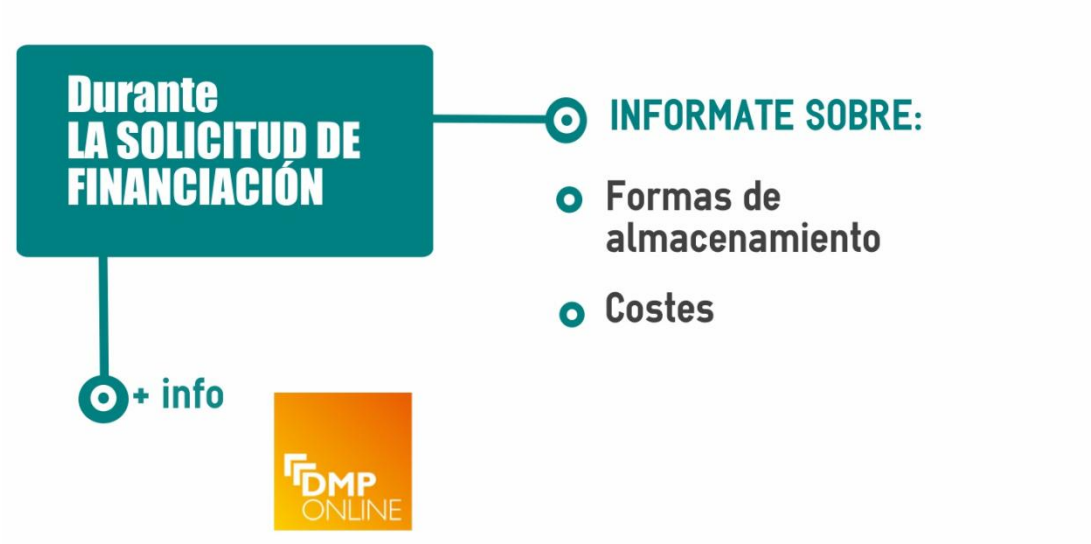


Criterios que deben cumplir los datos para ser abiertos (Royal Society, 2012):

- Accesibles (*accessible*): fáciles de encontrar y en una forma en que puedan ser usados [y preservados].
- Evaluados/certificados (*asessable*): credibilidad para diferentes grupos de interés.
- Inteligibles (*intelligible*): ser entendidos.
- Reutilizable (*useable*): en un formato para usar y con licencias adecuadas.

Gestiona y Comparte
Datos de Investigación



Un Data Management Plan podría aportar información sobre:

- La definición de los datos
  - ¿Cómo los obtienes y con qué instrumentos?
  - ¿Cuándo se actualizan?
  - ¿Cuántos generas y en qué formatos?
  - ¿Cuántas versiones almacenas?
- Establecimiento de mecanismos de control
  - ¿Cuánta información adicional es necesaria para entenderlos?
  - ¿Dónde los almacenas?
  - Directorios y nombres de archivos
  - Copias de seguridad: ¿cómo y cuándo? ¿testeas?
- Hay que definir una serie de detalles para posteriormente poder compartir los datos:
  - ¿De quién es la propiedad?
  - ¿Quién puede usarlos y quién podría?
  - ¿Qué compartes y qué no? ¿por qué?
- Para el archivo de datos definitivo se delimitará:
  - ¿Qué debe ser archivado?
  - ¿Por cuánto tiempo y dónde?
  - ¿Cuándo pasan al estado "archivo"?
  - ¿Quién es el responsable de moverlos?
  - ¿Quién tendrá acceso? ¿En qué condiciones?
- Y para finalizar, hay que establecer unos mecanismos de supervisión del plan
  - ¿Quién es responsable?
  - ¿Con qué frecuencia se actualizará?

Información más detallada sobre lo que debe contener un **Plan de Gestión de Datos y ejemplos** de planes se puede encontrar en: http://www.bath.ac.uk/research/data/planning/checklist.html en ella se pueden localizar también ejemplos de planes de gestión de datos.

En el Plan de Gestión de Datos hay que tener en cuenta los costes que acarrea su gestión. En estos costes determinaremos:

- Costes del almacenamiento de datos: si es en institución propia, en servicios externos contratados, el tiempo y el espacio requerido.
- Costes de transferencia de datos y acceso: si se necesitan características especiales para la transferencia de datos.
- Costes de copias de seguridad.
- Costes de seguridad para el acceso.

Para más información se pude consultar: http://www.data-archive.ac.uk/media/247429/costingtool.pdf

En relación con la organización hay que normalizar (Stuart McDonald[1]):

- La estructura de los directorios.
- El nombre de los ficheros.
- Cómo documentar los datos, anotando la procedencia, la frecuencia de actualización, cuántos datos se generan y los formatos de datos que se van actualizando. Hay también que documentar los datos de manera que sirvan para recordar, ayudar a otros, verificar, replicar, archivar, reclamar autorías…

  Ejemplos: cuadernos de laboratorio, notas de campo, metodologías cualitativas
  - Nivel de Proyecto: documentar la base del estudio, métodos, instrumentos, hipótesis de trabajo
  - Nivel de archivo o dataset: formatos, relaciones entre archivos
  - Nivel de variable o item: como fue generada la variable y descripción de los campos

- Controlar las versiones (redactar manuales). Es imprescindible identificar las diferentes versiones, especialmente si el dataset es actualizado por múltiples usuarios. Se recomienda:
  - Usar un sistema de número secuencial: v1, v2, v3, etc.
  - No utilizar nombres confusos como revisión, final, final2, etc.
  - Registrar todos los cambios, incluso los pequeños.
  - Descartar versiones obsoletas (nunca eliminar los datos brutos).
- Los backups automáticos.
- La información adicional para entender cada uno de los archivos.
- Las copias de seguridad que se establezcan.

En relación a la propiedad de los datos, el conocimiento legal necesario para tomar una decisión acertada en cuanto al almacenamiento en la nube no es tarea habitual en los equipos de investigación (Jahnke y Asher, 2013[2]). Los equipos han de lidiar situaciones altamente complejas y multidimensionales. La firma de acuerdos precisos sobre la titularidad y derecho de uso de los datos producidos no es trivial, especialmente si los equipos están formados por consorcios internacionales.

- En relación con el archivo hay que diferenciar:
  - Lo que se archivará al finalizar el proyecto
  - Durante cuánto tiempo
  - Dónde se almacenará
  - Quién es responsable de enviar los datos al archivo y mantenerlo
  - Quién puede tener acceso

Debe existir una persona responsable de que se cumpla el plan de gestión de datos. Se ocupará de todo el proceso y fundamentalmente de la integridad y seguridad de los datos.
En esta fase del proyecto donde aún se está trabajando con los datos, se pueden compartir a través de plataformas como Dropbox, Drive o Figshare.

[1] http://datalib.edina.ac.uk/mantra/libtraining.html

[2] Jahnke, Lori M.; Asher, Andrew. (2013) Dilemmas of Digital Stewardship: Research Ethics and the Problems ofData Sharing. En: Research Data Management: Principles, Practices, and Prospects. Washington, D.C.: CLIR pp. 80 – 99 http://www.clir.org/pubs/reports/pub160/pub160.pdf

Una vez finalizado el proyecto, llega el momento de pasar a la publicación en abierto de los datos.

Antes de conservarlos en abierto hay que seleccionar los ficheros a compartir, ya que no puede preservarse todo. Para ello, se pueden utilizar varios criterios, entre ellos, los propuestos por el DCC[3]. Además, hay que preparar los ficheros de manera adecuada, con el fin de proporcionarlos en formatos accesibles, lo que se denomina "abiertos".

No se debe olvidar que el hardware y software utilizados en el tratamiento de los datos puede quedar obsoleto. Para que los datos perduren a lo largo del tiempo, lo mejor es convertirlos a un formato estándar no propietario.

[3] Criterios DCC. *How to Select and Appraise Research Data for Curation*. http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services#Selection-criteria

Los formatos recomendados para la preservación de los archivos son:

Tipo de datos	Formato de archivo para compartir, reusar y preservar
Quantitative tabular data with extensive metadata **Datos tabulados con metadatos extensivos: un conjunto de datos con etiquetas de las variables, etiquetas de códigos, y los valores que faltan definir, y la matriz de datos**	SPSS en formato compatible (.por) Texto delimitado y comando ('setup') con archivo (SPSS, Stata, SAS, etc), que contenga información de los metadatos Texto estructurado o un archivo de marcado que contenga información de metadatos, por ejemplo, Archivo XML DDI
Datos tabulados cuantitativos con metadatos mínimos: **una conjunto de datos con o sin encabezados de las columnas o nombres de variables, pero que no contiene ningún otro metadato o etiqueta**	Valores delimitados por comas (CSV) (. csv) Valores delimitados por tabulaciones (tab). Texto delimitado por determinado conjunto de caracteres con instrucciones SQL donde sea apropiado
Datos geoespaciales **Datos vectoriales y raster**	ESRI Shapefile (essential: .shp, .shx, .dbf ; optional: .prj, .sbx, .sbn) Datos georreferenciados TIFF (.tif .tfw) datos CAD (.dwg) Datos de atributos de tabla GISS
Datos cualitativos **Textuales**	eXtensible Mark-up Language (XML) texto de acuerdo con Document Type Definition (DTD) o esquema (.xml) Rich Text Format (.rtf) texto plano, ASCII (.txt)
Datos de imágenes digitales	TIFF version 6 no comprimidos (.tif)
Datos de audio digitales	Free Lossless Audio Codec (FLAC) (.flac)
Datos de video digitales	MPEG-4 (.mp4) motion JPEG 2000 (.jp2)
Documentación	Rich Text Format (.rtf) PDF/A or PDF (.pdf) OpenDocument Text (.odt)

Conversión de los datos

Cuando se vayan a depositar los datos para su conservación, deben ser los mismos investigadores quienes los conviertan y revisen, pues ellos son quienes mejor conocen su integridad y pueden detectar los cambios y errores producidos por la conversión. Los datos de paquetes estadísticos, hojas de cálculo o bases de datos suelen perder metadatos internos como definiciones, decimales, fórmulas o etiquetas, incluso en algunos casos se pueden truncar los datos.

Deben contar con los metadatos suficientes que describan el origen de los datos, su propósito, cuándo se crearon, su ubicación geográfica, creador, condiciones de acceso y condiciones de uso.

No hay que olvidar que los metadatos son procesados por máquinas y se utilizan para el intercambio, exposición y consulta. Cuantos más metadatos se incluyan, mejor será la recuperación de la información. Pueden ser de distintos tipos:

- Descriptivos: DC, title, author, abstract
- Administrativos: de preservación, derechos, formatos
- Estructurales – Describe las relaciones entre ellos o entre tablas en las bases de datos

La documentación que acompaña al proyecto debe contener los siguientes apartados:

- El contexto en el que se ha realizado su recopilación: la historia del proyecto, objetivos e hipótesis.
- Los métodos de recolección: el muestreo, el proceso de recolección de datos, los instrumentos utilizados, el hardware y software utilizado, la escala y resolución, cobertura temporal y geográfica y las fuentes de datos secundarios utilizados.
- Estructura, casos de estudio, y las relaciones entre los archivos.
- La validación de datos, verificación, pruebas, limpieza y procedimientos de control de calidad llevado a cabo.
- Los cambios realizados desde su creación original y la identificación de diferentes versiones.
- Información sobre las condiciones de acceso, uso y confidencialidad.

Las licencias de uso de los datos que aparezcan deben ser lo menos restrictivas posibles, como la Creative Commons Zero o la Licencia al Dominio Público (Public Domain Dedication & Licence PDDL-OKF)

Depósito de los datos

Una vez ya preparados los datos, hay distintas opciones a la hora de depositarlos. En España todavía no existen políticas y normativas para la gestión y reutilización de los datos. Tampoco se han desarrollado infraestructuras técnicas preparadas para el almacenamiento de datos brutos de investigación, a excepción de Digital.CSIC, del Consejo Superior de Investigaciones Científicas, la Universidad Pompeu Fabra y la Fundación Juan March (Nina et al, 2013).

Las posibilidades son:

- Almacenamiento en las plataformas de las ediciones electrónicas de las revistas, habitualmente bajo la categoría de “material suplementario”.
- Almacenamiento en repositorios institucionales específicos para datos, especializados por disciplinas o de carácter multidisciplinar.
- Depositándolo en un banco de datos especializado.

El depósito en la edición electrónica de las revistas:

- Ofrecen una mayor facilidad para localizarlos y entenderlos pero no ofrecen suficientes garantías porque:
- la preservación a largo plazo no suele tener suficiente interés, para las entidades editoras, que dudan de su capacidad para generar beneficios económicos.
- la información se dispersa en múltiples fuentes no interoperables técnicamente.

El depósito en repositorios institucionales aporta dos ventajas:

- Su interoperabilidad.
- Existencia de una institución responsable que garantiza su calidad y su continuidad.

El depósito en bancos de datos especializados aporta:

- La preservación a largo plazo en formatos accesibles y estandarizados y la conversión de formatos cuando sea necesario debido a actualizaciones de software.
- Custodia en un entorno seguro, con la capacidad de controlar el acceso cuando sea necesario.
- Back-ups regulares.
- Búsqueda a través de los catálogos de datos.
- Acceso a los datos en formatos populares.
- Acuerdos de concesión de licencias para reconocer los derechos.
- Mecanismo de citación estandarizado.
- Promoción.
- Monitoreo del uso secundario.
- Gestión del acceso y consultas de los usuarios.

En qué repositorios puedo depositar mis datos

- Digital.CSIC (https://digital.csic.es/). Es un repositorio de la actividad investigadora del Consejo Superior
- Investigaciones Científicas. Su objetivo es organizar, archivar, preservar y difundir en modo de acceso abierto la producción intelectual resultante de la institución.
- Existen una serie de repositorios por disciplinas:
  - DRYAD. Es un repositorio abierto de artículos de revistas y de datos de la Universidad de Carolina del Norte.
  - Figshare. Es un repositorio con sede en Londres destinado a las ciencias biológicas, pero abierto a todas las disciplinas.
  - Zenodo . Es el repositorio europeo creado por el proyecto OpenAire que alberga publicaciones, datos, presentaciones, etc. Se encuentra alojado en el CERN, permite asignar licencias, añade DOIS a los objetos digitales.
- Para averiguar si tu disciplina tiene un repositorio organizado, hay algunos registros de repositorios de datos:
  - Databib. Es una iniciativa estadounidense con sede en Purdue University Libraries mantenida con bibliotecarios voluntarios.
  - Re3data. Es un sitio web alemán que tiene como objetivo realizar un registro completo de los repositorios de datos.