CIENCIA ABIERTA Y REPOSITORIO INSTITUCIONALES

Primer Congreso Internacional de Ciencia Abierta y Repositorios Institucionales

Preservación de datos digitales para investigadores

Andréa Gonçalves

Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (Icict/Fiocruz)

Basado en: Open Access Post-Graduate Teaching Materials for Research Data Management in Archaeology, Cambridge University Library

Agenda

1. **Introducción**

   Definiendo nuestros datos de investigación

2. **La teoría**

   Ciclo de vida de datos y planes de gestión: una visión general

3. **La práctica**

   Cosas prácticas: Estructura de archivos, nombres y formatos, etc.

   Cosas útiles**:** Derechos de propiedad intelectual y datos de investigación

   Cosas realmente útiles**:** E-tesis y datos digitales complementarios

   Cosas esenciales: Archivo de datos digitales

4. **Y por fín...**

   Redacción de planes de gestión de datos

Introducción

Definiendo nuestros datos de investigación

Datos de investigación digital

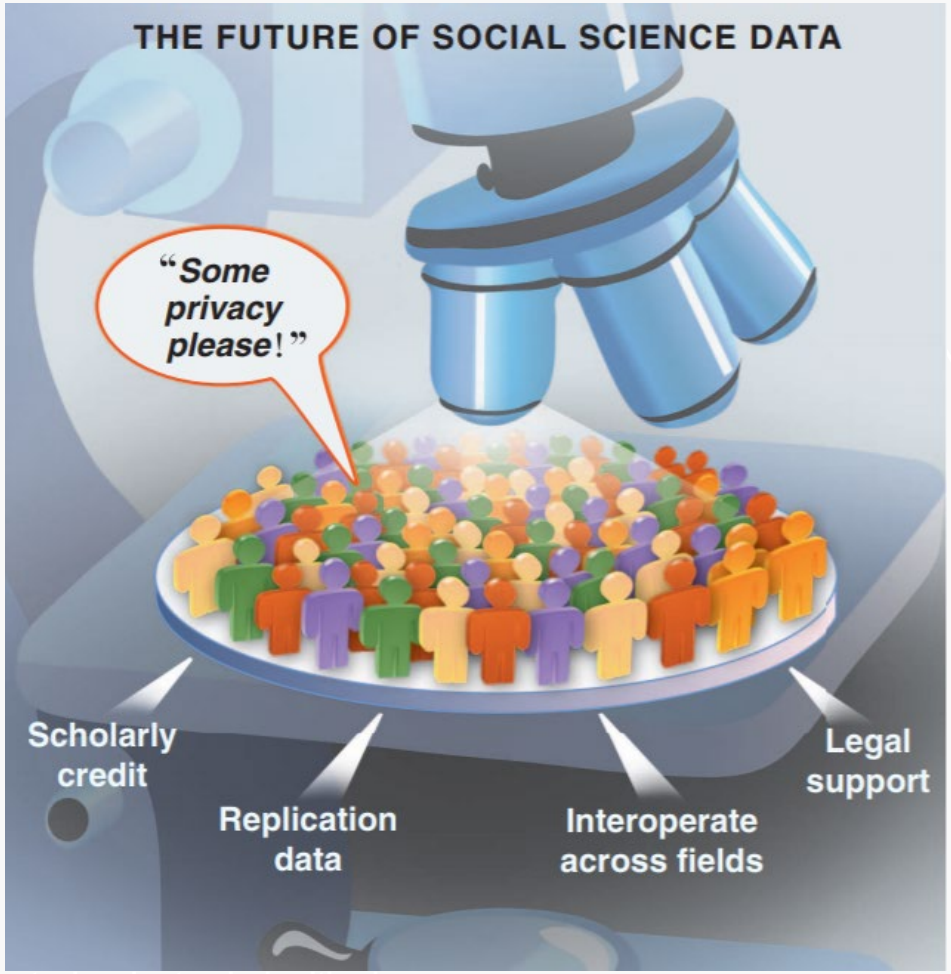
Los datos digitales son todo lo que es creado o manipulado en una computadora:

- Correspondencia por correo electrónico
- Archivos de texto
- Imágenes - desde escaneos digitales de objetos físicos hasta fotos e imágenes complejas en 3D
- Audio
- Video
- Hojas de cálculo y bases de datos - datos numéricos y textuales
- Datos de encuestas, desde encuestas simples hasta escaneos y encuestas geofísicas
- Mensajes de texto
- Sitios web: incluso YouTube puede ser información de investigación.
- Etc…

Algunos problemas con los datos digitales

- Volumen
  - Selección y retención: qué tirar y qué guardar?
- Manejo y preservación
  - Fragilidad de lo datos digitales, soportes, formatos y costos
- Temas legales
  - Autenticidad de datos digitales
  - Copyright (material en línea)
  - Datos confidenciales (datos personales, datos no públicos)
- Reutilización
  - Conocer los detalles técnicos de los datos digitales
  - Comprender el contexto de los datos digitales

Datos digitales en las ciencias sociales ¿Un caso especial?



Fuente: King, Gary. Ensuring the Data-Rich Future of the Social Sciences. *Science*, v. 331, 11 feb 2011.

¿Por qué digitalizar datos físicos?

- Los materiales de publicación son todos digitales
- Copia de seguridad de archivo en papel
- Poner en orden el archivo de papel: notas de campo, ilustraciones, etc.
- Permitir el análisis: buscar en hojas de cálculo/bases de datos, etc.
- Portátil, compartible y reutilizable

Gestión de datos de investigación en el posgrado

¿Por qué estamos hablando sobre gestión de datos?

- Los buenos datos apuntalan la investigación de alta calidad
- Interpretaciones creíbles y verificables
- Preservación a largo plazo
- Reconocimiento y reputación académica y profesional
- Requisitos de organismos de financiación, códigos de conducta legales y éticos

- Para ayudarlo a terminar su tesis a tiempo con el menor estrés

Gestión de datos de investigación: Un contexto más amplio

- Crecimiento exponencial de los datos digitales
- Responsabilidades institucionales y de los investigadores
- Políticas institucionales sobre gestión de datos de investigación
- Repositorios digitales
  - Repositorios digitales institucionales
  - Nacionales y específicos por disciplina

Puntos clave

- Nadie es perfecto!
- Piense en los datos digitales al inicio de la planificación de su proyecto
- La gestión de datos va de la mano con los resultados de la investigación
- Haga que sus datos de investigación sean comprensibles para otros
- Archivo y preservación de los datos para lograr mejor difusión

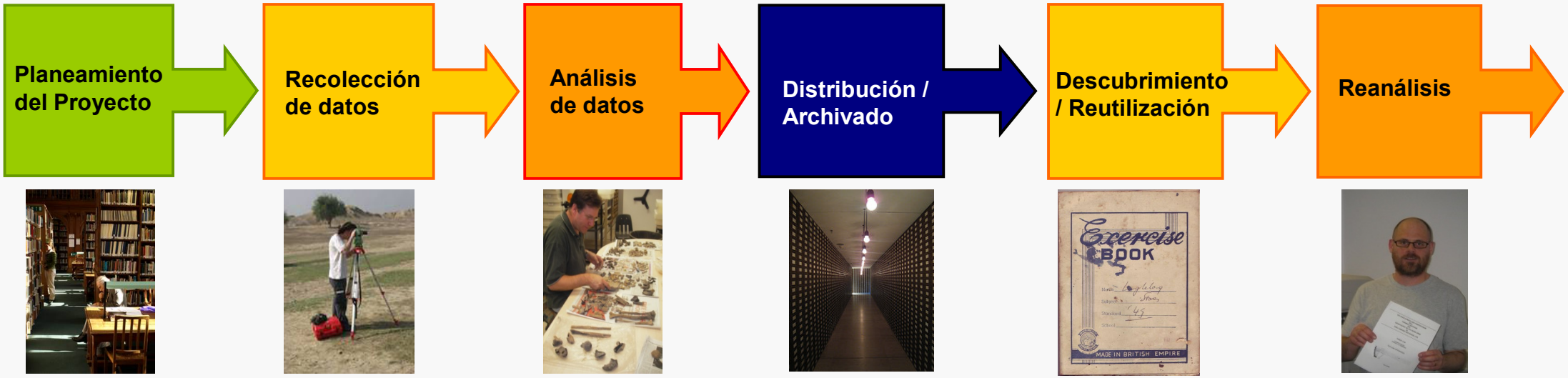
CIENCIA ABIERTA Y REPOSITORIO INSTITUCIONALES

Primer Congreso Internacional de Ciencia Abierta y Repositorios Institucionales

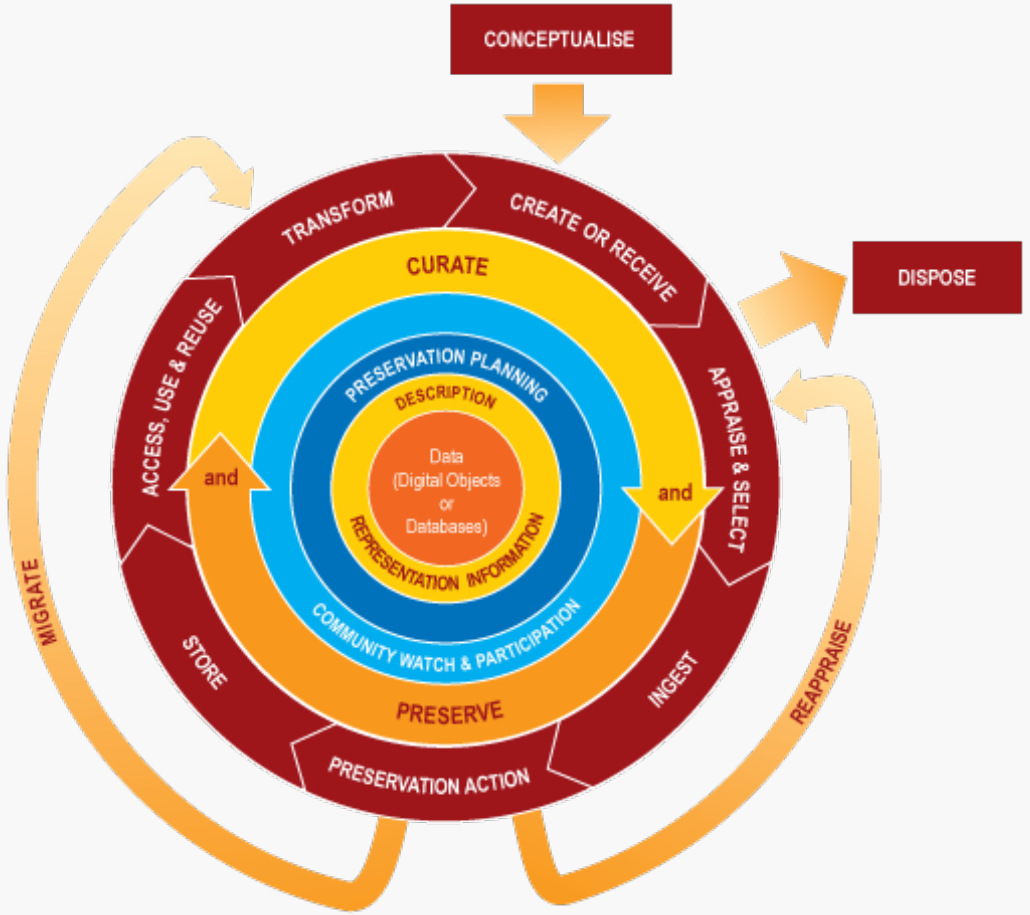
La teoría

Ciclo de vida de los datos y planes de gestión

Ciclo de vida de los datos



Digital Data Curation Centre
Modelo de ciclo de vida



www.dcc.ac.uk *Porque una buena investigación necesita buenos datos*

Algunas jergas y definiciones

Captura de datos digital	Los datos que se derivan de un objeto de datos físicos, por ejemplo, la medición introducida de artefactos, escaneos de dibujos, etc.
Creación de datos digitales	Datos que son natodigitales y que no se derivan de un objeto de datos físicos, por ejemplo, fotografías digitales, datos de encuestas geofísicas, algunas lecturas de análisis, etc.
Documentación	Explica cómo se crearon o digitalizaron los datos, qué significan los datos, cuál es su contenido y estructura, y cualquier manipulación que pueda haber tenido lugar.
Metadatos "Datos sobre datos"	Información estructurada estandarizada que explica el propósito, origen, referencias de tiempo, ubicación geográfica, creador, condiciones de acceso y términos de uso de una recopilación de datos.
Almacenamiento de datos / copia de seguridad	Sistema utilizado para cuidar datos digitales durante la vida de un proyecto. La copia de seguridad NO es preservación.
ingestión	Proceso mediante el cual los datos digitales son archivados por un depositario digital.
Preservación digital	Archivado de datos digitales a largo plazo para que sea accesible en el futuro.
Largo plazo	Período durante el cual las tecnologías cambiantes, los formatos y los medios impactan en el acceso y uso de los recursos digitales.
Migración	Transferencia de recursos digitales de una generación de formato de hardware / medios y software / archivo a la siguiente.

Ciclo de vida de datos y

Planes de gestión de datos

1. ¿Qué datos voy a producir?
2. ¿Cómo voy a organizar los datos?
3. ¿Qué datos voy a guardar? ¿Funciona bien mi gestión de datos?
4. ¿Qué datos se depositarán y dónde?
5. ¿Quién estará interesado en reutilizar los datos?

Ciclo de vida de datos y

Planes de gestión de datos

1. **¿Qué datos voy a producir?**
   - Documentos de texto
   - Análisis de artefactos
   - Análisis de muestra
   - Datos de encuestas
   - Dibujos
   - Fotografías
   - Entrevistas grabadas
   - Etc ..

Ciclo de vida de datos y Planes de gestión de datos

2. **¿Cómo voy a organizar mis datos?**
   - Estructura de archivo
   - Nombre de archivo
   - ¿Qué formatos de archivo usaré?
   - ¿Qué software usaré?
   - Aproximadamente, ¿cuántos archivos?
   - ¿Cómo describiré y documentaré mis datos?

Ciclo de vida de datos y Planes de gestión de datos

3. ¿Funciona bien mi gestión de datos?

- ¿Es la estructura del archivo / nomenclatura comprensible para otros?
- ¿Se requieren más datos?
- ¿Se requieren nuevos tipos de datos?
- ¿Qué datos se guardarán?
- ¿Qué datos se pueden descartar?

Ciclo de vida de datos y Planes de gestión de datos

4. ¿Qué datos se depositarán y dónde?

- Definir el conjunto de datos básicos del proyecto
- ¿Qué datos se incluirán en la tesis?
- ¿Qué datos son suplementarios?
- ¿Produciré una tesis electrónica?
- ¿Dónde depositaré mi tesis electrónica?
- ¿Depositaré datos suplementarios?

Ciclo de vida de datos y Planes de gestión de datos

5. Preservación y reutilización

- ¿Cómo garantizar el acceso a los datos en el largo plazo?
- ¿Quién estará interesado en reutilizar los datos?
- ¿Existe información suficiente para permitir una fácil reutilización de los datos?

La mejor manera de ayudar a preservar los datos es planificar su reutilización

[... en 10, 50, 100 o incluso 500 años ...]

Regresando al futuro ...

¿Quién posee los datos originales?

- ¿Los datos están cubiertos por derechos de propiedad intelectual?
- ¿Hay datos confidenciales en el proyecto?
- ¿Hay datos personales como parte del archivo del proyecto?
- ¿Tendré autoridad para archivar estos datos?
- ¿Cómo obtengo permiso para archivar estos datos?

Primer Congreso Internacional
de Ciencia Abierta y
Repositorios Institucionales

La práctica

Trabajando con los datos digitales

Una historia de horror en la gestión de datos

https://youtu.be/66oNv_DJuPc

Cosas prácticas

- Estructuras de archivos
- Nombramiento de archivos
- Control de versiones
- Formatos de archivo
- Documentación de datos
- Selección

Estructura de los archivos

Dónde poner las cosas para que no las pierdas

- Estructura lógica para uno mismo y fácilmente comprensible para otros
- Facilidad para recuperar, compartir o intercambiar datos en proyectos más grandes o entre miembros del equipo
- Manténgase libre de carpetas y archivos temporales
- Los diseños de investigación cambian y también la estructura de los archivos.
- Evite el uso excesivo de carpetas (aunque es más fácil decirlo que hacerlo).

Nombre de los archivos

Cómo llamar a las cosas para que sepas qué son

- Los nombres nos dicen qué es un archivo – Información contextual
- Nombres ordenan los archivos – Hace las cosas fáciles de encontrar
- Define tu sistema – Y mantente firme!

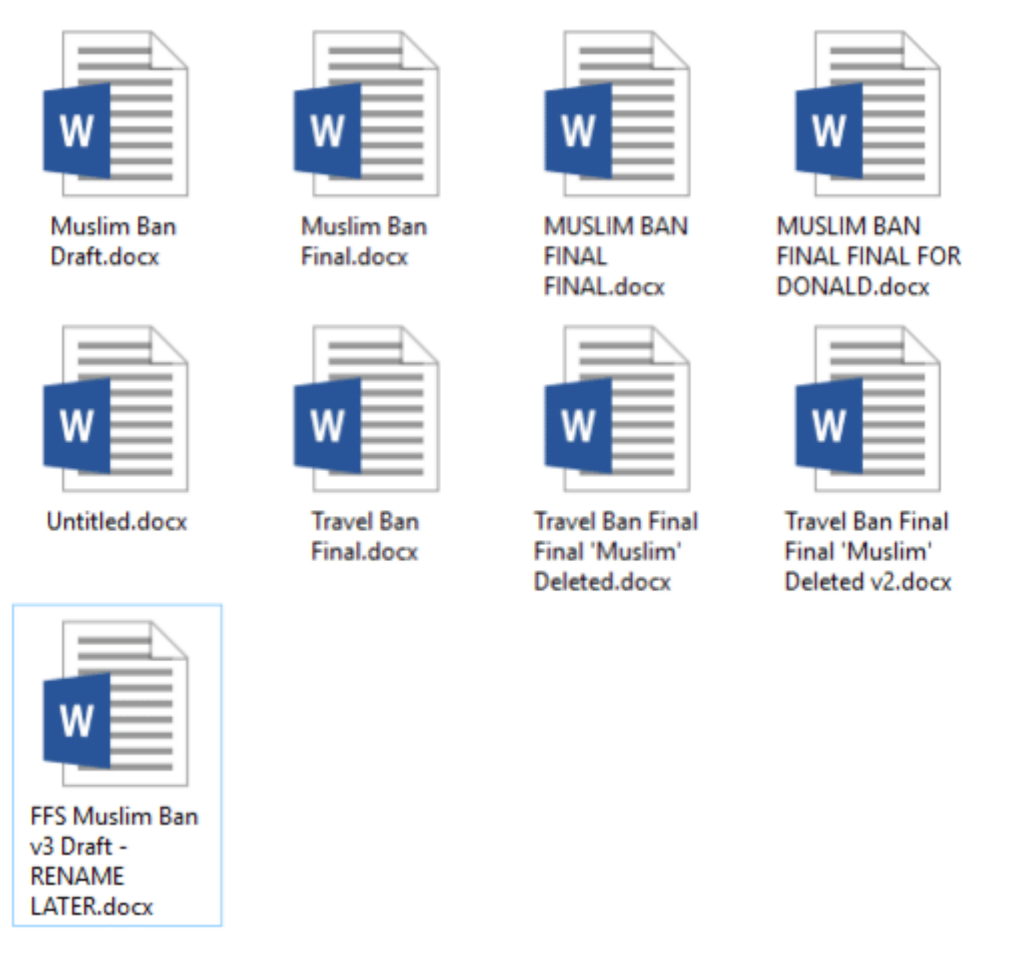
Nombrando archivos (i)

- Primero: Define los tipos de datos y formatos de archivo para la investigación.
- Datos diferentes pueden requerir diferentes convenciones de nomenclatura:
  - ¿Deben identificarse diferentes formatos de datos / archivos como parte del mismo proyecto?
- Ejemplos de información contextual en los nombres de archivo:
  - Fecha, autor o iniciales, sitio o proyecto, material.
- Las mayúsculas en los nombres de los archivos afectan el orden: sea coherente.
- Los números ordenan archivos solo si se usan ceros antes de unidades y decenas:
  - 001, 002, 003, etc. ordenará archivos hasta 999.

Nombrando archivos (ii)

- Las fechas son útiles para el control de versiones y para ordenar archivos.
  - AA-MM-DD (11-03-02) al final del nombre ordena archivos del mismo nombre por año.
  - El año en primer lugar sirve para ordenar archivos, por ejemplo, PDF de publicaciones
- Los espacios entre los nombres de archivo causan problemas en algunos sistemas. Use_underscores
- Tenga cuidado al mantener archivos en varias computadoras
- ¡LAS CAPITALES SON DIFICILES DE LEER!

Control de versiones



Control de versiones

Es MUY fácil perder el rastro de la versión más reciente de un archivo!

Especialmente:

- Archivos de Word con borradores de capítulos de tesis.
- Archivos de Word comentados por otros.
- Archivos de varios autores enviados de un lado a otro por correo electrónico.
- Gráficos y archivos de imagen.

- Sea coherente con los nombres de archivos: número de versión, iniciales, fecha.
- Ponga las versiones anteriores en una carpeta separada de "Borradores".
- Considere eliminar borradores antiguos cuando finalice la versión actual.

Formatos de archivo

En qué formato guardar las cosas para que sea seguro

- Formatos estándares adecuados para archivos de texto, imágenes, hojas de cálculo, bases de datos, audio/video
- Facilita el intercambio de datos
- Facilita el trabajo en diferentes computadoras / paquetes de software
- Adecuados para la preservación y reutilización en el futuro

Formatos de archivo: cuestiones clave

Propietario	vs	Abierto
No estándar	vs	Estándar ISO
Binario (no legible para humanos)	vs	XML Lenguaje de marcado extensible (legible para humanos)
Comprimido	vs	Sin compresión

Documentación y metadatos

Permita que otros entiendan sus datos

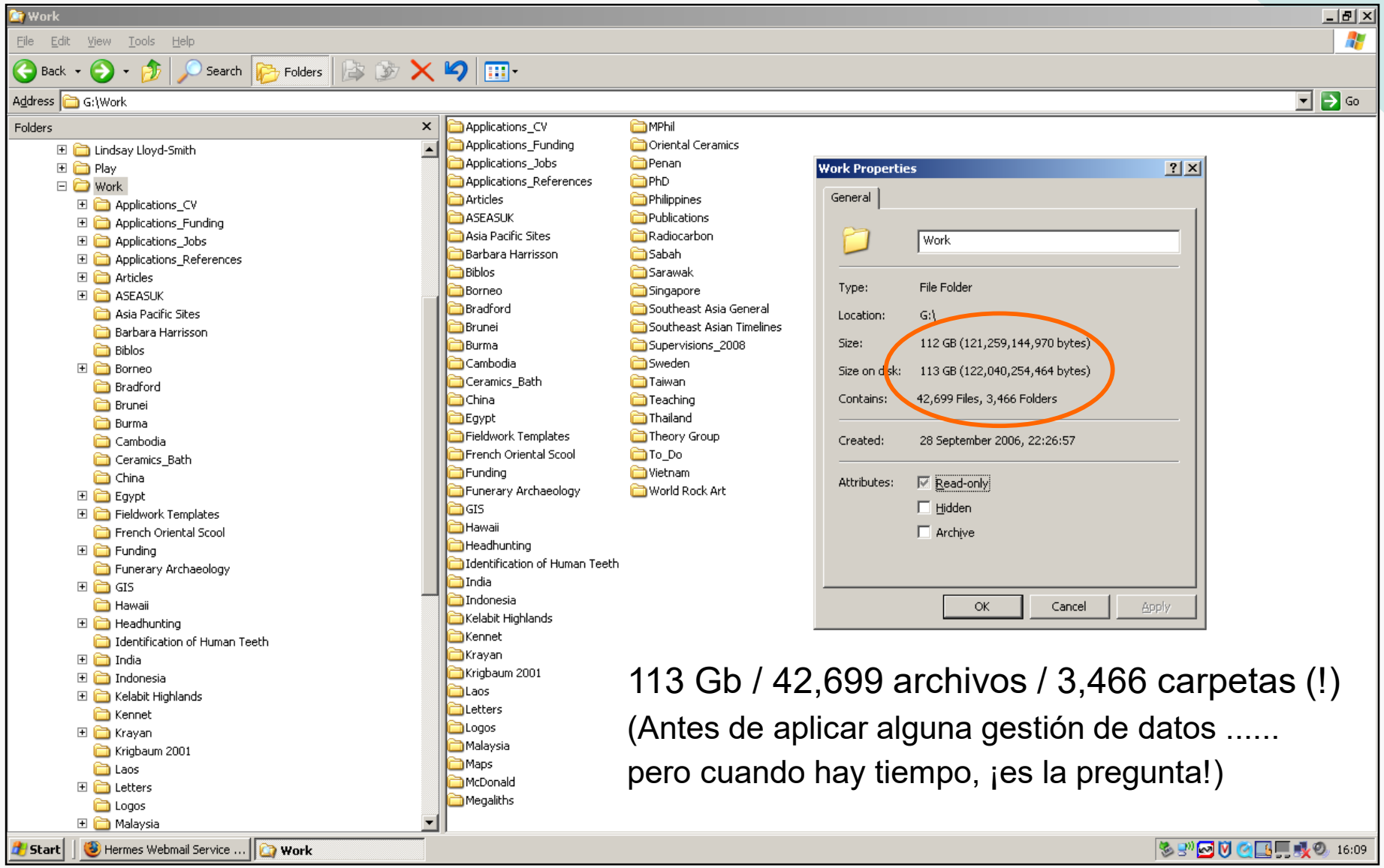
- Documentación del proyecto
  - Metodología de la Tesis: información general, estándares utilizados, etc.
  - Introducción / Apéndices: información técnica detallada, por ejemplo, explicación de los nombres de archivos y formatos utilizados, métodos y estándares de captura de datos digitales (configuraciones de escaneo, etc.).

- Documentación de archivos individuales: en metadatos o por separado
  - Datos descriptivos sobre imágenes, archivos audiovisuales, etc.
  - Explicación de encabezados, códigos, estructura y formato de hojas de cálculo y bases de datos.

Selección - ¡Tirando cosas!

- ¿Deberíamos guardar todo?
- Defina los datos principales que formarán el archivo del proyecto.
- Mantenga limpios los datos principales.
- ¿Podemos conservar los datos que otras personas nos envían?
- Tire cosas durante el proyecto.
  - Intenta no acumular varias versiones del mismo archivo.
- Almacene borradores anteriores en una carpeta separada como copia de seguridad.
  - Eliminar borradores de documentos cuando finalice el archivo.
  - Las propuestas de proyectos de investigación pueden ser útiles para consultar más adelante.
- ¿Qué hacer con los correos electrónicos?

Un disco duro típico seis años después de comenzar una maestría
(L. Lloyd-Smith)



113 Gb / 42,699 archivos / 3,466 carpetas (!)
(Antes de aplicar alguna gestión de datos .....
pero cuando hay tiempo, ¡es la pregunta!)

Selección - ¡Tirando cosas!

- ¿Deberíamos guardar todo? **NO!**

- Defina los datos principales que formarán el archivo del proyecto.
- Mantenga limpios los datos principales.
- Elimine archivos durante el proyecto, intenta no acumular varias versiones del mismo archivo.
- Almacene borradores anteriores en una carpeta separada como copia de seguridad.
- ¿Debemos conservar los datos que otras personas nos envían?
- ¿Qué hacer con los correos electrónicos?

Cosas útiles

- Derechos de propiedad intelectual
- Copyright
- Datos personales y sensibles

Derechos de propiedad intelectual

Importante descargo de responsabilidad: ***lo que sigue es una introducción muy básica****.*

- Las cuestiones con respecto a los datos de investigación son muy importantes.
- Piense cómo pueden afectar su investigación y sus datos de investigación.
- Consulte más información en sitios web de repositorios digitales, políticas de copyright del editor, contrato de trabajo.

Copyright y datos de investigación

- Copyright protege la expresión de una idea
  - no la idea en sí.
- Los datos no están cubiertos por derechos de autor
  - pero la disposición de los datos en una hoja de cálculo o base de datos, sí.
- Los derechos de autor no necesitan estar registrados
  - se asigna automáticamente cuando se produce un trabajo creativo.
- Las diferentes formas de trabajo creativo tienen derechos de autor por diferentes períodos de tiempo.
- Diferentes instituciones tienen diferentes cláusulas de derechos de autor en sus contratos de trabajo.
- Diferentes países tienen diferentes leyes de derechos de autor.

Datos personales y sensibles

- Datos relacionados con personas vivas que los identifican: nombre, edad, sexo, dirección, etc.
- Datos personales sensibles
- Datos que pueden incriminar a una persona:
  - Raza, origen étnico, opinión política, creencias religiosas, salud física / mental, orientación sexual, procesos penales o condenas.
- Datos personales que pueden considerarse confidenciales:
  - Datos conectados a una persona que los proporciona.
  - Datos que identifican a una persona (nombre, direcciones, ocupación, fotografías).
  - Datos proporcionados de forma confidencial o acordados en mantenerse confidenciales (es decir, no divulgados al dominio público).
  - Datos cubiertos por pautas éticas, requisitos legales o formularios de consentimiento de investigación.

Regreso al futuro ...

¿Qué datos producirá el proyecto?

Planifique con anticipación cuestiones de:

- Propiedad original de los datos
- Derechos de propiedad intelectual
- Datos sensibles
- ¿Qué datos serán depositados?
- ¿Dónde se pueden depositar los datos?

Consulte el repositorio digital con antelación

Cosas realmente útiles

- Tesis electrónicas
- Tesis doctorales y copyright
- Datos complementarios

Tesis doctorales electrónicas

Requisitos diferentes para las tesis electrónicas en cada universidad:

- Manuscrito en papel sin oportunidad de presentar tesis electrónicas
- Manuscrito en papel con depósito voluntario de tesis electrónicas en la biblioteca universitaria
- Presentación obligatoria de tesis electrónicas digitales

- Repositorios de tesis electrónicas
  - Hace los resultados de la investigación disponibles al público.
  - Preservación a largo plazo con identificador único.
  - Proporciona una interfaz de búsqueda y recuperación de e-tesis no embargadas.
  - Acepta texto y datos digitales complementarios para su difusión en línea.

Tesis Doctorales y Copyright

- Puede incluir material con derechos de autor.
- Una tesis manuscrita en papel sigue siendo una obra literaria inédita.
- Una tesis en formato digital que está disponible en línea es una obra literaria publicada y tiene que cumplir con la ley de derechos de autor.
- Sin embargo, esto no debería inhibir el ejercicio intelectual de un doctorado:
  - El material con derechos de autor se puede colocar en un apéndice restringido.
  - El material de copyright en el manuscrito en papel puede retirarse de la versión en línea de la tesis electrónica.
  - Se puede imponer un embargo a la difusión de la tesis.
- Consulte las orientaciones sobre tesis electrónicas y derechos de autor en las bibliotecas universitarias o repositorios digitales.

Datos complementarios

- Datos que apoyan la interpretación.
- Formatos digitales necesarios para la presentación y/o interpretación de datos.
- Los documentos digitales son más baratos y rápidos que imprimir un segundo volumen de datos.
- La tesis puede contener datos sensibles, por ejemplo, investigación de patrimonio cultural o cuestiones políticas.
- Tesis contiene una cantidad significativa de material de copyright de otra parte.

Archivo de datos complementarios: Opciones en línea

- Laboratorio de investigación / Sitios web académicos
  - Difusión de datos digitales a corto plazo.

- Repositorios digitales
  - Para archivo a largo plazo
  - Institucional (por ejemplo, bibliotecas universitarias)
  - Específico por disciplina (por ejemplo, SSRN)

Cosas esenciales

- Archivando los datos digitales

¿Por qué depositar los datos?

- Garantizar la preservación
- Proporcionar acceso
  - Potencial para vincular datos a artículos relacionados
  - Simplificar la reutilización de datos para investigación y enseñanza
- Reconocimiento profesional
  - Mayor visibilidad de su investigación
- Requisitos de organismos de financiación o de la revista

Cuándo depositar?

- Autoridad para depositar los datos
  - Permisos obtenidos en términos de Derechos sobre los datos.
  - Capaz y dispuesto a otorgar al repositorio una licencia para difundir los datos.
- El material está 'completo'
  - Archivos de proyectos terminados.
  - La entidad digital individual está completa, es decir, no es una versión de borrador.
- Formato de archivo preferido - consultar sitios web de repositorios
  - Formatos de archivo más comunes aceptados para tesis y datos
  - Formatos abiertos preferidos para preservación
- Documentación suficiente del proyecto y metadatos de archivo
  - Estructura de datos, convenciones de nomenclatura de archivos
  - Sistema operativo (plataforma), software y versión utilizada.
  - Sigue estándares y recomendaciones para conjuntos de datos

Principios de datos FAIR

- FAIR = Findable, Accesible, Interoperable, Reutilizable
  - Buenas prácticas para la publicación de datos científicos

- “Guía de Principios FAIR para el manejo de datos científicos” (https://www.nature.com/articles/sdata201618)

- “Tan abierto como sea posible, tan cerrado como sea necesario”

¿Dónde depositar?

- Repositorios institucionales
- Repositorios específicos por disciplina
- Repositorios de datos
- Repositorios de datos de revistas
- Repositorios gubernamentales

- Consultar políticas y restricciones para cada tipo de repositorio (estándares, formatos de archivo, seguridad, acceso, capacidad de descubrimiento, derechos, reutilización, costos)

Primer Congreso Internacional de Ciencia Abierta y Repositorios Institucionales

Y ahora te toca hacer…

Plan de gestión de datos para proyectos de investigación de posgrado

Plan de gestión de datos para proyectos de posgrado

Un plan de gestión de datos es un documento de proyecto formal que abarca todas las etapas alrededor del ciclo de vida de los datos:

- Formaliza la definición de sus datos de investigación;
- Documenta los detalles contextuales y técnicos de sus datos;
- Explica la estructura de archivos y protocolos de nombres;
- Describe su propuesta de lo que sucederá con sus datos en el futuro;
- Describe qué planes tiene para compartir y archivar sus datos.

Ciclo de vida de datos y

Plan de gestión de datos

1. Qué datos voy a producir
2. Cómo voy a organizar los datos
3. Qué datos voy a guardar y cómo los voy a gestionar
4. Qué datos se depositarán y dónde
5. Quiénes estarán interesado en reutilizar los datos

Regreso al futuro ...

Planifique con antelación:

- Propiedad original de los datos
- Derechos de propiedad intelectual
- Datos sensibles
- ¿Qué datos serán depositados?
- ¿Dónde se pueden depositar los datos?

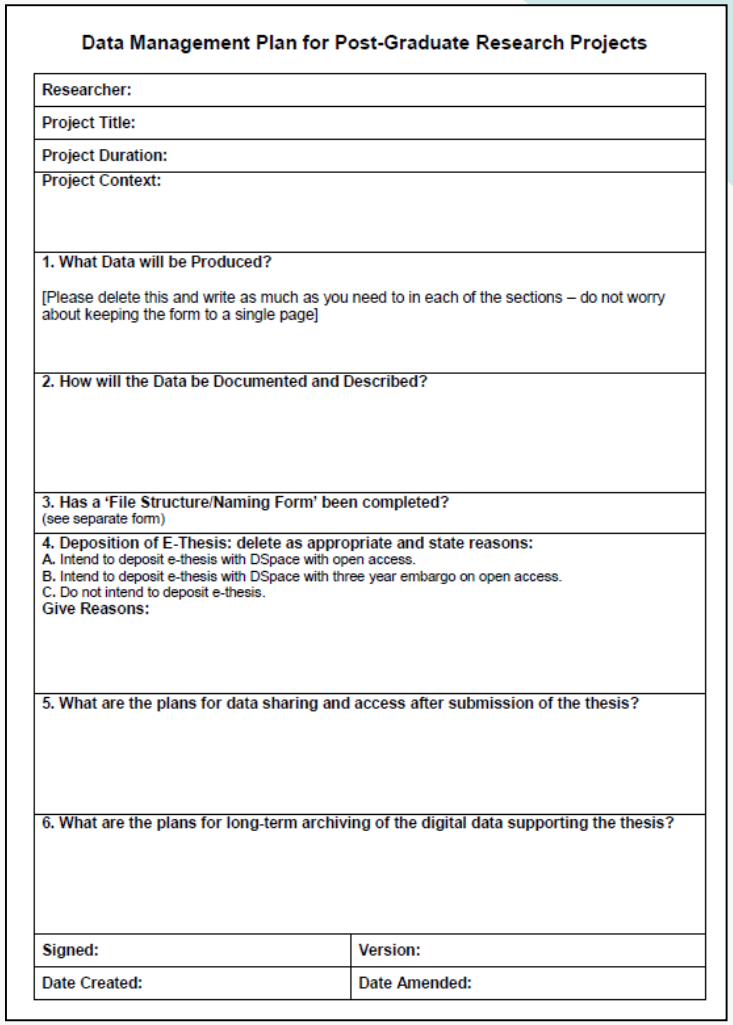
Consulte con el repositorio digital

"La mejor manera de ayudar a preservar sus datos a largo plazo es planificar su reutilización"

Plan de gestión de datos para proyectos de posgrado

- Hay varios *templates* de Plan de Gestión de Datos en internet.
- Elije uno que refleje el ciclo de vida de tus datos y empieza a llenarlo ya.
- Cuanto más detallado, mejor. Pero manténlo sencillo.
- Revisar y actualizar a cada año.

- Modelo disponible en: https://archaeologydataservice.ac.uk/resources/attach/Post-Graduate_DMP_Form.rtf



Data Management Plan for Post-Graduate Research Projects

Researcher:
Project Title:
Project Duration:
Project Context:
1. What Data will be Produced? [Please delete this and write as much as you need to in each of the sections – do not worry about keeping the form to a single page]
2. How will the Data be Documented and Described?
3. Has a 'File Structure/Naming Form' been completed? (see separate form)
4. Deposition of E-Thesis: delete as appropriate and state reasons: A. Intend to deposit e-thesis with DSpace with open access. B. Intend to deposit e-thesis with DSpace with three year embargo on open access. C. Do not intend to deposit e-thesis. Give Reasons:
5. What are the plans for data sharing and access after submission of the thesis?
6. What are the plans for long-term archiving of the digital data supporting the thesis?

Signed:	Version:
Date Created:	Date Amended:

Más referencias…

- DMPTool
  https://dmptool.org/

- DMP Template for the Social Sciences
  https://zenodo.org/record/1291816#.X661lshKg2w

- How to draft a DMP from the perspective of the social sciences
  https://forscenter.ch/fors-guides/fg-2019-00007/

- PDGOnline
  https://dmp.consorciomadrono.es/

CIENCIA ABIERTA Y REPOSITORIO INSTITUCIONALES

Primer Congreso Internacional de Ciencia Abierta y Repositorios Institucionales

Muchas gracias!

Andréa Gonçalves
http://about.me/andreafg

Fuente original:

Open Access Post-Graduate Teaching Materials for Research Data Management in Archaeology

Creado por Lindsay Lloyd-Smith (2011)

Agradecimientos

Este material fue financiado por JISC y creado por el Proyecto DataTrain con sede en la Biblioteca de la Universidad de Cambridge.
Jefe del proyecto: Elin Stangeland (Biblioteca de la Universidad de Cambridge)
Asesores de proyecto: Stuart Jeffrey (Servicio de Datos de Arqueología), Sian Lazar (Departamento de Arqueología, Universidad de Cambridge), Irene Peano (Oficial del Proyecto DataTrain: Antropología Social), Cameron Petrie (Departamento de Arqueología, Universidad de Cambridge), Grant Young (Biblioteca de la Universidad de Cambridge) y Anna Collins (responsable de datos y investigación digital de DSpace @ Cambridge Research).