La Gestión de Datos de Investigación

Marisa Pérez Aliende
Universidad Autónoma de Madrid
mp.aliende@uam.es
Sevilla, 13 de Junio 2017

Organización de los datos de investigación

- Buenas prácticas
- Gestión de ficheros
- Transformación de los datos
- Almacenamiento, *backups* y seguridad

Organización de los datos

Una buena práctica en la gestión de ficheros ayuda a: **identificar**, **localizar** y **usar** los datos de una manera efectiva, así como ayuda a otros a entender, colaborar y/o reusar los datos.

Una buena gestión de ficheros tiene que ser:

- Distinguible:

  En una carpeta, los ficheros se deben de distinguir.

- Fácil de localizar y ojear

  Las carpetas y los ficheros necesitan estar etiquetados y organizados de manera sistemática, identificables y accesibles para usuarios actuales o futuros.

- No fácilmente sobrescritos o borrados

  Organizados según una secuencia lógica.

- Fácil colaboración

  Un buen nombre de archivo de datos evita la confusión cuando mucha gente trabaja con ficheros compartidos.

Gestión de ficheros: Nombres I

Existen tres criterios que asisten a la hora de nombrar los ficheros de datos de investigación:

- **ORGANIZACIÓN:**
  Importante para el acceso futuro y la recuperación. Tener en cuenta las restricciones de nombres de archivo del sistema donde el fichero se ubica.
- **CONTEXTO:**
  Se puede incluir contenido específico o descriptivo independientemente de donde los datos estén almacenados.
- **CONSISTENCIA:**
  Importante elegir la convención de nombres y asegurarse de que las reglas se siguen sistemáticamente. Incluyendo siempre la misma información.
  Ejemplo: Fecha y hora, en el mismo orden YYYYMMDD

Elementos a considerar a la hora de nombrar carpetas y los archivos:

- Número de versión
- Fecha de creación o de publicación
- Nombre de autor / nombre del grupo de investigación / del departamento asociado con los datos
- Descripción del contenido
- Número del proyecto

Gestión de ficheros: Nombres II

Reglas para nombrar los ficheros:

- Escalables:
  - Si se incluyen nombres de proyectos no limitar a 2 dígitos. Ej.; 001 vs 01

    20150101_example_099
  - No utilizar nombres genéricos que puedan causar conficto si se cambian de lugar
  - Asegurarse de la sincronización de ficheros cuando se trabaja en más de un PC
- Nombres cortos y relevantes. Normalmente con más de 20 caracteres y no más de 32
- No usar caracteres especiales:
  - Guión bajo en lugar de punto
  - Barras en lugar de espacios
- Si se incluyen fechas que sean consistentes
- Se podría incluir información sobre el estatus del fichero (draft, final).
- Evitar el uso de etiquetas confusas como:

  Revision2

  Final_revisada

Gestión de ficheros: Nombres III

www.ed.ac.uk/records-management/records-management/staff-guidance/electronic-records/naming-conventions

Why use naming conventions?

Naming records consistently, logically and in a predictable way will distinguish similar records from one another at a glance, and by doing so will facilitate the storage and retrieval of records, which will enable users to browse file names more effectively and efficiently. Naming records according to agreed conventions should also make file naming easier for colleagues because they will not have to 're-think' the process each time.

File naming conventions

The conventions comprise the following 13 rules. Follow the links for examples and explanations of the rules.

- Keep file names short, but meaningful
- Avoid unnecessary repetition and redundancy in file names and file paths.
- Use capital letters to delimit words, not spaces or underscores
- When including a number in a file name always give it as a two-digit number, i.e. 01-99, unless it is a year or another number with more than two digits.
- If using a date in the file name always state the date 'back to front', and use four digit years, two digit months and two digit days: YYYYMMDD or YYYYMM or YYYY or YYYY-YYYY.
- When including a personal name in a file name give the family name first followed by the initials.
- Avoid using common words such as 'draft' or 'letter' at the start of file names, unless doing so will make it easier to retrieve the record.
- Order the elements in a file name in the most appropriate way to retrieve the record.
- The file names of records relating to recurring events should include the date and a description of the event, except where the inclusion of any of either of these elements would be incompatible with rule 2.
- The file names of correspondence should include the name of the correspondent, an indication of the subject, the date of the correspondence and whether it is incoming or outgoing correspondence, except where the inclusion of any of these elements would be incompatible with rule 2.
- The file name of an email attachment should include the name of the correspondent, an indication of the subject, the date of the correspondence, 'attch', and an indication of the number of attachments sent with the covering email, except where the inclusion of any of these elements would be incompatible with rule 2.
- The version number of a record should be indicated in its file name by the inclusion of 'V' followed by the version number and, where applicable, 'Draft'.
- Avoid using non-alphanumeric characters in file names.

Gestión de ficheros: Versionado

Siempre hay que registrar los cambios en los ficheros aunque parezca innecesario:

- Números ordinales para cambios mayores
  V1, V2, V3
- Números decimales para cambios menores
  V1.1, V1.2,…V1.6
- Borrar las versiones obsoletas
- Utilizar softwares para controlar las versiones:
  Subversion
  TortoiseSVN
  Box

File name	Changes to file
Interviewschedule_1.0	Original document
Interviewschedule_1.1	Minor revisions made
Interviewschedule_1.2	Further minor revisions
Interviewschedule_2.0	Substantive changes

Renombrar ficheros:
Existen herramientas para el *batch renaming,* renombrar por lotes*:*

NameChanger. Diseñado para renombrar listas de ficheros de manera fácil y rápida.

Gestión de ficheros: Ejemplos

Best Practice	Example
Limit the file name to 32 characters (preferably less!)	32CharactersLooksExactlyLikeThis.csv
When using sequential numbering, **use leading zeros** to allow for multi-digit versions For a sequence of 1-10: 01-10 For a sequence of 1-100: 001-010-100	**NO** ProjID_1.csv ProjID_12.csv **YES** ProjID_01.csv ProjID_12.csv
Don't use special characters & , * % # ; * ( ) ! @$ ^ ~ ' { } [ ] ? < > -	**NO** name&date@location.doc
Use only one period and use it before the file extension	**NO** name.date.doc **NO** name_date..doc **YES** name_date.doc
Avoid using generic data file names that may conflict when moved from one location to another	**NO** MyData.csv **YES** ProjID_date.csv

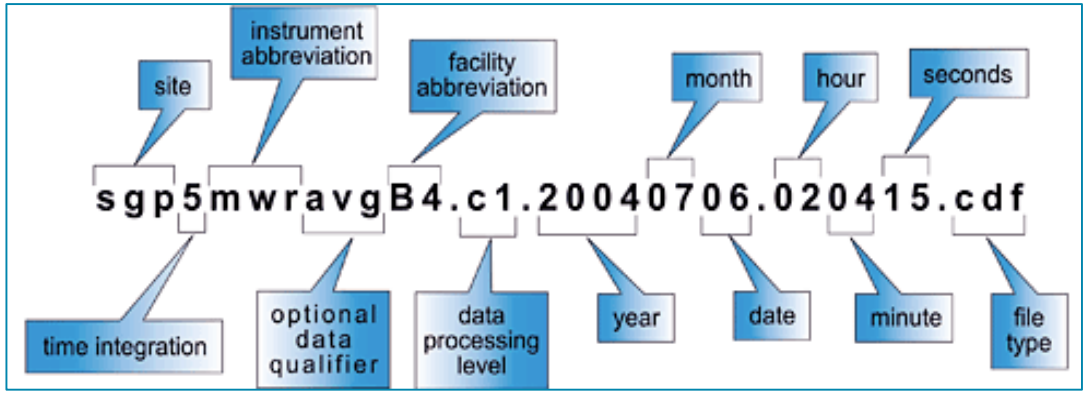
(Malinowski "Data Management. File organization. 2017. http://libraries.mit.edu/data-management/files/2014/05/FileOrgSlides_20170118sm.pdf)

For example…

sam_monarch_wing_20160115_CM_001.tif
[instrument]_[item]_[date]_[collector]_[ascension#].ext

FileOrgSlides_20170118.pptx
[class][material]_[date].ext

SevilletaLTER_NM_2001_NPP.csv
[project name]_[state]_[year]_[dataset].ext
SevilletaLTER_NM_2001_NPP_20170117.csv
[project name]_[state]_[year]_[dataset]_[analysisID].ext



Convenciones por disciplinas en:

- **DOE's Atmospheric Radiation Measurement (ARM) program**
- **GIS datasets from Massachusetts**
- **The Open Biological and Biomedical Ontologies**

Gestión de ficheros: Directorio

Example file structure systems/directory hierarchy conventions:

/[Project]/[Grant Number]/[Event]/[Date]
/[Project]/[Sub-project]/[Run of an experiment]/[Person]/[Date]
/[Research area]/[Project]/[Data vs. documentation]/[Date]
/[Project]/[Type of file]/[Person]/[YYYYMMDD]
/[Instrument]/[Date]/[Sample]

For the butterfly project:

/butterfly/images/mcneill/20140117
/butterfly/tabular/mcneill/20140117
/butterfly/projectDocs/
/butterfly/literature/subject/

IITLibraries

(McNeill, D. y Bailey, H. “Research Data Management: File organization. 2014. https://libraries.mit.edu/data-management/files/2014/05/file-organization-july2014.pdf)

Gestión de ficheros: Formatos y Migración

Codificar información sobre un fichero que permite ser reconocido por un programa o aplicación informática.

- **Formatos propietarios** vs **abiertos:**
  - **Propietarios**: Sólo se pueden abrir con el software usado para crear el fichero. Ej.: Excel, Word, GIF
  - **Abiertos:** Pueden ser reconocidos por otras aplicaciones. Ej.: CSV, PDF/A, TIFF

    Si se quiere que los datos se reusen, lo mejor es utilizar estándares internacionales.

    Obsolescencia del formato:
    - Cambios en la tecnología
    - Actualización del software
    - Colaboración entre plataformas
- Ampliamente utilizados por la comunidad investigadora
- Codificados en ASCIII, UTF-8 y Unicode
- Formatos no comprimidos

MIGRACIÓN VS **NORMALIZACIÓN:**

MIGRACIÓN: Se refiere a la conversión de ficheros cuando el formato corre el riesgo de quedarse obsoleto.

NORMALIZACIÓN: Es la práctica de convertir formatos de ficheros cara a la preservación a largo plazo.

Ambos implican convertir ficheros de un formato a otro, generalmente son *preservation-friendly,* y formatos abiertos.

Suele ser una buena idea normalizar los ficheros para asegurar la preservación y evitar la migración.

Gestión de ficheros: Formatos

Type	Recommended	Avoid for data sharing
Tabular data	CSV, TSV, SPSS portable	Excel
Text	Plain text, HTML, RTF PDF/A only if layout matters	Word
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data	XML, RDF	RDBMS

(Sarah Jones. "Managing Research Data and H2020". 2015)

Some preferred file formats

- Containers: TAR, GZIP, ZIP
- Databases: XML, CSV
- Geospatial: SHP, DBF, GeoTIFF, NetCDF
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabular data: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC

(Stanford University, best practices for file formats)

Otros recursos:

UK Data Archive
DPC- File Formats Standards
Open Data Handbook – File Formats
Edinburgh DataShare: Recommended File Formats

Proprietary Format	Alternative/Preferred Format
Excel (.xls, .xlsx)	Comma Separated Values (.csv) ASCII
Word (.doc, .docx)	plain text (.txt), XML, PDF/A, HTML, ODF or if formatting is needed, PDF/A (.pdf)
PowerPoint (.ppt, .pptx)	PDF/A (.pdf), ODP, JPEG 2000, PDF, PNG
Photoshop (.psd)	TIFF (.tif, .tiff),
Quicktime (.mov)	MPEG-4 (.mp4), MOV, AVI, MXF
Sounds	WAVE, AIFF
Containers	TAR, GZIP, ZIP
Databases	XML, CSV

Documentación

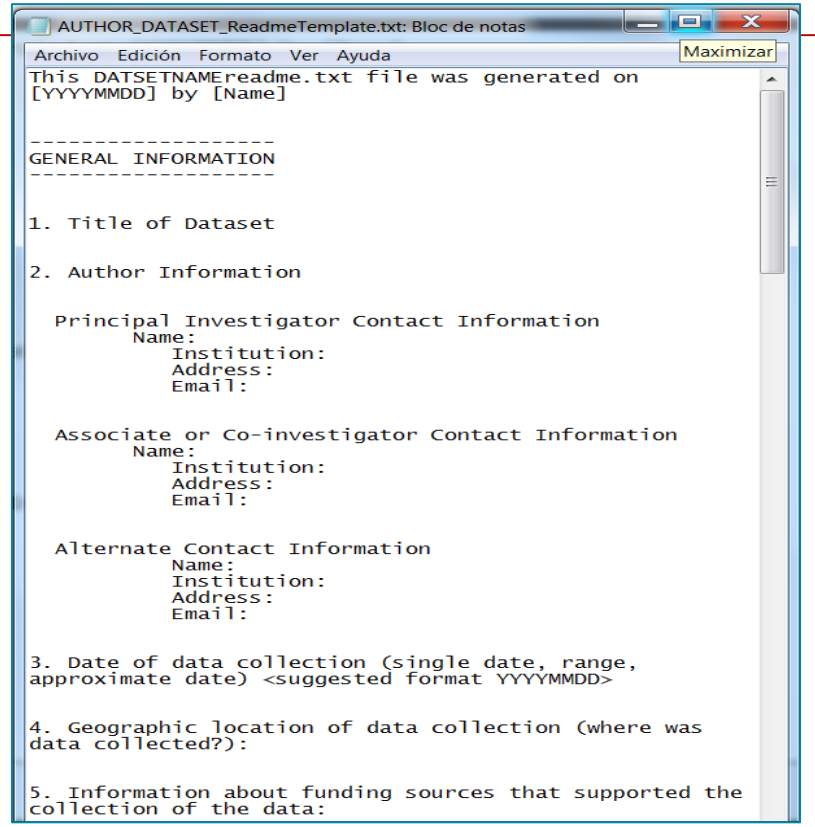
Documentación de datos son:

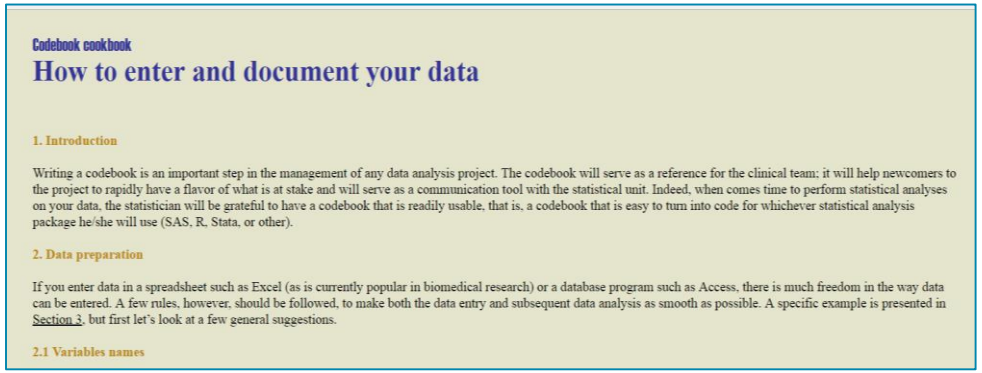
cuadernos de laboratorio y protocolos experimentales, cuestionarios, libros de códigos, diccionarios de datos, software y su sintaxis, información sobre la configuración del equipo y calibración, esquemas de las bases de datos, metodología, etc.

- Se suelen resumir en 2 tipos:
  - **Readme files (readme.txt)**:
    - Fichero que ayuda al lector a identificar, evaluar, usar y comprometerse con el proyecto.
    - Proporciona información sobre el dataset para que sea correctamente interpretado por personas y máquinas.
    - Un readme por dataset, o uno muy desarrollado para varios

      Plantilla para escribir un readme file
      (University of Cornell. Guidelines for writing “readme” style metadata)

  - **Codebooks y data dictionary (**proporciona descripción detallada de cada elemento o variable del dataset)




Codebook cookbook

How to enter and document your data

1. Introduction

Writing a codebook is an important step in the management of any data analysis project. The codebook will serve as a reference for the clinical team; it will help newcomers to the project to rapidly have a flavor of what is at stake and will serve as a communication tool with the statistical unit. Indeed, when comes time to perform statistical analyses on your data, the statistician will be grateful to have a codebook that is readily usable, that is, a codebook that is easy to turn into code for whichever statistical analysis package he/she will use (SAS, R, Stata, or other).

2. Data preparation

If you enter data in a spreadsheet such as Excel (as is currently popular in biomedical research) or a database program such as Access, there is much freedom in the way data can be entered. A few rules, however, should be followed, to make both the data entry and subsequent data analysis as smooth as possible. A specific example is presented in Section 3, but first let's look at a few general suggestions.

2.1 Variables names

Gestión de ficheros: *Backup, a*lmacenamiento y seguridad

Backup preservar copias adicionales de los datos mientras se realiza la investigación

- Se recomiendan 3 copias, en al menos, 2 localizaciones diferentes. Chequeadas regularmente
- Incremental (de lo que va cambiando) vs *full (* de todos los datos)
- Preparar una programación ej.: rotación abuelo-padre-hijo permite que estén disponible más tiempo antes de que se sobrescriban

Almacenamiento

- Preservar los ficheros de datos en una localización segura fácilmente accesible
- En pc o portátiles personales, en dispositivos extenos (cd, dvd, usb...no utilizar para copias máster, aunque los hay de calidad y hay que revisarlos regularmente y copiar de un disco a otro), servidores de las propias instituciones y servicios de *cloud* (Dropbox, OneDrive, GoogleDrive).

  ¡¡¡Cuidado!!!: los servidores pueden estar fuera de la UE, y tener otra legislación en relación con los datos

Seguridad

- Evitar la corrupción de los datos y controlar el acceso para impedir:
  - Modificar los datos accidental o maliciosamente
  - Robo
  - Romper con la confidencialidad

para ello ENCRIPTAR:
Proceso de convertir tus datos en código que no se pueda leer

La Gestión de Datos de Investigación

Marisa Pérez Aliende
Universidad Autónoma de Madrid
mp.aliende@uam.es
Sevilla, 13 de Junio 2017