The University of Edinburgh

EDINA

Organising and Documenting Data

Stuart Macdonald
EDINA & Data Library
stuart.macdonald@ed.ac.uk

DIY Research Data Management Training Kit for Librarians

Organising your data

- RDM is one of the essential areas of responsible conduct of research.
- Research data files and folders need to be organised in a systematic way to be:
  - identifiable and accessible for yourself,
  - identifiable and accessible for colleagues, and for future users.
- Thus it is important to plan the organisation of your data before a research project begins.
- Doing so will prevent any confusion while research is underway or when multiple individuals will be editing and / or analysing the data.

This can be achieved through:

- Directory structure & file naming conventions
- (File naming conventions for specific disciplines)
- File renaming
- File version control

For this to be successful a consistent and disciplined approach is required.

Easier to accomplish as and when data files are generated rather than retrospectively attempting to implement.

When organization methods become too time consuming, consider automated methods.

File Naming conventions

- Naming datasets according to agreed conventions should make file naming easier for colleagues because they will not have to ‘re-think’ the process each time.
- File names should provide context for the contents of the file, making it distinguishable from files with similar subjects or different versions of the same file.
- Many files are used independently of their file or directory structure, so provide sufficient description in the file name.
- Suggested strategies: identify the project; avoid special characters; use underscores rather than spaces; include date of creation or modification in a standard format (e.g. YYYY_MM_DD or YYYYMMDD): use project number
- Be consistent! Avoid being cryptic!

Batch (or bulk) renaming

- Software tools exist that can organise data files and folders in a consistent and automated way through batch renaming.
- There are many situations where batch renaming may be useful, such as:
  - where images from digital cameras are automatically assigned filenames consisting of sequential numbers
  - where proprietary software or instrumentation generate crude, default or multiple filenames
  - where files are transferred from a system that supports spaces and/or non-English characters in filenames to one that doesn't (or vice versa). Batch renaming software can be used to substitute such characters with acceptable ones.

Benefits of consistent data file labelling are:

- Data files are not accidentally overwritten or deleted
- Data files are distinguishable from each other within their containing folder
- Data file naming prevents confusion when multiple people are working on shared files
- Data files are easier to locate and browse
- Data files can be retrieved both by creator and by other users
- Data files can be sorted in logical sequence
- Different versions of data files can be identified
- If data files are moved to other storage platform their names will retain useful context

Version Control

It is important to consistently identify and distinguish versions of data files.

This ensures that a clear audit trail exists for tracking the development of a data file and identifying earlier versions especially if data is frequently updated by multiple users.

Suggested strategies:

- Use a sequential numbered system: v1, v2, v3, etc.
- Don't use confusing labels: revision, final, final2, etc.
- Record all changes -- no matter how small
- Discard obsolete versions (but never the raw copy)
- Use auto-backup instead of self-archiving, if possible

The alternative is to use version control software. (Bazaar, TortoiseSVN, SubVersion)

Documenting Data

There are many reasons why you need to document your data:

- To help you remember the details later
- To help others understand your research
- Verify your findings
- Replicate your results
- Archive your data for access and re-use

Some examples of data documentation are:

- Laboratory notebooks
- Field notes
- Questionnaires
- SOPs
- Methodologies

Documenting Data

Laboratory or field notebooks, for example play an important role in supporting claims relating to intellectual property developed by University researchers, and even defending claims against scientific fraud.

Research data need to be documented at various levels:

- Project level
  - study background, methodologies, instruments, research hypothesis
- File or database level
  - formats, relationships between files
- Variable or item level
  - How variable was generated & label descriptions

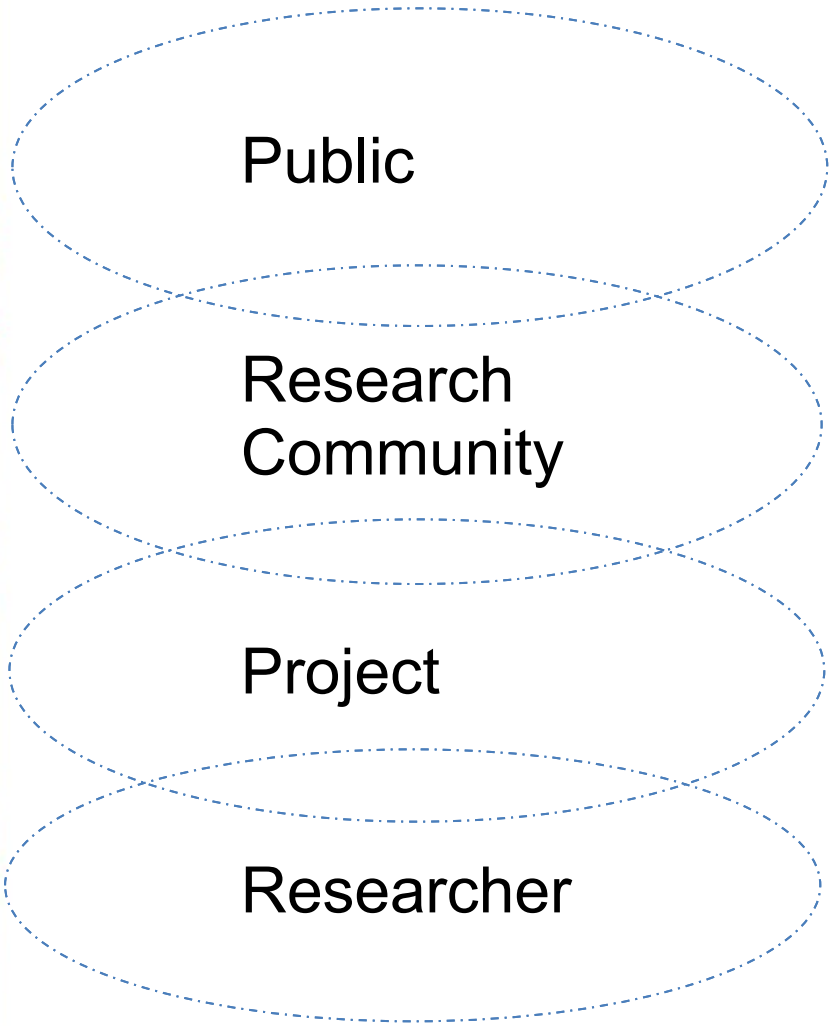
Metadata – ‘data about data’

The difference between documentation and metadata is that the first is meant to be read by humans and the second implies computer-processing (though may also be human-readable) to assist location and access to data through search interfaces.

Three broad categories of metadata are:

- *Descriptive* - common fields such as title, author, abstract, keywords which help users to discover online sources through searching and browsing e.g. DC, MARC
- *Administrative* - preservation, rights management, and technical metadata about formats.
- *Structural* - how different components of a set of associated data relate to one another, such as a schema describing relations between tables in a database.

Need for metadata



Metadata may not be required if you are working alone on your own computer, but become crucial when data are shared online.

Metadata help to place your dataset in a broader context, allowing those outside your institution, discipline, or research environment to understand how to interpret your data.

THANK YOU!