Selecting data: what to keep, what to delete?

Why select data?

A selection or appraisal process is essential in order to manage your research data effectively. Deciding what should be kept and what can be disposed of saves on storage costs, enables you to find meaningful data more easily, and ensures you abide by any legal requirements for the data's retention and reuse.

How do I decide what to keep?

Choosing which research data to keep will always involve subjective judgement, as no one can know exactly what information could be useful in the future. The following checklist outlines some of the things to consider when assessing your data:

1.   Does the data underpin a research publication?

Data that underpin research publications should be kept to ensure the integrity, transparency and robustness of the research, allowing others to confirm or challenge research results.

2.   Are there any relevant institutional or legal requirements?

- Does the University/research group have a retention policy for this type of data?
- Does your funding body have a data policy that specifies a retention period for the project's data?
- Is the data affected by legislation such as Data Protection, Freedom of Information or copyright?

3.   Does the data have scientific or historical value?

- Is the data vital to your project?
- Has it been used again in subsequent projects or research?

Adapted from: http://www.lib.cam.ac.uk/dataman/resources/Incremental_factsheet_Cambridge_selection.pdf

4.   Is the dataset unique?

- Does it duplicate existing work or is it unique?
- Do other copies exist elsewhere, and if so will they be preserved?

5.   Can the data be reused?

- Are there any intellectual property rights (IPR) issues relating to sharing or reuse of the data?
- Are human subjects involved and was consent given for archiving or reusing the dataset?
- Is the dataset in a format that allows others to reuse it without cost or other restrictions?

6.   Can the data be easily recreated or replicated?

- Is the data recording one-off events that cannot be recreated?
- Can the data be replicated or re-measured without considerable cost or new external funding?

7.   Is retaining the data going to be cost effective?

- Have you considered the costs of long-term preservation of this data?
- Do you have the funds available to do this?

8.   Has the data been effectively documented?

- Is there a "readme.txt" file explaining things such as field names and the context of the data?
- Is there sufficient documentation to allow the data to be found wherever it is stored?

For further help and advice

Contact the Open Access and Data Curation team or email rdm@exeter.ac.uk

Adapted from: http://www.lib.cam.ac.uk/dataman/resources/Incremental_factsheet_Cambridge_selection.pdf