Data Management Checklist

Managing research data throughout its lifecycle ensures its long-term value and prevents data from falling into digital obsolescence. Proper data management is a key prerequisite for effective data sharing throughout the scientific community. This, in turn, increases the visibility of scholarly work and is likely to increase citation rates.

Many funding organizations prescribe the use of data management plans and insist on open access publication of the research results they funded.

Even if a funding body does not explicitly demand data management, following professional curation and preservation concepts has numerous advantages:

(a) It greatly facilitates the reuse of research data.

(b) As a result, this increases the impact of research results.

(c) It saves precious research funds and ultimately natural and human resources by avoiding unnecessary duplication of work.

Today, the availability of well-managed data is part of good scientific practice and ensures the reproducibility of research results, a key requirement at the core of the research process.

The following data management checklist is based on a generalised research data lifecycle, and is flexible enough to be applied to requirements from different funding organisations.

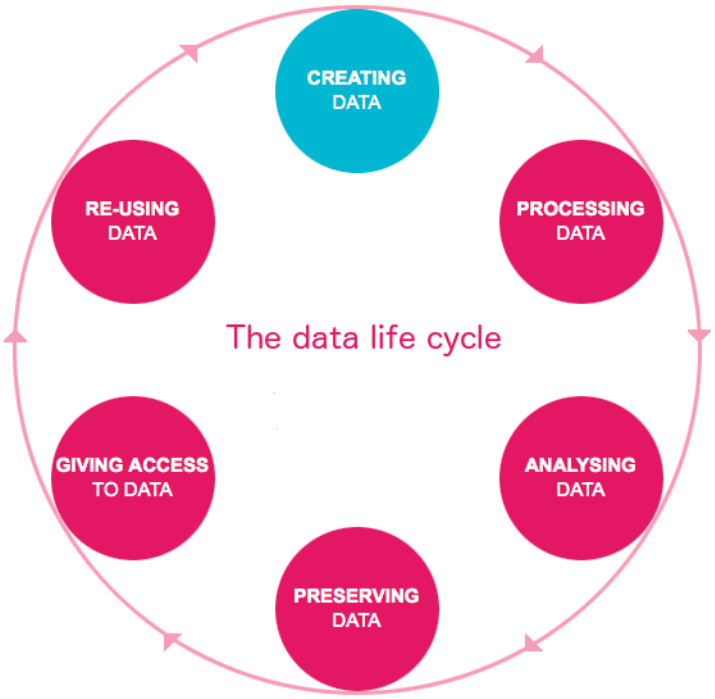

Figure 1 - Data lifecycle according to UK Data Archives (www.data-archive.ac.uk/create-manage/life-cycle)

Data management checklist

Planning

- State the project title, aim of the research and project duration
- List the principal investigator, researchers and project members as well as collaborators and partner institutions.
- Who is responsible for (which part of) data management?
- What (additional) resources will be required to implement the necessary data management (e.g. training for employees, data repository fees, funding for special hard- or software)?

Data collection and creation

- What type (e.g. observational data, experimental data, statistical data, survey results…) will be created?
- Will you be reusing existing datasets, and if so, from where?
- What is the expected volume (in GB or TB) of data created within the project?
- How is data creation distributed over time?
- Have you ensured that adequate resources for handling the data are available when needed?
- Is the created data reproducible (e.g. experiment) or irreproducible (e.g. observations)?
- Will quality assurance processes be adopted?
- How will the versioning be handled?
- Will special standards or methodologies be used?

Appraisal and selection

- By which criteria will data be appraised and selected for further work within the project?
- What tools, if any, will be used for appraising and selecting data?

Documentation and metadata

- What information will be needed for the data to be read and interpreted in the future?
- Are special tools or software needed to read and work with the data?
- How will the data be labelled and organized (file and folder naming conventions)?
- Are your data self-explanatory in terms of variable names, codes and abbreviations used?

- How will metadata be captured, created and managed? Is there a discipline-specific standard?
- What other documentation and contextual information will be available in order to help others understand the data (e.g. data dictionaries, questionnaires)?
- What metadata standards will be used?
- Will any unique identifiers be used?

File Formats

- What file formats of data will be produced?
- Do your chosen formats and software enable sharing and long-term access to the data?
- Are you using a format that is standard in your field? If not, how will you document the alternative format you are using?
- Do these formats conform to an open standard and/or are they proprietary?
- When converting across formats, how will it be ensured that no data, annotation or internal metadata have been lost or altered?

Storage

- Where and on what media will the data be stored?
- Which data will be stored and for how long?
- What data must be retained/ destroyed for contractual, legal, or regulatory purposes?
- How will data be filtered, appraised and selected to effectively separate data to be retained from data to be destroyed?
- How will the back-up be organized (frequency and responsibilities)?
- What are the risks to data security?
- How will data security be guaranteed (e.g. encryption or recovery after an incident)?
- Is the available storage sufficient or will you need to invest in additional services?

Ethics

- Are there any ethical and privacy issues concerning your data in general and sharing it in particular?
- If so, have you sought guidance from your institution's contact person(s) for research ethics and/or data protection issues?
- Does your data contain confidential or sensitive information? If so, have you discussed data sharing with the respondents from whom you collected the data and gained their written consent if needed?

- Does the data need to be anonymized prior to sharing?
- How will the sensitive data be handled to ensure it is stored and transferred securely?

Copyright and intellectual property

- Who owns the data arising from your research, and the intellectual property rights relating to them?
- Are there requirements of your institution regarding the Exploitation of Research Results to be followed?
- How will the data be licensed for reuse (e.g. Creative Commons)?
- Are there any restrictions on the reuse of third-party data?
- Will data sharing be postponed or restricted due to e.g. publishing or patenting?

Sharing

- Do you intend to make all your data available for sharing or will you select certain data only (if so, on which grounds)?
- Will there be any limits to data sharing (e.g. embargo periods, contracts, non-disclosure agreements)?
- What tools/software will be needed to work with the data?
- How will the data be discovered and shared?
- In which repository do you plan to deposit and share your data? Is this a trusted and sustainable repository?
- Do you wish to make available your data through a certain publisher according to a specific data policy?
- Does your institution or funding agency mandate data sharing?
- In addition to the owners of the data you generate, who else has a right to see or use this data? And who else should have access? Who will be the audience for your data?
- Will a data sharing agreement (or equivalent) be required?
- With whom will you share data and under what conditions?

Long term management

- What data will be kept or destroyed after the end of the project?
- Are there requirements on how long data needs to be preserved?
- In which repository or data archive will the data be stored in the long run?
- Are the chosen file formats long-lived?

- Who will manage the long-term data?
- What is needed to prepare the data for preservation or future sharing (e.g. after an embargo period or after the data creators passed away)?
- What related information will be deposited with the data?
- What are the foreseeable research uses for the data?
- Are there additional costs that come with using the repository or data archive of your choice?
- Did you anticipate these costs for using the repository or data archive? How will you cover such costs?

Contact

ETH Zürich
ETH-Bibliothek
Digital Curation Team
data-archive@library.ethz.ch
http://www.library.ethz.ch/Digital-Curation

EPFL
EPFL Library
Research Data Team
datamanagementplan@epfl.ch
http://library.epfl.ch/research-data