Twenty Questions for Research Data Management

These twenty questions are designed to prompt and assist your thinking, as a research student, a postdoc or an academic researcher at the beginning of a research project, and to form the basis of a workable research data management plan that can both guide your on-going data management activities and inform others about the nature and availability of your research data.

They will help you determining how best to safeguard your data from loss, how to describe your datasets in ways that assist both yourself when returning to them in the future and others in their subsequent interpretation, and how to publish your data in ways that maximize their usefulness to others and bring maximum academic scholarly credit to yourself, to reward your efforts in acquiring, analysing, describing, interpreting and publishing them in the first place.

You may not have immediate answers to all these questions. But, by seeking advice from your research supervisor, colleagues and others in your institution with responsibilities for data management, you should endeavour to discover them. Then, once in a while, you should revisit these questions and see whether your data management practices can be improved, updating your answers.

The nature of your data

1 What is the subject discipline (domain, field) to which your research data relates?

Possible responses:

Quantum physics.

Cell biology.

Ornithology.

2 What is the exact nature (range, scope) of your research data?

Possible responses:

Long-distance quantum communication using entangled photons.

Protein chemistry and electron microscopy of cell membrane proteins.

Video field recordings of avian behaviour, and their quantitative analysis.

3 In what format(s), will you store your data in the short term after acquisition?

Possible responses:

Questionnaire response data will be stored on my laptop in a Microsoft Office Access 2007 database.

Raw video recording on digital video tapes on the shelf above my desk, edited videos in .mov format on my laptop. numerical analyses in a spreadsheet (Microsoft Office Excel 2007 format) on my laptop.

On my research group's cloud-based secure DataStage research data file store, in Zeiss confocal 3D image format.

4 Who owns the data arising from your research, and the intellectual property rights relating to them?

Possible responses:

Myself alone.

Myself and my research group leader.

My university.

Data descriptions (metadata, "data about data")

5 How will your research datasets be described?

Possible responses:

The only description will be the filenames on my hard drive.

The only description will be the column and row labels in my spreadsheets.

The data will be described in handwritten notes in my lab notebook.

I will save metadata describing the data files in electronic form.

6 How will these descriptive metadata be created or captured?

Possible responses:

Instrument metadata are automatically included in each data file.

The only metadata will be the title and short textual description that I will manually complete in the Web submission form, when depositing each dataset in my university's data repository.

My data descriptions will be saved in spreadsheets or word processor documents.

Rich metadata conforming to a Minimal Information Standard appropriate to my research field will be recorded at the time of data acquisition, using a metadata entry form, and will thus be available as a metadata file to accompany my datasets during submission of the data to a data repository.

Data sharing

7 With whom will you share your research data in the short term, before publication of any papers arising from their interpretation?

Possible responses:

My research supervisor only.

Members of my research group and trusted external collaborators.

Anyone who asks for them.

Everyone, by publishing the data online, since our research community is committed to the rapid sharing of research results.

Data storage and backup

8 Where will you store your data in the short term, after acquisition?

Possible responses:

On my laptop.

On the computer connected to the microscope.

On the research group's DataStage filestore.

9 Who is responsible for the immediate day-to-day management, storage and backup of the data arising from your research?

Possible responses:

Myself alone.

My research group's data manager.

Our departmental IT staff, who manage our research group's DataStage research data management system.

10 How frequently will your research data be backed up for short-term data security?

Possible responses:

Whenever I remember to do so.

Nightly, using our research group's DataStage research data management system connected to the University's automated backup service.

Data archiving

11 Where will your research data be archived for long-term preservation?

Possible responses:

Selected data will be included in the figures and tables of research papers published by my research group, but we have no plans to archive and publish the full datasets.

As supplementary files attached to my journal articles on the publisher's web site.

In the University's DataBank data repository, run by the library service.

In appropriate genomics databases run by the European Bioinformatics Institute.

12 When will your research data be moved to a secure archive for long-term preservation and publication?

Possible responses:

Our research data are already securely stored in an institutional data server.

Nightly.

Upon completion of each set of experiments.

When my research group leader decides it is appropriate.

Immediately after publication of my thesis.

Upon submission of our *Nature* paper, so that the data are available for reviewers.

13 Who will decide which of your research data are worth preserving?

Possible responses:

Myself alone.

Myself, in consultation with my research supervisor.

My research supervisor alone.

14 How (i.e. by what physical or electronic method) will you transfer your research datasets to their long-term archive, under the curatorial care of a separate third-party, e.g. a data repository?

Possible responses:

On physical hard drives that I will bring back from my field site by air.

By e-mailing files to our librarian.

By completion of the Web-based database submission form and uploading of the data files over the Internet.

By automated data packaging and repository submission over the Web from my local DataStage filestore, using the SWORD repository submission protocol.

Data publication

15 For how long will you embargo your research data before it is published for others to see and use?

Possible responses:

We will allow immediate public access to the data.

For one year, to permit us to exploit our hard-won research results.

Until the journal article describing our results has been published.

16 Why is public access to your research data to be restricted (if indeed it is)?

Possible responses:

We intend to make a patent application, and must avoid prior disclosure.

Don't want to make locations of members of endangered species available to poachers.

The research data are confidential because of the arrangement my research group has made with the commercial partner sponsoring our research.

My data form part of a long-term study upon which my research group is entirely reliant for its on-going research publications and academic reputation. We only share this with trusted colleagues.

Confidential human patient data.

Questionnaire data collected in confidence from individuals – anonymized averaged data *will* be published.

17 Under what data-sharing license will you publish your research data?

Possible responses:

What is a data-sharing license?

Under a Creative Commons Open Data CC Zero public domain dedication and waiver, since my research data are not covered by copyright.

Using a Creative Commons Attribution License, since my image data are copyrightable.

18 What persistent identifiers will be used to permit correct citation of your datasets?

Possible responses:

A Digital Object Identifier (DOI) issued by DataCite.

The accession number for the dataset issued by the database to which it is submitted.

19 What metadata will be published with the data to make them interpretable and reusable?

Possible responses:

I will expect users to be able to interpret the column and row labels in my spreadsheets.

The dataset will be described in the journal article we will publish, but will have no other metadata beyond those required by the repository for data citation: Author, Date, Title, Source, Identifier.

An XML metadata file created in conformance with a Minimal Information standard will be submitted to the repository as part of the data package, along with the data files.

Future data management

20 Who will be responsible for your data, once you have left your present research group?

Possible responses:

At this stage, I have no idea.

I'll take my data with me and maintain responsibility.

My supervisor will make appropriate arrangements.

I hope the journal will maintain access to the supplementary information files associated with my article.

My University will assume long-term responsibility for the data I have chosen to preserve in its data archive.

- - - - -

Notes

Creative Commons: Creative Commons is a non-profit organization that has developed a legal and technical infrastructure for the licensing of copyright material and data in a standardised and machine-readable manner, thereby facilitating open publication, sharing and innovation in the digital age.

DataCite: DataCite is an international organization that manages the issue of DOIs (Digital Object Identifiers) for datasets. (DOIs are more commonly used to identify journal articles).

DataStage and DataBank: DataStage is a simple research data filestore and repository data submission system, designed for deployment at the research group level. DataBank is a data repository for archiving and publishing research data, designed for deployment at the institutional level. Both are open-source services for local or cloud deployment developed together at Oxford University within the JISC University Modernization Fund DataFlow Project, and both are now available in beta versions for third-party installation and use. Full Version 1.0 releases of both DataStage and DataBank are scheduled for May 2012.

European Bioinformatics Institute: The EBI houses Europe's primary databases for molecular sequence data, genomics and bioinformatics, and shares data daily with similar institutions in the United States and Japan.

Minimal Information Standards for life science research specify minimal metadata requirements for certain types of research data, are integrated by the MIBBI Project (Minimum Information for Biological and Biomedical Investigations), and are described in [1].

SWORD2: The SWORD2 repository submission protocol is a standard protocol for on-line submissions to text or data repositories.

Reference

[1] Taylor *et al*. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology* **26** (8): 889-896. doi:10.1038/nbt0808-889.

Twenty Questions for Research Data Management was created by David Shotton, University of Oxford. The original of this document is available from http://datamanagementplanning.wordpress.com/2012/03/07/twenty-questions-for-research-data-management/.