Postgraduate Data Management Plan: Guidance for Reviewers

Introduction

The purpose of a Data Management Plan

The University's Research Data Policy requires that Data Management Plans should be written for all new research projects.[1] There are many reasons why this requirement extends to postgraduate projects, as well as to externally funded projects:

1. Many funders ask for Data Management Plans either as part of any bids for funding or as part of setting up a research project. Data Management Plans are also used to establish data management procedures in industry. Writing a plan at postgraduate level therefore helps students acquire skills they will need later in their career.
2. Writing a plan prompts students to discover the services that the University offers to assist them in research data management, as well as appropriate services offered externally.
3. Writing a plan exposes students to best practice in research data management, helping them to avoid data breaches and other incidents that may occur through ignorance of relevant requirements and risks.
4. Reflecting the mood of the global academic community, it is also University policy that research data of value – whether as evidence supporting published findings or as a resource supporting future research – should be retained and shared as openly as possible. By planning ahead, researchers can broaden those possibilities for sharing, enabling additional value to be extracted from data that would otherwise have to be kept strictly private or destroyed.
5. The assessment process for Data Management Plans not only broadens the basis on which students' understanding of the research process may be judged, but also allows University staff to correct any misunderstandings that may lead to problems later in the project.
6. The contents of Data Management Plans may also be significant for how the University develops its research data management support infrastructure.

[1] University of Bath Research Data Policy: http://www.bath.ac.uk/research/data/policy/research-data-policy.html

Null Data Management Plans

If a project will not produce/use any data, nor develop any software – as may be the case in, say, pure mathematics – the candidate should provide a statement to that effect rather than attempt to fill out a full Data Management Plan template. Such statements need only be assessed on their veracity, not their quality!

Assessing a Data Management Plan

When assessing a Data Management Plan, you should look for the following qualities:

- The plan should be realistic and appropriate to nature of the research project.
- The plan should demonstrate awareness of all relevant requirements, and indicate how the project will comply with those requirements.
- The plan should identify any issues that restrict opportunities for data sharing. For any issues thus identified, the plan should demonstrate best efforts to minimise the issue and maximise data sharing.
- Any extraordinary support requirements identified in the plan (e.g. previously unplanned time in facilities) should be clearly specified and justifiable.
- If any aspect of the plan is incomplete, but still applicable, the gap should be justifiable and the plan should indicate how this will be rectified in future iterations. There is an expectation that the plan will evolve over time to reflect how the project is progressing.

The numbered sections below correspond to the sections in the University's postgraduate research Data Management Plan template. Each section contains suggestions for assessing whether a response is poor, satisfactory or excellent.

1 Overview

This section is largely factual. The project description should be coherent with the rest of the confirmation report; excellent answers should highlight the role that data will play in the project.

2 Compliance

2.1 With what legislative, contractual and policy requirements must the project comply?

The purpose of this question is to check whether the candidate has identified, read and understood the legislation, contracts and policies that require them to manage their data in a particular way. Only those documents and requirements that relate to research data management should be mentioned.

At the very least, the plan should mention the University's Research Data Policy.[2] Other sources of requirements may include the following:

- Code of Good Practice in Research Integrity (general)
- Data Protection Act (personal data)
- University of Bath Information Classification Framework (sensitive data of any kind)
- University of Bath IT Security Policy (collecting or using sensitive data off-campus)
- Agreements with organisational participants
- Requirements inherited from a larger project of which this is a part, e.g. collaboration agreements, funder data policy
- Other funding or studentship agreements
- Licences to use third-party data sources, if known

A satisfactory answer should identify all legislation, contracts and policies known to be relevant at the time of writing, but need not be exhaustive in the data management requirements it identifies. If the project is only affected by University policies, the candidate should mention at least three key requirements. The higher priority requirements are usually as follows:

- matters relating to informed consent;
- matters relating to data security;
- matters relating to retention of data (which should be retained, for how long, when deposit in an archive should occur);
- matters relating to data sharing (openly or with restrictions).

An excellent answer should provide a clear and concise summary of all relevant data management requirements.

3 Gathering data

3.1 What data will the project require?

From a satisfactory answer to this question, you should be able to deduce the following, even if the matters are not addressed explicitly:

- whether the data will be qualitative or quantitative;
- what the main classes of data will be;
- which formats the main classes of data will use;

[2] University of Bath Research Data Policy: http://www.bath.ac.uk/research/data/policy/research-data-policy.html

- approximately how much data will be needed.

You should also be satisfied that the data will be appropriate and sufficient to answer the research aims given in the project description and confirmation report. Regarding file formats, the answer should give a sufficient level of detail and show consideration for how readable files will be in future:

- For tabular data, CSV and TSV (comma- and tab-separated values) are excellent choices for long-term usability. MS Office Open XML Excel spreadsheets (.xlsx) and Open Document Spreadsheets (.ods) are good choices, and more functional. Older binary Excel spreadsheets (.xls) are less good but still acceptable. Merely specifying ‘Excel format’ is rather vague.
- For textual data, plain text and RTF (Rich Text Format) are both excellent choices for the long term. MS Office Open XML Word documents (.docx) and Open Document Text (.odt) are good choices, and more functional. Older binary Word documents (.doc) are less good but still acceptable. Merely specifying ‘Word format’ or ‘.doc format’ is rather vague (the .doc extension has also been used for non-Microsoft formats).
- PDF (Portable Document Format) is not usually acceptable as a data format, since many things can go wrong when trying to extract content from it.
- For other types of data (e.g. audiovisual), consider how widely used the format is, and whether more than one software product is able to read/write it.
- It is acceptable for candidates to use obscure or poorly supported formats for their data analysis so long as the files they plan to retain will (also) be saved in a long-lived/well-supported format.

Excellent answers will give additional detail:

- a thorough breakdown of the types of data files/records the project will use, including raw, processed, and non-digital data;
- file formats and quantity estimates for every type of data identified.

3.2 How will these data be gathered?

A satisfactory answer should provide a brief summary of the methodology used to gather the data. Usually a single sentence per class of data will suffice. For primary data collection, it should be clear which methodological step or equipment produces which class of data from the previous question; for example:

> ‘I will take high-resolution digital photographs of artefacts recovered in the field, and send some samples off for analysis.’

For secondary data use, it should be clear which data will come from which source, or which sources the candidate expects to check for relevant data.

A poor answer will either provide excessive methodological detail, or leave it unclear how some or all of the data will be gathered.

Excellent answers should demonstrate that the candidate has already checked sources of existing data, and will therefore not create new data when an existing resource could be reused. An answer may also be considered excellent if it references standard protocols or methodologies.

3.3 What original software, if any, will the project create?

If applicable, a good answer to this question should indicate

- what sort of software the project will create, e.g. script, model, module, library, prototype application, software product;
- what programming languages or frameworks will be used;
- whether the software will be uniquely fitted to this project or potentially usable in new contexts.

An excellent answer would also include details of dependency management and quality control, e.g. unit testing, continuous integration, user/usability testing.

4 Working with data

4.1 Where and how will the data be stored?

A satisfactory answer should select an appropriate means of storing research data, and include details of how those data will be backed up.

Generally speaking, the University's managed data storage should be used as the primary location of the data. Use of a supervisor's or project X Drive storage area is preferred to use of the H Drive. Valid reasons for considering alternative storage solutions include

- use of non-digital data;
- quantities of data in excess of available storage space;
- use of highly sensitive data (requiring non-networked storage);
- use of data held by a third party.

Any alternative solution must provide adequate data protection and security, and at least two members of the University must have access.

Unless backups are performed by Computing Services (or by whichever third party holds the data), the plan should specify the details of how they will be carried out:

- at what time interval will backups be taken?
- how many backup copies will be made?

- on what media will the backups be stored?
- where will the backups be kept?
- for how long will each backup be kept?

A backup plan for digital data can be considered excellent if it conforms to the 3-2-1 rule: at least 3 copies, on at least 2 different types or makes of medium, with at least 1 kept physically distant from the others. It should also indicate how the effectiveness of the backups will be tested (e.g. by performing test restorations), and the various timescales should be appropriate to the importance of the data and the capacity of the backup storage media.

Indicators of a poor plan include

- a failure to mention the selected storage location;
- a failure to mention how backups will be made;
- use of cloud storage for personal or commercially sensitive data (refer to the Data Protection Team and University IT Security officer if an argument is presented why an exception should be made);
- a mismatch between the capacity of the selected storage solution and the amount of data to be gathered.

Some discussion of data security may be expected here, but that issue should mainly be dealt with under the following question.

4.2 How will access be controlled?

A satisfactory answer should give details of a security regime that is appropriate to the sensitivity of the data. An excellent answer provides justification for the security regime with reference to, for example, the University's Information Handling Protocol,[3] the Data Protection Act,[4] or other requirements identified earlier in the plan.

For non-sensitive data, an appropriate security regime might involve restricting access to the project folder on the X Drive to the candidate and their supervisor, and only accessing the data through secure network connections on University devices. Files may be shared with collaborators through files.bath, and received from them through the other institution's equivalent service.

For sensitive data, the answer should indicate that higher levels of security have been considered:

- Identification keys and participant information should be stored securely and separately from the data.
- Portable devices (e.g. laptops) should only store sensitive data in an encrypted folder or partition, or the whole device should be encrypted.

[3] University of Bath guidance on information security: http://www.bath.ac.uk/university-secretary/guidance-policies/information_security.html

[4] University of Bath guidance on data protection: http://www.bath.ac.uk/data-protection/guidance/

- Data should be transferred to collaborators in encrypted form; for example, on an encrypted drive shipped through a mutually trusted courier, through files.bath in an encrypted ZIP file, or in an encrypted email or email attachment.

Indicators of a poor answer include disproportionate or entirely absent security controls; no consideration over who should have access to the data; and reliance on simple password protection of devices (or the accounts on those devices) to secure sensitive data.

4.3 How will the data be organised?

A satisfactory answer should outline a folder structure and file naming convention for organising the data files of the project, and indicate how version control will be handled.

An excellent answer should provide this information in enough detail to be implemented immediately.

The proposed folder structure should have either a functional or a semantic significance, or a mixture of both; in other words it should reflect how the files within will be managed (e.g. based on access permissions) or group files together that relate to the same aspect of the project (e.g. work package, task, sample, cohort, run of an experiment or simulation).

The proposed naming convention should name folders to reflect what the files/folders within have in common, in contrast to the files outside. File names should contain elements of key information about the contents of the file, such as the date of creation, the data subject or the type of file. These elements should appear in a consistent order and format.

Version control should be achieved either through the use of numerical version numbers appended to the file name (but before the extension) or through use of a dedicated version control system. The choice should be appropriate to the type of data/software and the methodology.

Regarding version control systems, an excellent answer should specify which system will be used (e.g. Git, Mercurial) and where the upstream repository will be hosted (e.g. University's private GitHub, GitLab, BitBucket).

4.4 What documentation will accompany the data?

A satisfactory answer should specify either or both of

- a 'readme' file (containing information about methodology, file organisation, formats, abbreviations/codes, and other contextual information);
- metadata conforming to a disciplinary norm or standard (e.g. DDI, MIAME).

An excellent answer should also talk about when and how this information will be captured and recorded; for example,

- automated embedding of metadata within files by instruments and data processing tools;
- writing comments alongside software source code at the same time;
- adding to the 'readme' file throughout data gathering and processing.

5 Archiving data

5.1 Which data should be retained long-term? Which will be deleted at the end of the project?

A satisfactory answer should indicate that data will be retained if they

- underlie published research findings;
- would be impossible or prohibitively expensive to reproduce;
- are otherwise of value as a resource for future research.

It should also nominate classes of data that may be deleted (e.g. temporary or auxiliary files, raw recordings).

An excellent answer will indicate or imply how these decisions were made, perhaps with reference to the requirements documents identified in the Compliance section of the plan.

A poor answer might specify, without adequate justification, that all data will be deleted at the end of the project. An inability to share data is not, in and of itself, adequate reason for none to be retained.

5.2 How will retained data be preserved? For how long?

A satisfactory answer should name a data archive or repository to which the retained data will be submitted. Note that the University's Research Data Archive is a suitable option, but the X Drive and Institutional Repository are not.

It is acceptable for the candidate to be tentative about this and supply a small number of alternatives. It may also be appropriate for the data to be split into subsets, with different data types submitted to different repositories. For small quantities of data, it is acceptable for them to be preserved as supplementary data to published papers, though this is not appropriate for high-value, complex or extensive datasets.

The length of time for preservation should match the requirements identified in the Compliance section. In the absence of any contrary requirement, this length of time should be at least ten years.

An excellent answer should include a few notes on how data will be prepared for archiving; for example, it might specify conversion of proprietary files to an open format, or packaging files in a community-supported format. The chosen repository should have been selected with care: if possible, it should specialise in the type of data that will be submitted there; otherwise the rationale for the choice should be given.

5.3 How will any original software be maintained after the project?

The quality of the answer should be judged on how well it matches the nature of the software, script or model being proposed.

Archiving a snapshot of the code alongside the data is a good option for scripts that are peculiar to the project, but not for software applications intended for many users. For the latter, a source code repository service would be more appropriate.

An excellent answer should address, if applicable, issues of community uptake and sustainability. For example, it might aim to have the software included in existing software distributions or catalogues (e.g. myExperiment for workflows, PyPI for Python modules), or explain how an open source community might be built around the software.

If the software is to be kept entirely private, this is usually indicative of a poor answer, though it might be justified if the research will be commercialised and the applicable policies permit this.

6 Sharing data

6.1 Will access be restricted to any retained data? Why, and how?

Indicators of a poor answer include the following:

- Access restrictions are planned (e.g. no sharing at all, access on request) with no justification, and no consideration of approaches that might permit wider sharing.
- Sensitive or confidential data will be shared openly.
- Data mined from, say, social media will be shared without respect for the terms of service of those sites.
- Data will be shared from a project website even though an archived copy is available.

A satisfactory answer should indicate that the data selected for retention will be shared as openly as possible. Where data will not be made available openly and immediately upon publication of the respective results, the reasons should be made clear, and the access restrictions should be proportionate to those reasons.

In general, it is acceptable to embargo data to allow certain other processes to complete (e.g. patent applications), but not in the speculative hope that new applications or submissions might be made in future.

An excellent answer will show evidence that, faced with potential barriers to data sharing, the candidate has planned steps to overcome those barriers (e.g. obtaining appropriate informed consent, negotiating terms with collaborators or rights holders, setting up a restricted access regime).

7 Implementation

7.1 How will this plan be kept up to date?

A poor answer might rely on the candidate making adjustments on an ad hoc basis as improvements occur to them.

A satisfactory answer should name another person with which the candidate will discuss the plan: normally this should be the candidate's supervisor. Some indication of when the meetings will occur should be given. If the candidate has already scheduled one or more meetings and quoted them in the plan, this points to an excellent answer.

7.2 What special resources will this plan require, if any?

If any previous answers imply that resources will be needed beyond normal University provision, these resources should be clearly specified here. They should nevertheless be reasonable and realistic in the context of the project.

7.3 What training or further information will you need, if any?

This question is primarily provided to assist in the candidate's professional development as a researcher, by identifying training opportunities that would help them improve their data management skills.

If the candidate identifies actions that would help them complete under-developed parts of the plan, consider whether they should have completed these actions prior to submitting their confirmation report. If it is reasonable that they did not, please consider the answer here as mitigation for the earlier gaps.