DEVELOPING AN INSTITUTIONAL RESEARCH DATA MANAGEMENT PLAN: GUIDELINES

Unisa 2014 LIS Research Symposium
Research Data Management, July 24 2014
Dr Heila Pienaar
Deputy Director: Innovation & Technology
University of Pretoria Library
Research data (management) cycle
What is a data management plan?
Benefits of a research data management (RDM) plan
The two best-known international RDM plans
Examples of university RDM plans
Guidelines
Living in an Ivory Basement
Stochastic thoughts on science, testing, and programming.

My Data Management Plan - a satire

Dear NSF,

I am happy to respond to your request for a 2-page Data Management Plan.

First of all, let me say how enthusiastic I am that you have embraced this new field of "large scale data analysis". Ever since I started working with large Avista data sets in 1993, then with large meteorological data sets in 1995, and then again with large sequence data sets in 1999, I have seen the need for a systematic plan to manage the data. It is nice to see NSF stepping up to the plate in such a timely manner, and I am happy to comply.

Now, as to my actual data management plan, here is how I plan to deal with research data in the future.

I will store all data on at least one, and possibly up to 50, hard drives in my lab. The directory structure will be custom, not self-explanatory, and in no way documented or described. Students working with the data will be encouraged to make their own copies and modify them as they please, in order to ensure that no one can ever figure out what the actual real raw data is.

Backups will rarely, if ever, be done.

When required to make the data available by my program manager, my collaborators, and ultimately by law, I will grudgingly do so by placing the raw data on an FTP site, named with UUIDs like 4e283d36-61c4-11df-9a26-63f1d420002d. I will under no circumstances make any attempt to provide analysis source code, documentation for formats, or any metadata with the raw data. When requested (and only when requested), I will provide an Excel spreadsheet linking the names to data sets with published results. This spreadsheet will likely be wrong -- but since no one will be able to analyze the data, that won't matter.

Did I mention the click-through license? "You are provided this data for the sole purpose of reproducing our published results. Any attempt to publish your own analyses of this data will be rejected, if necessary during the anonymous review process, by pointing out all of the data cleanup steps you forgot to do correctly in your analysis. (We don't remember all of them ourselves, but there sure were a lot!) Give up now."

Mon 17 May 2010
By C. Titus Brown
In science.
tags: science
‘I will store all data on at least one, and possibly up to 50, hard drives in my lab.

The directory structure will be custom, not self-explanatory, and in no way documented or described.

Students working with the data will be encouraged to make their own copies and modify them as they please, in order to ensure that no one can ever figure out what the actual real raw data is.

Backups will rarely, if ever, be done.’
Research Data Life Cycle

Creating data
- design research
- plan data management (formats, storage etc.)
- plan consent for sharing
- locate existing data
- collect data (experiment, observe, measure, simulate)
- capture and create metadata

Processing data
- enter data, digitise, transcribe, translate
- check, validate, clean data
- anonymise data where necessary
- describe data
- manage and store data

Analysing data
- interpret data
- derive data
- produce research outputs
- author publications
- prepare data for preservation

Preserving data
- migrate data to best format
- migrate data to suitable medium
- back-up and store data
- create metadata and documentation
- archive data

Giving access to data
- distribute data
- share data
- control access
- establish copyright
- promote data

Re-using data
- follow-up research
- new research
- undertake research reviews
- scrutinise findings
- teach and learn

Based on UK Data Archive documentation:
http://www.hsrc.ac.za/Page-156.phtml
http://www.data-archive.ac.uk/media/2894/managingsharing.pdf
http://www.data-archive.ac.uk/
A data management plan (DMP) is a document that describes how you will collect, organise, manage, store, secure, back-up, preserve, and where applicable, share your data.

A data management plan should be created at the start of a research project so that good practices are established early.

http://www.sheffield.ac.uk/library/rdm/dmp
Helping you to identify issues and strategies early in your research project
Ensuring that you have documented your compliance with institutional and funder policies and ethics approval requirements
Helping you think about data sharing and reuse opportunities of your research data
Ensuring that your research data remains useful and stored safely for future use

http://www.sheffield.ac.uk/library/rdm/dmp
THE TWO BEST-KNOWN INTERNATIONAL RDM PLANS

https://www.flickr.com/photos/rosefirerising/6776182890/

https://dmponline.dcc.ac.uk/
What data will you collect or create?
How will the data be collected or created?
What documentation and metadata will accompany the data?
How will you manage any ethical issues?
How will you manage copyright and Intellectual Property Rights (IPR) issues?
How will the data be stored and backed up during the research?
How will you manage access and security?
Which data should be retained, shared, and/or preserved?
What is the long-term preservation plan for the dataset?
How will you share the data?
Are any restrictions on data sharing required?
Who will be responsible for data management?
What resources will you require to deliver your plan?

Welcome.
DMPonline has been developed by the Digital Curation Centre to help you write data management plans.

Sign in

Sign up

New to DMPonline? Sign up today.
# DMPTool: Requirements

Use the A-Z links below to narrow down the list by institution or use the search box to search for specific DMP Templates. Sample plans are not necessarily from the funder.

<table>
<thead>
<tr>
<th>Template</th>
<th>Funder</th>
<th>Funder Links</th>
<th>Sample Plans (if available)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gordon and Betty Moore Foundation</td>
<td>Gordon and Betty Moore Foundation</td>
<td>Guidelines</td>
<td>GBMF: Sample Plan #1, GBMF: Sample Plan #2, GBMF: Sample Plan #3</td>
</tr>
<tr>
<td>Gulf of Mexico Research Initiative</td>
<td>Gulf of Mexico Research Initiative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institute of Education Sciences (US Dept of Education)</td>
<td>Institute of Education Sciences (US Dept of Education)</td>
<td>IES Data Sharing Implementation Guide</td>
<td></td>
</tr>
<tr>
<td>Institute of Museum and Library Services</td>
<td>Institute of Museum and Library Services</td>
<td>Guidance</td>
<td></td>
</tr>
<tr>
<td>Joint Fire Science Program</td>
<td>Joint Fire Science Program</td>
<td></td>
<td></td>
</tr>
<tr>
<td>National Institutes of Health</td>
<td>National Institutes of Health</td>
<td>Guidance</td>
<td>NIH: Sample Plans</td>
</tr>
<tr>
<td>NEH-ODH</td>
<td>University of Connecticut</td>
<td>NEH Digital Humanities Start-Up Grants</td>
<td></td>
</tr>
<tr>
<td>NEH-ODH: Office of Digital Humanities</td>
<td>National Endowment for the Humanities</td>
<td>Guidelines</td>
<td>NEH-ODH Sample</td>
</tr>
<tr>
<td>NSF-AGS: Atmospheric and Geospace Sciences</td>
<td>National Science Foundation</td>
<td>AGS Advice</td>
<td></td>
</tr>
</tbody>
</table>

https://dmptool.org/guidance
This is an example Data Management Plan created by participants of a DataONE best Practices Workshop.

Data Management Plan

I. Products of Research

Every two days, we will subsample *E. affinis* populations growing at our treatment conditions. We will use a microscope to identify the stage and sex of the subsampled individuals. We will document the information first in a laboratory notebook, then copy the data into an Excel spreadsheet. For quality control, values will be entered separately by two different people to ensure accuracy. The Excel spreadsheet will be saved as a comma-separated value (.csv) file daily and backed up to a server. After all data are collected, the Excel spreadsheet will be saved as a .csv file and imported into the program R for statistical analysis. Strasser will be responsible for all data management during and after data collection.

II. Data Storage and Preservation

Our short-term data storage plan, which will be used during the experiment, will be to save copies of 1) the .txt metadata file and 2) the Excel spreadsheet as .csv files to an external drive, and to take the external drive off site nightly. We will use the Subversion version control system to update our data and metadata files daily on the University of Alberta Mathematics Department server. We will also have the laboratory notebook as a hard copy backup.
PURDUE UNIVERSITY, USA

What is PURR?
The Purdue University Research Repository (PURR) provides an online, collaborative working space and data-sharing platform to support the data management needs of Purdue researchers and their collaborators.

Start Your Research Project
- Create a Data Management Plan
  Learn about the detailed requirements for your data management plan (DMP). Funding agency requirements are very specific and our DMP resources can help you to clear up any confusion. Get Started
- Upload Research Data to Your Project
  Create a project to upload and share your data with collaborators using our step-by-step form to guide you through the process. Invite collaborators from other institutions to join your project. Create a Project
- Publish your Dataset
  Package, describe, and publish your dataset with a Datadice DOI. Publishing will ensure your dataset is citable, reusable, and archived for the long-term. See Published Datasets

Featured Dataset
- Maize grain yield record for the WQPS (1995-2012)
  By S.M. Brooker, N. Da Armond, R.P. Turco, J.I. Volkenc
  Purdue University
  The data included here are for the WQPS maize grain yields only, see companion publications for other data including yields of rotter crops, maize sugar biomass, nitrate loss in drainage water and other environmental impact data.

Do you have a question?
- Chat with Us
  Send us your email question:
  Enter your email address
  Your question/answer
  Send
Data Management Plan (DMP)

Data management planning (DMP) is an essential element of research, and seen as a necessity by many funders. Because “sharing” or “publishing” data may not have always been an integral part of research, PURR offers several levels of assistance in data management planning. For those who have a data management plan and intend to use PURR, there is a guide to get you going. For those who are new to preparing a data management plan, here are a couple of options:

- Review the Data Management Plan overview
- Go directly to the Data Management Planning Self-Assessment Tool
- Use the DMP Tool to create ready-to-use data management plans for specific funding agencies
- Contact a Libraries faculty member for consultation on discipline specific questions using the “Ask a Librarian” form below.

NOTE: If you need more information about PURR as a repository solution for data management planning, please see the links under Contact Us.

Overview:
This is where you’ll find links to more details about writing your Data Management Plan.
- Overview
- DMP Examples
- DMP Self-Assessment Tool
- DMP Workshop

Guides and Tutorials:
In this section, we will eventually provide tutorials to help you get more acquainted with PURR, and how to use groups, projects and much more for collaboration and publication.
- Boilerplate Text for DMPs planning to use PURR
- Guidance on developing a DMP for NSF proposals

Contact Us:
If you can’t seem to find what you need, or you’ve noticed a bug on our site, please contact us and let us know.
- For questions on getting started with a grant proposal: Contact SPS
- For information using PURR and data management, contact Courtney Matthews, 496-2770, courtneyearlmatthews@purdue.edu, PURR Digital Data Repository Specialist
- For technical support or to report a bug: open a trouble ticket
Text for Your Proposal

The following boilerplate text is a great way to describe PURR in your Data Management Plan:

Support for data management for this project throughout its lifecycle will occur using the Purdue University Research Repository (PURR), Purdue’s institutional data repository. PURR utilizes HUBzero®, a web-mediated software platform designed for scientific collaboration and sharing of scientific data that was developed with support from the National Science Foundation and Purdue University. PURR provides workflows and tools for ingestion, identification and dissemination of data as well as services to ensure data security, fidelity, backup, and mirroring. Purdue Libraries will consult with investigators to facilitate selection and ingestion of data with the application of appropriate descriptive metadata and data standards as well as to provide appraisal of data for long-term preservation and stewardship. PURR is working towards the ISO 16363 process to become a certified Trusted Digital Repository. PURR comes with a set of default policies and functionality that addresses privacy and confidentiality, intellectual property and copyright, and access and sharing of data. Datasets published using PURR will be assigned Digital Objects Identifiers (DOIs) and will be exposed to the web using open standards to maximize discoverability and scholarly reuse of data. An allocation of resources from PURR has been reserved for this project and will be appropriated upon its award.

Suggested citation for proposal bibliography:

Research Data Planning Checklist (Higher Degree by Research Student)

Name
Faculty / division / unit
Contact details
Title / description of the Higher Degree by Research (HDR) project

A. OWNERSHIP, COPYRIGHT, INTELLECTUAL PROPERTY (IP)

Copyright protection
1. ☐ The data is protected by copyright. 
   This will apply to most research data.
2. ☐ The data will be collected, created or compiled
   □ in Australia - Australian copyright applies.
   □ outside of Australia.

Ownership of copyright and IP
3. The copyright and other IP in the data is owned by:
   ☐ the Higher Degree Research Student
   Research by Monash HDR student in the normal course of study, which does not fall into any of the other categories prescribed under the Statutes and Regulations.
   ☐ Monash University
   I have assigned IP to the University because it falls into one of the categories prescribed under the Statutes and Regulations.
   ☐ Monash University (joint ownership)
   Research project is collaborative, copyright and IP ownership is shared between the member institutions.

KEY DOCUMENTS ON THIS TOPIC
Research data management guidelines: ownership, copyright and IP
Intellectual Property Framework
Statute 11.2 IP and Copyright and IP Regulations
Explanatory Memorandum for IP Statute and Regulations
Copyright at Monash website
Practical Data Management: A Legal and Policy Guide (national guide)

Consult the Copyright Advisers or University Solicitors.

MIGR Handbook Chapter 6: Intellectual Property - Assignment and Licensing
Provide a copy of MIGR IP and Assignment Forms to help clarify ownership of the data.
Data sets

The HSRC Research Data Service provides a digital repository facility for the HSRC’s research data in support of evidence based human and social development in South Africa and the broader region. Access to data is dependent on ethical requirements for protecting research participants, as well as on legal agreements with the owners, funders or in the case of data owned by the HSRC, the requirements of the depositors of the data. We facilitate data use by preparing comprehensive metadata and disseminating data and related documents to appropriate target audiences. Data sharing is subject to an End User License agreement. Data sets from the following projects have been curated and are available for use:

- **A comparative analysis of self-Identifying Gender Non-Conforming youth against the realities of Gender Non-Conforming youth In Cape Town, South Africa (NCG)**
  The study, funded through the International Congress of Psychology, provides a comparative analysis of self-identifying gender non-conforming youth residing on the urban peripheries of Cape Town, against the realities of gender non-conforming youth in more affluent, resourced communities in Cape Town, South Africa.
  The study explored the dynamics that make for different possibilities to express gender non-conformity among youth, as well as the opportunities available for these young people to recreate their identities in ways that make sense to themselves, and allow for a more fulfilling sense of self in contemporary South Africa.

- **A longitudinal view of unemployment in South Africa (UNEMPL)**
  Global unemployment has risen in the past few years. Spatial data are required to address the problem effectively. South African unemployment literature focused mostly at a national level of spatial analysis. Some literature refer to spatial aspects that affect unemployment trends, but does not assign a location, e.g. a suburb or municipality. The research was conducted to obtain an understanding of geographical unemployment changes over time. The objectives were to provide an overview of South African unemployment spatial data sets, assess the usefulness of unemployment data for the geographical analysis of unemployment and recommend data sets that should be used for the spatial analysis of unemployment.

- **Agrarian Reform and rural poverty reduction (AR)**
  Phase one of the project 'Agrarian Reform and Poverty Reduction' produced a comprehensive status report on this topic, including a review of the conceptual approaches and methodological best practices to guide further empirical research. This desktop synthesis of the literature identified prominent mechanisms or pathways through which reforms are likely to interact with poverty and to define meaningful indicators of human well-being as measures of impacts.
Research Data Management Plan

Please supply the following information about how the sharing of the research data related to the project will be done. Provide the completed form as part of the submission of the ethic application to the HSRC Research Ethics Committee.

Note: For clarity about the required information, please refer to the footnotes provided.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Study / collection title</td>
</tr>
<tr>
<td>2</td>
<td>Person who will be responsible for the data</td>
</tr>
<tr>
<td>3</td>
<td>Research data files</td>
</tr>
</tbody>
</table>
| 3.1 | What will be deposited for preservation?  
Select one or more of the data types indicated below. | Give a short description of the contents of the data files. | Indicate the electronic file format, e.g. .sav |
| Quantitative tabular data | | |
| GIS and CAD data | | |
| Qualitative data | | |
| Digital image data | | |
| Digital audio data | | |
| Digital video data | | |
| Other if other, please describe. | | |
GUIDELINES

- Do not try to create a data management plan from scratch: many good examples are available.
- Use current examples to co-create with role players a data management plan framework with minimum requirements for your institution.
- Each research group can adapt the plan as long as the minimum requirements are included.
- Plan should be short, easy and available online.
- Must link with or include ethical requirements.
- Must form an integral part of research planning & process.