

Productivity through data management

a.k.a. Writing an effective data management plan

Anna J Dabrowski

Introduction

Focus on planning for effective data management in advance of research projects.

Goal: Gain an overview of information to include in a plan.

Slides: <http://hdl.handle.net/1969.1/166284>

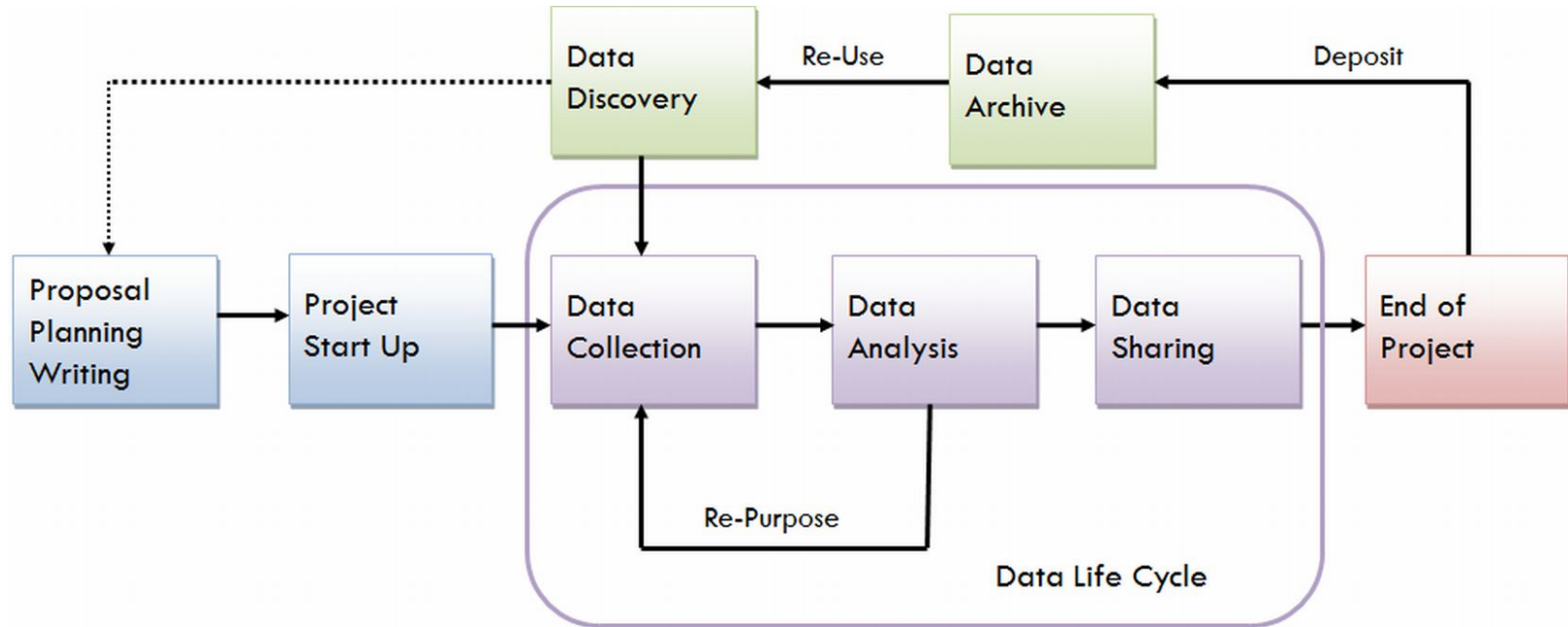
Defining “data”

“Recorded factual material commonly accepted in the scientific community as necessary to document and support research findings.” ([National Institutes of Health](#))

“...determined by the community of interest through the process of peer review and program management.” ([National Science Foundation](#))

“...materials generated or collected during the course of conducting research.” ([National Endowment for the Humanities](#))

The research data lifecycle



Time is the enemy

- Accumulating large quantities of disorganized files.
- Lack of information describing content in digital files.
- Changes to hardware, software, and file formats in common use.
- File corruption.
- Failure of storage media.
- Data leaving with collaborators.

Research data management (RDM)

The practices in place for

- organizing,
 - documenting,
 - storing,
 - sharing,
 - and preserving
- data collected during a research project.



Data management plans (DMPs)

Structured documents created in advance of collecting data that:

- Help you think about data over time and in context;
- Provide a framework for documenting your RDM practices;
- Are required by many grant agencies that fund research.

Scope of DMPs

- A research group or collaboration;
- An individual researcher;
- A specific research project.

Two perspectives

Concerned with **efficient** practices that help to maintain **access** to usable data.

Concerned with ensuring results can be **validated** and data **reused** in the future.



For researchers

The DMP is a “living” document adjusted over time. Allows you to:

- Identify relevant practices that improve research efficacy.
- Plan in advance to reduce later costs in time and effort.
- Easily share an overview with collaborators and stakeholders.
- Make incremental changes and speed up the grant application process.

For funders


The DMP is written as part of a grant proposal and is limited in scope:

- Describes the expected practices for a particular research project.
- Approximately 2–3 pages of text written according to a funder's guidelines.

Stebbins, M. 2013. Expanding Public Access to the Results of Federally Funded Research. <https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research>

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren 
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

I. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and

- b) Ensure that all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long-term preservation and access cannot be justified;

growth and job creation.

The Administration also recognizes that publishers provide valuable services, including the coordination of peer review, that are essential for ensuring the high quality and integrity of many scholarly publications. It is critical that these services continue to be made available. It is also important that Federal policy not adversely affect opportunities for researchers who are not funded by the Federal Government to disseminate any analysis or results of their research.

To achieve the Administration's commitment to increase access to federally funded published research and digital scientific data, Federal agencies investing in research and development must have clear and coordinated policies for increasing such access.

What do funders want?

Depends on the funder.

May have specific data sharing expectations or requirements.



The Scholarly Publishing and Academic Resources Coalition (SPARC) tracks article and data sharing requirements for federal agencies.

What do funders want?

Findable

Accessible

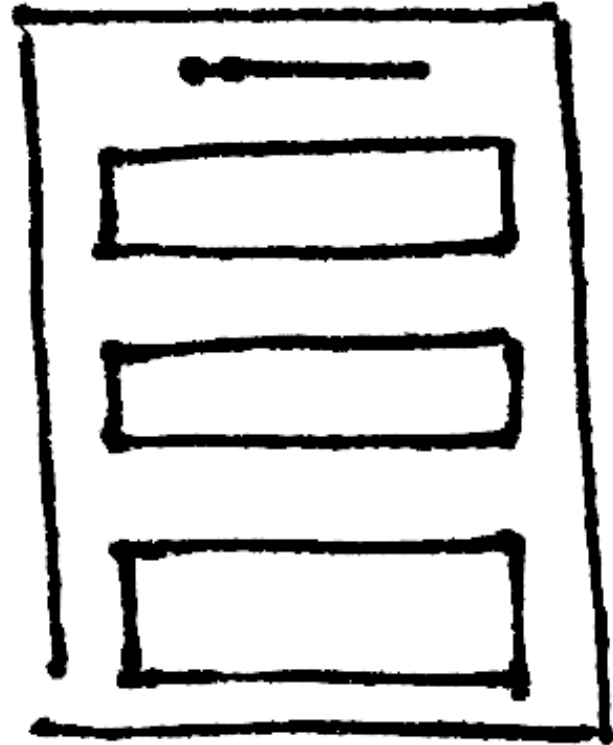
Interoperable

Reusable

Wilkinson, M.D. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. <https://www.nature.com/articles/sdata201618>

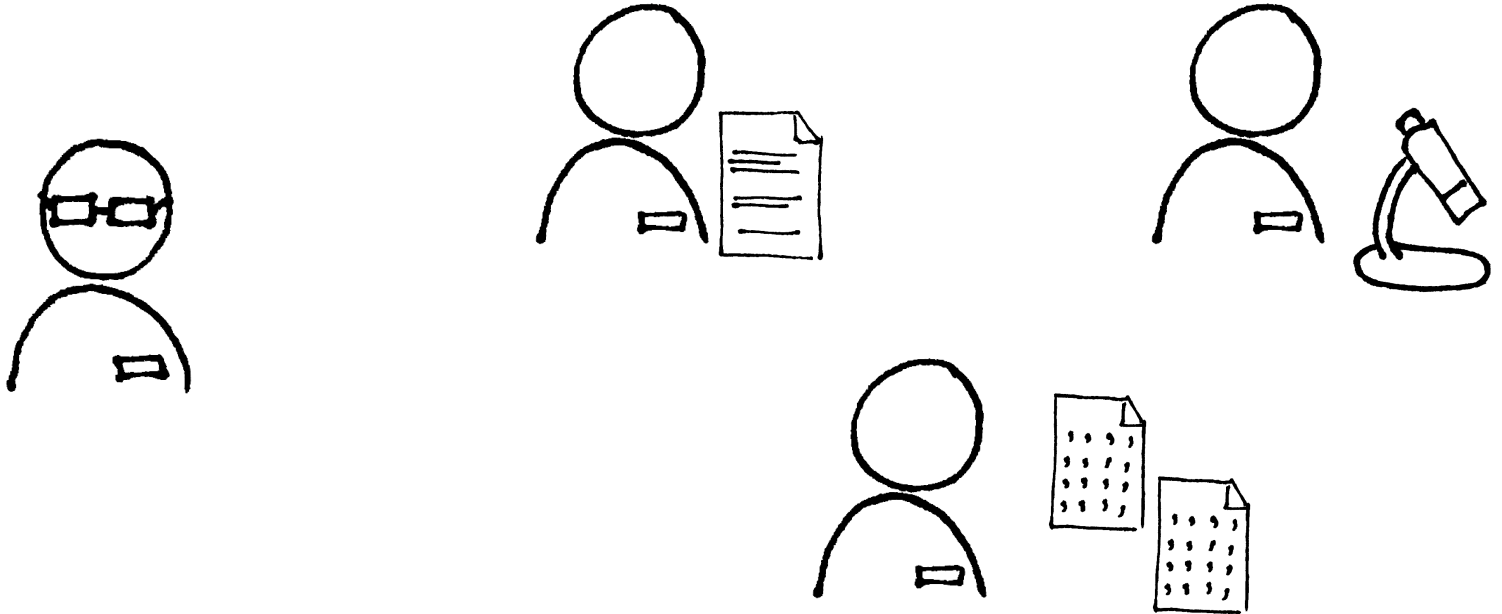
Components of a plan

1. Roles and responsibilities
2. Expected data
3. Data formats and standards
4. Data storage
5. Data sharing, access, and rights
6. Archiving and preservation



1. Roles and responsibilities

Outline the roles and responsibilities for all RDM activities.



2. Expected data

Describe the information to be gathered, including the nature and scale of data that will be generated or collected.

Existing data: Will existing data be reused? How might new data complement existing data?

Data volume: At the scale of MB, GB, TB. Are there implications for storage, sharing, and transfer?

Data types: Which data are of long-term value and should be shared and/or preserved?

2. Examples of data types

Text: Field or laboratory notes, survey responses.

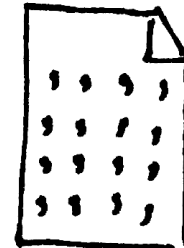
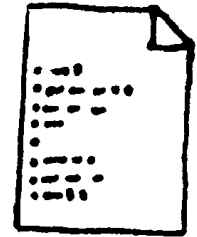
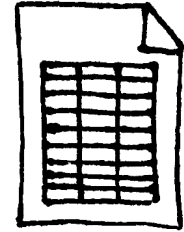
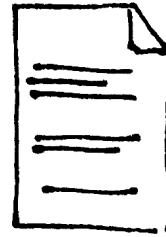
Numeric: Tables, counts, measurements.

Audiovisual: Images, sound recordings, video.

Code and models: Python, MATLAB, R, etc.

Discipline-specific: FITS in astronomy, CIF in chemistry.

Instrument-specific: Equipment outputs.



3. Data formats and standards

Explain **digital file formats** chosen for all data types.

&

Explain the **documentation** that will be created for understanding the data and enabling reuse.

&

Mention **metadata standards** that will be used to enable finding data that are shared.

3. Recommended file formats

Focus on interoperability and long-term usability. Features of formats that last:

- In common usage by the research community.
- Non-proprietary.
- Documented standards.
- Use standard character encodings (i.e. ASCII, UTF-8).
- Uncompressed (space permitting).

3. Recommended file formats by extension

Content type	File formats
Text	PDF/A, HTML, XML, TXT
Tabular data (spreadsheets and databases)	XML, CSV
Numbers and statistics	TXT, DTA, POR, SAS, SAV
Geospatial	SHP, DBF, GeoTIFF, NetCDF
Audio	WAVE, AIFF, MP3, MXF
Images	TIFF, JPG, JP2, PNG, GIF, BMP
Moving Images	MOV, MPEG-4, AVI, MXF
Web Archive	WARC
Containers	TAR, GZIP, ZIP

3. Documentation

Consider how you will capture this information and where it will be recorded:

- Data collection methodology;
- Analytical and procedural information;
- Definitions of variables and units of measurement;
- Assumptions made and quality indicators;
- Software used to collect and/or process the data.

3. Metadata standards

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.

Metadata is recorded according to one of many standards.

<http://rd-alliance.github.io/metadata-directory/standards/>

The Research Data Alliance (RDA) maintains a discipline-based metadata standards directory.

3. Distinguishing documentation and metadata

Documentation

- Can be informal.
- Often created while working on a project.
- May cover many levels (project, datasets, data files, variables and values).
- May provide general context.

Metadata

- Formally describes a particular object (can be a data file or dataset).
- Often created upon “publication” and linked to an object.
- Formatted into a record and structured according to standards.
- May derive from the documentation.

4. Data storage

Describe where the data will be stored and backed up during the course of research activities.

- Consider how **data security** will be managed, particularly if you are working with sensitive data.
- Discuss your **backup** strategy.
- Note the main risks and how these will be managed for **compliance**.

5. Data sharing, access, and rights

Explain who will have access to which data, and when. Note ethical, privacy, or legal issues.

- Who will own copyright and intellectual property rights?
- Which data will be shared with others?
- When will you make the data available?
- How will data be made available to others?
- How will the data will be licensed for reuse?

5. Common data sharing approaches

Informal sharing: provide access to, or send research data, upon request.

Supplemental information: provide research data in support of published articles.

Data repository: deposit research data in an openly accessible repository.

5. Data repositories

Repositories provide a means of openly sharing (publishing) data online and archiving digital files.

May also be called *data centers*, *data archives*, or *scientific databases*.

5. Finding data repositories

- Registry of Research Data Repositories: re3data.org
- FAIRSharing: <https://fairsharing.org/databases/>
- Nature Scientific Data recommended repositories: <https://www.nature.com/sdata/policies/repositories>
- Texas Data Repository: <https://dataverse.tdl.org/dataverse/tamu>

6. Archiving and preservation

Outline the plans for data archiving and preservation.

How long will the data be retained and where will they be archived?

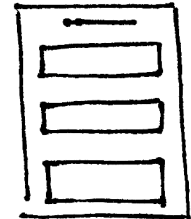
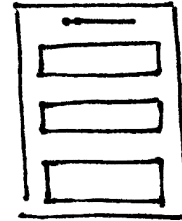
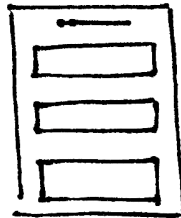
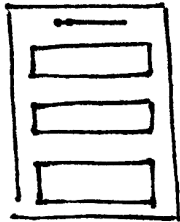
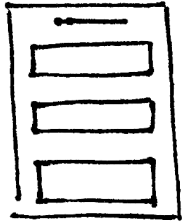
Texas A&M University. SAP 15.99.03.M1.03: The Responsible Stewardship of Research Data. <http://rules-saps.tamu.edu/PDFs/15.99.03.M1.03.pdf>

Questions to answer in a DMP

- What data will you collect or create, and how?
- What types of documentation and metadata will you produce to support the data?
- How and where will you store data files?
- How will ethical and legal issues be handled?
- Which data will be retained and shared?
- How do you intend to archive and share your data, and why those options?
- Who will act as the responsible steward for the data?
- What resources, including monetary, will be required?

Example DMPs

- Briefly read over the DMPs in the packet.
- What strikes you about the plans?
- Choose one to focus on for the next exercise.



Review a DMP

Use the *Data Management Planning (DMP) themes* to review how well your chosen example answers relevant questions.

- Is it thorough?
- What does it omit?
- Once the project is over, how likely is it that data will be **F**indable, **A**ccessible, **I**nteroperable, **R**eusable?

Writing a plan for a funder



<https://dmptool.org>

The DMPTool is an online tool that walks you through writing a DMP for specific funding agencies.

It will ask you to answer the questions that a particular funder cares about.

Conclusion

DMPs serve to:

- Aid thinking through how you will treat the data in your project.
- Show others (funders) that you've thought about it.

Data Management Plans are also living documents that change as you work with data over time.

Additional resources

- Carroll, M. W. 2015 Sharing Research Data and Intellectual Property Law: A Primer. <http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002235>
- Dabrowski, A.J. Research Data Management: Data Management Planning. <https://tamu.libguides.com/research-data-management/dmps>
- DataONE. Best Practices. <https://www.dataone.org/best-practices>
- DMPTool. Data management general guidance. https://dmptool.org/general_guidance
- ICPSR. Framework for creating a data management plan. <http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/framework.html>

Contact information



Anna J Dabrowski
Data Management Librarian
ajdabrowski@tamu.edu
(979) 845-8847

NIH Data Sharing Policy and Implementation Guidance

From: https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#ex

Example 1

The proposed research will involve a small sample (less than 20 subjects) recruited from clinical facilities in the New York City area with Williams syndrome. This rare craniofacial disorder is associated with distinguishing facial features, as well as mental retardation. Even with the removal of all identifiers, we believe that it would be difficult if not impossible to protect the identities of subjects given the physical characteristics of subjects, the type of clinical data (including imaging) that we will be collecting, and the relatively restricted area from which we are recruiting subjects. Therefore, we are not planning to share the data.

NIH Data Sharing Policy and Implementation Guidance

Example 2

The proposed research will include data from approximately 500 subjects being screened for three bacterial sexually transmitted diseases (STDs) at an inner city STD clinic. The final dataset will include self-reported demographic and behavioral data from interviews with the subjects and laboratory data from urine specimens provided. Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics. Thus, we will make the data and associated documentation available to users only under a data-sharing agreement that provides for: (1) a commitment to using the data only for research purposes and not to identify any individual participant; (2) a commitment to securing the data using appropriate computer technology; and (3) a commitment to destroying or returning the data after analyses are completed.

NIH Data Sharing Policy and Implementation Guidance

Example 3

This application requests support to collect public-use data from a survey of more than 22,000 Americans over the age of 50 every 2 years. Data products from this study will be made available without cost to researchers and analysts.

User registration is required in order to access or download files. As part of the registration process, users must agree to the conditions of use governing access to the public release data, including restrictions against attempting to identify study participants, destruction of the data after analyses are completed, reporting responsibilities, restrictions on redistribution of the data to third parties, and proper acknowledgement of the data resource. Registered users will receive user support, as well as information related to errors in the data, future releases, workshops, and publication lists. The information provided to users will not be used for commercial purposes, and will not be redistributed to third parties

Project documentation

- Rationale and context for data collection.
- Research questions, goals, and hypotheses.
- Data sources, collection methodology, protocols.
- Data validation and quality assurance actions.
- Transformation of raw or derived data for integration or analysis.
- Data confidentiality, access, and use conditions.

Dataset documentation

- Variable names and descriptions.
- Codes and classification schemes.
- Algorithms used to transform data.
- Structure and organization of files.
- Relationship among data files or tables in a database schema.
- Version information.
- File formats and software used.

How to document

A few documentation tools:

- Laboratory and field notebooks
- README files
- Codebooks

Example: README template

Cornell University.

<https://cornell.app.box.com/v/ReadmeTemplate>

Codebooks and data dictionaries

Document at the level of variables and data within the dataset.

Stand-alone or part of other files.

Primarily aimed at an external audience and your future self.

Good practices and tips

- Variable name.
- Variable meaning.
- Variable format and how the variable was recorded.
- Units of measurement for scale variables.
- Numeric codes for categorical variables, and what they represent.
- Known issues and relationships.