Data Management Plan Template: **Open Science Workflows**

Abstract

This template guides the writing of a research data management plan (DMP) with an open science/open scholarship workflow, which uses mixed social sciences research methods, producing quantitative as well as qualitative datasets. The questions emphasize data sharing and reuse throughout the project, not only at the final stage of publication. This DMP template will be most useful to researchers who are working in a multi-institutional partnership and who have already completed a funding application and an ethics review protocol. The DMP is a living document: don't forget to revisit your DMP throughout the research project to update or review your responses.

Not all of these questions will apply to all research projects. We encourage you to respond to as many as possible but ultimately, you and your team have to decide which questions and answers apply to your workflow.

This template is also available on the Meaningful Data Counts Zenodo community space: https://doi.org/10.5281/zenodo.4092122.

Administrative Details

Template Author(s): Erica Morissette, Lina Harper, University of Ottawa; Isabella Peters, ZBW Leibniz Information Centre for Economics, CAU Kiel University; Felicity Tayler, Stefanie Haustein, University of Ottawa

Published: April 19, 2021

DOI: 10.5281/zenodo.4701021

Contact: Portage Network - portage@engagedri.ca, portagenetwork.ca

Version:

Version	Date	Changes
1.0	2021-04-09	Formatted for inaugural publication.
1.1	2021-04-19	Attribution adjusted.

Responsibilities and Resources

Who will be responsible for data management? Will the Principal Investigator (PI) hold all responsibility during and beyond the project, or will this be divided among a team or partner organizations?

Assign responsibilities: Once completed, your data management plan will outline important data activities in your project. Identify who will be responsible -- individuals or organizations -- for carrying out these parts of your data management plan. This could also include the time frame associated with these staff responsibilities and any training needed to prepare staff for these duties.

In the event that the PI leaves the project, who will replace them? Who will take temporary responsibility until a new PI takes over?

Succession planning: The PI is usually in charge of maintaining data accessibility standards for the team. Consider who will field questions about accessing information or granting access to the data in the event the PI leaves the project. Usually the Co-PI takes over the responsibilities.

Indicate a succession strategy for these data in the event that one or more people responsible for the data leaves (e.g. a graduate student leaving after graduation). Describe the process to be followed in the event that the Principal Investigator leaves the project. In some instances, a co-investigator or the department or division overseeing this research will assume responsibility.

List all expected resources for data management required to complete your project. What hardware, software and human resources will you need? What is your estimated budget?

Budgeting: Common purchases are hard drives, cloud storage or software access. TU Delft's Data Management Costing Tool is helpful to determine the share of human labor (FTE) that should be allocated for research data management.

Data Collection

What types of data will you collect, create, link to, acquire and/or record?

Data types: Your research data may include digital resources, software code, audio files, image files, video, numeric, text, tabular data, modeling data, spatial data, instrumentation data.

Answer the following regarding file formats:

A. **What file formats do you expect to collect (e.g. .doc, .csv, .jpg, .mov)?**
B. **Are these file formats easy to share with other researchers from different disciplines?**
C. **In the event that one of your chosen file formats becomes obsolete (or is no longer supported) how will you ensure access to the research data?**
D. **Does your data need to be copied to a new media or cloud platform, or converted to a different file format when you store or publish your datasets?**

Proprietary file formats requiring specialized software or hardware to use are not recommended, but may be necessary for certain data collection or analysis methods. Using open file formats or industry-standard formats (e.g. those widely used by a given community) is preferred whenever possible. Read more about file formats: Library and Archives Canada, UBC Library or UK Data Archive.

Answer the following regarding naming conventions:

A. **How will you structure, name and version-control your files to help someone outside your research team understand how your data are organized?**
B. **Describe your ideal workflow for file sharing between research team members step-by-step.**
C. **What tools or strategies will you use to document your workflow as it evolves during the course of the project?**

File naming and versioning: It is important to keep track of different copies or versions of files, files held in different formats or locations, and information cross-referenced between files. This process is called 'version control'. Logical file structures, informative naming conventions, and clear indications of file versions, all contribute to better use of your data during and after your research project. These practices will help ensure that you and your research team are using the appropriate version of your data, and minimize confusion regarding copies on different computers and/or on different media. Read more about file naming and version control from UBC Library or the UK Data Service.

Remember that this workflow can be adapted, and the DMP updated throughout the project.

Using a file naming convention worksheet can be very useful. Make sure the convention only uses alphanumeric characters, dashes and underscores. In general, file names should be 32 characters or less and contain the date and the version number (e.g.: “P1-MUS023_2020-02-29_051_raw.tif” and “P1-MUS023_2020-02-29_051_clean_v1.tif”).

Document workflows: Have you thought about how you will capture, save and share your workflow and project milestones with your team? You can create an onboarding document to ensure that all team members adopt the same workflows or use workflow management tools like OSF or GitHub.

Documentation and Metadata

What support material and documentation (e.g. ReadMe) will your team members and future researchers need in order to navigate and reuse your data without ambiguity?

Data documentation: It is strongly encouraged to include a ReadMe file with all datasets (or similar) to assist in understanding data collection and processing methods, naming conventions and file structure.

Typically, good data documentation includes information about the study, data-level descriptions, and any other contextual information required to make the data usable by other researchers. Other elements you should document, as applicable, include: research methodology used, variable definitions, vocabularies, classification systems, units of measurement, assumptions made, format and file type of the data, a description of the data capture and collection methods, explanation of data coding and analysis performed (including syntax files). View a useful template from Cornell University’s “Guide to writing ‘ReadMe’ style metadata.”

How will you undertake documentation of data collection, processing and analysis, within your workflow to create consistent support material? Who will be responsible for this task?

Assign responsibilities for documentation: Individual roles and workflows should include gathering data documentation as a key element.

Establishing responsibilities for data management and documentation is useful if you do it early on, ideally in advance of data collection and analysis, Some researchers use CRediT taxonomy to determine authorship roles at the beginning of each project. They can also be used to make team members responsible for proper data management and documentation.

Consider how you will capture this information and where it will be recorded, to ensure accuracy, consistency, and completeness of the documentation.

Do you plan to use a metadata standard? What specific schema might you use?

Metadata for datasets: DataCite has developed a metadata schema specifically for open datasets. It lists a set of core metadata fields and instructions to make datasets easily identifiable and citable.

There are many other general and domain-specific metadata standards. Dataset documentation should be provided in one of these standard, machine readable, openly-accessible formats to enable the effective exchange of information between users and systems. These standards are often based on language-independent data formats such as XML, RDF, and JSON. There are many metadata standards based on these formats, including discipline-specific standards. Read more about metadata standards at the UK Digital Curation Centre's Disciplinary Metadata resource.

How will you make sure that a) your primary data collection methods are documented with transparency and b) your secondary data sources (i.e., data you did not collect yourself) — are easily identified and cited?

Document your process: It is useful to consult regularly with members of the research team to capture potential changes in data collection/processing that need to be reflected in the documentation.

Storage and Backup

List your anticipated storage needs (e.g., hard drives, cloud storage, shared drives). List how long you intend to use each type and what capacities you may require.

Estimating data storage needs: Storage-space estimates should take into account requirements for file versioning, backups, and growth over time.

If you are collecting data over a long period (e.g. several months or years), your data storage and backup strategy should accommodate data growth. Include your back-up storage media in your estimate.

What is your anticipated backup and storage schedule? How often will you save your data, in what formats, and where?

Follow the 3-2-1 rule to prevent data loss: It's important to have a regular backup schedule — and to document that process — so that you can review any changes to the data at any point during the project. The risk of losing data due to human error, natural disasters, or other mishaps can be mitigated by following the 3-2-1 backup rule: Have at least three copies of your data; store the copies on two different media; keep one backup copy offsite. Read more about storage and backup practices from the University of Sheffield Library and the UK Data Service.

Keeping ethics protocol review requirements in mind, what is your intended storage timeframe for each type of data (raw, processed, clean, final) within your team? Will you also store software code or metadata?

Ask for help: Your institution should be able to provide guidance with local storage solutions. Seek out RDM support at your Library or your Advanced Research Computing department.

Third-party commercial file sharing services (such as Google Drive and Dropbox) facilitate file exchange, but they are not necessarily permanent or secure, and servers are often located outside Canada. This may contravene ethics protocol requirements or other institutional policies.

An ideal solution is one that facilitates cooperation and ensures data security, yet is able to be adopted by users with minimal training. Transmitting data between locations or within research teams can be challenging for data management infrastructure. Relying on email for data transfer is not a robust or secure solution.

- **Raw data** are directly obtained from the instrument, simulation or survey.
- **Processed data** result from some manipulation of the raw data in order to eliminate errors or outliers, to prepare the data for analysis, to derive new variables, or to de-identify the human participants.
- **Analyzed data** are the results of qualitative, statistical, or mathematical analysis of the processed data. They can be presented as graphs, charts or statistical tables.
- **Final data** are processed data that have, if needed, been converted into a preservation-friendly format.

Sharing, Reuse and Preservation

How will your data (both raw and cleaned) be made accessible beyond the scope of the project and by researchers outside your team?

Help others reuse and cite your data: Did you know that a dataset is a scholarly output that you can list on your CV, just like a journal article?

If you publish your data in a data repository (e.g., Zenodo, Dataverse, Dryad), it can be found and reused by others. Many repositories can issue unique Digital Object Identifiers (DOIs) which make it easier to identify and cite datasets.

re3data.org is a directory of potential open data repositories. Consult with your colleagues to determine what repositories are common for data sharing in your field. You can also enquire about RDM support at your local institution, often in the Library or Advanced Research Computing. In the absence of local support, consult this Portage repository options guide.

Is digital preservation a component of your project and do you need to plan for long-term archiving and preservation?

How long should you keep your data? The length of time that you will keep or share your data beyond the active phase of your research can be determined by external policies (e.g. funding agencies, journal publishers), or by an understanding of the enduring (historical) value of a given set of data. The need to publish data in the short-term (i.e. for peer-verification purposes), for a longer-term access for reuse (to comply with funding requirements), or for preservation through ongoing file conversion and migration (for data of lasting historical value), will influence the choice of data repository or archive.

If you need long-term archiving for your data set, choose a preservation-capable repository. Digital preservation can be costly and time-consuming, and not all data can or should be preserved.

What data will you be sharing publicly and in what form (e.g. raw, processed, analyzed, final)?

Consider which types of data need to be shared to meet institutional or funding requirements, and which data may be restricted because of confidentiality/privacy/intellectual property considerations.

Have you considered what type of end-user license to include with your data?

Use open licenses to promote data sharing and reuse: Licenses determine what uses can be made of your data. Consider including a copy of your end-user license with your DMP.

As the creator of a dataset (or any other academic or creative work) you usually hold its copyright by default. However, copyright prevents other researchers from reusing and building on your work. Open Data Commons licenses and Creative Commons licenses are a free, simple and standardized way to grant copyright permissions and ensure proper attribution (i.e., citation). CC-BY is the most open license, and allows others to copy, distribute, remix and build upon your work —as long as they credit you or cite your work.

Even if you choose to make your data part of the public domain (with no restrictions on reuse), it is preferable to make this explicit by using a license such as Creative Commons' CC0. It is strongly recommended to share your data openly using an Open Data or Creative Commons license.

Learn more about data licensing at the Digital Curation Centre.

What tools and strategies will you take to promote your research? How will you let the research community and the public know that your data exists and is ready to be reused?

Data sharing as knowledge mobilization: Using social media, e-newsletters, bulletin boards, posters, talks, webinars, discussion boards or discipline-specific forums are good ways to gain visibility, promote transparency and encourage data discovery and reuse.

One of the best ways to refer other researchers to your deposited datasets is to cite them the same way you cite other types of publications. Publish your data in a repository that will assign a persistent identifier (such as a DOI) to your dataset. This will ensure a stable access to the dataset and make it retrievable by various discovery tools. Some repositories also create links from datasets to their associated papers, increasing the visibility of the publications.

Ethics and Legal Compliance

Are there institutional, governmental or legal policies that you need to comply with in regards to your data standards?

Determine jurisdiction: If your study is cross-institutional or international, you'll need to determine which laws and policies will apply to your research.

Compliance with privacy legislation and laws that may impose content restrictions in the data should be discussed with your institution's privacy officer or research services office.

If you collaborate with a partner in the European Union, for example, you might need to follow the General Data Protection Regulation (GDPR).

If you are working with data that has First Nations, Métis, or Inuit ownership, for example, you will need to work with protocols that ensure community privacy is respected and community harm is reduced. For further guidance on Indigenous data sovereignty, see OCAP Principles, and in a global context, CARE Principles of Indigenous Data Governance.

Will you encounter protected or personally-identifiable information in your research? If so, how will you make sure it stays secure and is accessed by approved team members only?

Get informed consent before you collect data: Obtaining the appropriate consent from research participants is an important step in assuring Research Ethics Boards that the data may be shared with researchers outside your project. Your informed consent statement may identify certain conditions clarifying the uses of the data by other researchers. For example, your statement may stipulate that the data will either only be shared for non-profit research purposes (you can use CC-by-NC — the non-commercial Creative Commons licence with attribution) or that the data will not be linked with other personally-identifying data. Note that this aspect is not covered by an open license. You can learn more about data security at the UK Data Service.

Inform your study participants if you intend to publish an anonymized and de-identified version of collected data, and that by participating, they agree to these terms. For sample language for informed consent: ICPSR Recommended Informed Consent Language for Data Sharing.

You may need to consider strategies to ensure the ethical reuse of your published dataset by new researchers. These strategies may affect your selection of a suitable license, and in some cases you may not be able to use an open license.

Before publishing or otherwise sharing a dataset are you required to obscure identifiable data (name, gender, date of birth, etc), in accordance with your jurisdiction's laws, or your ethics protocol? Are there any time restrictions for when data can be publicly accessible?

Privacy protection: Open science workflows prioritize being “as open as possible and as closed as necessary.” Think about any privacy concerns you may have in regards to your data, or other restrictions on access outlined in your ethics protocol. If your institution or funder regulates legal or ethical guidelines on what information must be protected, take a moment to verify you have complied with the terms for consent of sharing data. In the absence of local support for anonymization or de-identification of data, you can reach out to the Portage DMP Coordinator at support@portagenetwork.ca.