Data Management Plan Template: **Studying Molecular Interactions**

Abstract

The following template is intended for research projects combining experimental (scientific) methods with computer modeling/simulation to study molecular interactions.

Administrative Details

Template Author(s): Tatiana Zaraiskaya, University of New Brunswick

Published: April 9, 2021

DOI: 10.5281/zenodo.4683647

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

Version:

Version	Date	Changes
1.0	2021-04-09	Formatted for inaugural publication.

Data Production

What type(s) of data will be produced and in what format(s)?

Describe the type(s) of data that will be collected, such as: text, numeric (ASCII, binary), images, and animations.

List the file formats you expect to use. Keep in mind that some file formats are optimal for the long-term preservation of data, for example, non-proprietary formats such as text ('.txt'), comma-separated ('.csv'), TIFF (‘.tiff’), JPEG (‘.jpg’), and MPEG-4 (‘.m4v’, ‘mp4’). Files converted from one format to another for archiving and preservation may lose information (e.g. converting from an uncompressed TIFF file to a compressed JPG file), so changes to file formats should be documented. Guidance on file formats recommended for sharing and preservation are available from the UK Data Service, Cornell’s Digital Repository, the University of Edinburgh, and the University of British Columbia.

Examples of experimental image data may include: Magnetic Resonance (MR), X-ray, Fluorescent Resonance Energy Transfer (FRET), Fluorescent Lifetime Imaging (FLIM), Atomic-Force Microscopy (AFM), Electron Paramagnetic Resonance (EPR), Laser Scanning Microscope (LSM), Extended X-ray Absorption Fine Structure (EXAFS), Femtosecond X-ray, Raman spectroscopy, or other digital imaging methods.

Does this project involve the use or analysis of secondary data? What is the data-sharing arrangement for these data?

Describe any secondary data you expect to reuse. List any available documentation, licenses, or terms of use assigned. If the data have a DOI or unique accession number, include that information and the name of the repository or database that holds the data. This will allow you to easily navigate back to the data, and properly cite and acknowledge the data creators in any publication or research output.

For data that are not publicly available, you may have entered into a data-sharing arrangement with the data owner, which should be noted here. A data-sharing arrangement describes what data are being shared and how the data can be used. It can be a license agreement or any other formal agreement, for example, an agreement to use copyright-protected data.

Examples:

An example of experimental data from a secondary data source may be structures or parameters obtained from Protein Data Bank (PDB).

Examples of computational data and data sources may include: log files, parameters, structures, or trajectory files from previous modeling/simulations or tests.

What tools, devices, platforms, and/or software packages will be used to generate and manipulate data during the project?

List all devices and/or instruments and describe the experimental setup (if applicable) that will be utilized to collect empirical data. For commercial instruments or devices, indicate the brand, type, and other necessary characteristics for identification. For a home-built experimental setup, list the components, and describe the functional connectivity. Use diagrams when essential for clarity.

Indicate the program and software package with the version you will use to prepare a structure or a parameter file for modeling or simulation. If web-based services such as FALCON@home, ProModel, CHARMM-GUI, etc. will be used to prepare or generate data, provide details such as the name of the service, URL, version (if applicable), and description of how you plan to use it.

If you plan to use your own software or code, specify where and how it can be obtained to independently verify computational outputs and reproduce figures by providing the DOI, link to GitHub, or another source. Specify if research collaboration platforms such as CodeOcean, or WholeTale will be used to create, execute, and publish computational findings.

Describe the data flow through the entire project. What steps will you take to increase the likelihood that your results will be reproducible?

Reproducible computational practices are critical to continuing progress within the discipline. At a high level, describe the steps you will take to ensure computational reproducibility in your project, including the operations you expect to perform on the data, and the path taken by the data through systems, devices, and/or procedures. Describe how an instrument response function will be obtained, and operations and/or procedures through which data (research and computational outputs) can be replicated. Indicate what documentation will be provided to ensure the reproducibility of your research and computational outputs.

Examples:

Some examples of data procedures include: data normalization, data fitting, data convolution, or Fourier transformation.

Examples of documentation may include: syntax, code, algorithm(s), parameters, device log files, or manuals.

Documentation and Metadata

What documentation will be needed for the data to be read and interpreted correctly in the future?

Documentation can be provided in the form of a README file with information to ensure the data can be correctly interpreted by you or by others when you share or publish your data. Typically, a README file includes a description (or abstract) of the study, and any other contextual information required to make data usable by other researchers. Other information could include: research methodology used, variable definitions, vocabularies, units of measurement, assumptions made, format and file type of the data, a description of the data origin and production, an explanation of data file formats that were converted, a description of data analysis performed, a description of the computational environment, references to external data sources, details of who worked on the project and performed each task. Further guidance on writing README files is available from the University of British Columbia (UBC) and Cornell University.

Consider also saving detailed information during the course of the project in a log file. A log file will help to manage and resolve issues that arise during a project and prioritize the response to them. It may include events that occur in an operating system, messages between users, run time and error messages for troubleshooting. The level of detail in a log file depends on the complexity of your project.

How will you make sure that documentation is created or captured consistently throughout your project? What approaches will be employed by the research team to access, modify, and contribute data throughout the project?

Consider how you will capture this information in advance of data production and analysis, to ensure accuracy, consistency, and completeness of the documentation. Often, resources you've already created can contribute to this (e.g. publications, websites, progress reports, etc.). 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. Individual roles and workflows should include gathering data documentation as a key element.

Describe how data will be shared and accessed by the members of the research group during the project. Provide details of the data management service or tool (name, URL) that will be utilized to share data with the research group (e.g. Open Science Framework [OSF] or Radiam).

If you are using a metadata standard and/or online tools to document and describe your data, please list them here.

There are many general and domain-specific metadata standards. Dataset documentation should be provided in a standard, machine-readable, openly-accessible format 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. Read more about metadata standards here: UK Digital Curation Centre's Disciplinary Metadata.

Storage and Backup

What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it?

Storage-space estimates should consider 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. Similarly, a long-term storage plan is necessary if you intend to retain your data after the research project.

How and where will your data be stored and backed up during your research project?

Data may be stored using optical or magnetic media, which can be removable (e.g. DVD and USB drives), fixed (e.g. desktop or laptop hard drives), or networked (e.g. networked drives or cloud-based servers). Each storage method has benefits and drawbacks that should be considered when determining the most appropriate solution.

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.

Sharing and Archiving

What data will you be sharing and in what form?

Consider which data need to be shared to meet institutional, funding, or industry requirements. Consider which data will need to be restricted because of data-sharing arrangements with third parties, or other intellectual property considerations.

In this context, data might include research and computational findings, software, code, algorithms, or any other outputs (research or computational) that may be generated during the project. Research outputs might be in the form of:

- **raw data** are the data directly obtained from the simulation or modeling;
- **processed data** result from some manipulation of the raw data in order to eliminate errors or outliers, to prepare the data for analysis, or to derive new variables; or
- **analyzed data** are the results of quantitative, statistical, or mathematical analysis of the processed data.

What steps will be taken to help the research community know that your research data exist?

Possibilities include: data registries, data repositories, indexes, word-of-mouth, and publications. If possible, choose to archive your data in a repository that will assign a persistent identifier (such as a DOI) to your dataset. This will ensure stable access to the dataset and make it retrievable by various discovery tools.

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 (articles, books, proceedings). The Digital Curation Centre provides a detailed guide on data citation. Note that some data repositories also create links from datasets to their associated papers, thus increasing the visibility of the publications.

If you will use your own code or software in this project, describe your strategies for sharing it with other researchers.

Some strategies for sharing include:

1. Research collaboration platforms such as Code Ocean, Whole Tale, or other, will be used to create, execute, share, and publish computational code and data;
2. GitHub, GitLab or Bitbucket will be utilized to allow for version control;
3. A general public license (e.g., GNU/GPL or MIT) will be applied to allow others to run, study, share, and modify the code or software. For further information about licenses see, for example, the Open Source Initiative;
4. The code or software will be archived in a repository, and a DOI will be assigned to track use through citations. Provide the name of the repository;
5. A software patent will be submitted.

Which data (research and computational outputs) will be retained after the completion of the project? Where will your research data be archived for the long-term? Describe your strategies for long-term data archiving.

In cases where only selected data will be retained, indicate the reason(s) for this decision. These might include legal, physical preservation issues or other requirements to keep or destroy data.

There are general-purpose data repositories available in Canada, such as Scholars Portal Dataverse, and the Federated Research Data Repository (FRDR), which have preservation policies in place. Many research disciplines also have dedicated repositories. Scientific Data, for example, provides a list of repositories for scientific research outputs. Some of these repositories may request or require data to be submitted in a specific file format. You may wish to consult with a repository early in your project, to ensure your data are generated in, or can be transformed into, an appropriate standard format for your field. Re3data.org is an international registry of research data repositories that may help you identify a suitable repository for your data. The Repository Finder tool hosted by DataCite can help you find a repository listed in re3data.org. For assistance in making your dataset visible and easily accessible, contact your institution's library or reach out to the Portage DMP Coordinator at support@portagenetwork.ca.

Ethics and Intellectual Property

Are there any ethical or legal concerns related to your data that you will need to address? Are there any ownership or intellectual property concerns that could limit if/how you can share your research outputs?

Consider the end-user license, and ownership or intellectual property rights of any secondary data or code you will use. Consider any data-sharing agreements you have signed, and repository 'terms of use' you have agreed to. These may determine what end-user license you can apply to your own research outputs, and may limit if and how you can redistribute your research outputs. If you are working with an industry partner, will you have a process in place to manage that partnership, and an understanding of how the research outputs may be shared? Describe any foreseeable concerns or constraints here, if applicable.

Roles and Responsibilities

Identify who will be responsible for managing data during and after the project. Indicate the major data management tasks for which they will be responsible.

Identify who will be responsible -- individuals or organizations -- for carrying out these parts of your project. Consider including the time frame associated with these staff responsibilities, and document any training needed to prepare staff for data management duties.

How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator?

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

What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

This estimate should incorporate data management costs incurred during the project as well as those required for the longer-term support for the data after the project has completed. Items to consider in the latter category of expenses include the costs of curating and providing long-term access to the data. It might include technical aspects of data management, training requirements, file storage & backup, the computational environment or software needed to manage, analyze or visualize research data, and contributions of non-project staff. Read more about the costs of RDM and how these can be addressed in advance: “What will it cost to manage and share my data?” by OpenAIRE. There is also an online data management costing tool available to help determine the costs and staffing requirements in your project proposal.