DMP TEMPLATE FOR A GENOMICS STUDY

(Examples were extracted from a fictional study and the template was adapted from the Portage general template - https://dmp-pgd.ca/template_export/878086536.pdf)

This document is a working draft and will continue to undergo piloting and refinement. It represents an ongoing effort to improve data management plans in the biomedical sciences.

The individuals involved in the development of the template include Dr. Anna Catharina V. Armond, Dr. Theodore J. Perkins, Dr. Michael Hoffman, and Dr. Kelly D. Cobey.

1. DATA DESCRIPTION AND COLLECTION

1a. Describe the study for which the data are being collected.

Guidance:
Add a brief description identifying the study design, population, interventions, control and outcome.

1b. What types of data will you collect, create, link to, acquire and/or record?

Guidance:
Provide a list of the following:
- The type of data collected (Examples of data types include numeric data/metadata, high-througput sequencing data, mass spectrometry data, imaging/microscopy, audio, video, text, tabular data, modeling data, spatial data, instrumentation data)
- A description of study instruments along with their reliability and validity, if known.
- List the names/types of the variables collected
- Specify where data collection forms can be found

1c. How will new data be collected or produced and/or how will existing data be re-used?

Guidance:
- Explain which methodologies or software will be used if new data are collected or produced.
- State any constraints on re-use of existing data if there are any.
- Explain how data provenance will be documented.
- Briefly state the reasons if the re-use of any existing data sources has been considered but discarded

1d. What file formats will your data be collected in? Will these formats allow for data reuse, sharing, and long-term access to the data?

Guidance:
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 preferred file formats: UBC Library, UK Data Service, MIAME/MINSEQE, NCI GDC, or HUPO PSI.

1d. What conventions and procedures will you use to structure, name and version-control your files to help you and others better understand how your data are organized?

Guidance:
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: UBC Library or UK Data Service.

1e. How will the research team and other collaborators access, modify, and contribute data throughout the project?

Guidance:
An ideal solution is one that facilitates co-operation and ensures data security yet is able to be adopted by users with minimal training. Include who will have access to the data and how the access will be granted. 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. Third-party commercial file sharing services (such as Google Drive and Dropbox) facilitate file exchange, but they are not necessarily permanent or secure, and are often located outside Canada. Please contact your institution to develop the best solution for your research project.

2. DOCUMENTATION AND METADATA

2a. What documentation will be needed for the data to be read and interpreted correctly in the future?

Guidance:
Typically, good 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), and details of who has worked on the project and performed each task, etc.

2b. Describe in detail the data-level descriptions, codes and definitions.

Guidance:
List the variables 'names and add units, codes and definitions.

2c. How will you make sure that documentation is created and captured consistently throughout your project?

Guidance:
Consider how you will capture this information and where it will be recorded, ideally in advance of data collection 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.

2d. If you are using a metadata standard and/or tools to document and describe your data, please list here.

Guidance:
There are many general and domain-specific metadata standards. Dataset documentation should be provided in one of these standards, 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.

Dataset documentation may also include a controlled vocabulary, which is a standardized list of terminology for describing information. Examples of controlled vocabularies include the Library of Congress Subject Headings (LCSH),NASA's Global Change Master Directory (GCMD) Keywords, Sequence Ontology, Uber-Anatomy Ontology (Uberon), and the Disease Ontology. Read more about metadata standards: UK Digital Curation Centre's Disciplinary Metadata.

3. STORAGE AND BACKUP

3a. 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?

Guidance:
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. Similarly, a long-term storage plan is necessary if you intend to retain your data after the research project. It is worth verifying the institutions 'infrastructure and storage options.

3b. How and where will your data be stored and backed up during your research project?

Guidance:
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: The rule consists of, when possible, have at least three copies of your data. Store the copies on two different media, and keep one backup copy offsite.

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.
Further information on storage and backup practices is available from the University of Sheffield Library and the UK Data Service. Where data, especially raw omics or imaging data, is too large to follow the 3-2-1 back up rule, or where privacy issues prevent data leaving the institution, two duplicate copies should be retained on different, secure, internal systems.

4. PRESERVATION

4a. Where will you deposit your data for long-term preservation and access at the end of your research project?

Guidance:
The issue of data retention should be considered early in the research lifecycle. Data-retention decisions can be driven by external policies (e.g. funding agencies, journal publishers), or by an understanding of the enduring value of a given set of data. The need to preserve data in the short-term (i.e. for peer-verification purposes) or long-term (for data of lasting value), will influence the choice of data repository or archive. A helpful analogy is to think of creating a 'living will' for the data, that is, a plan describing how future researchers will have continued access to the data.
If you need assistance locating a suitable data repository or archive, please contact your Library.
re3data.org is a directory of potential open data repositories. Verify whether or not the data repository will provide a statement agreeing to the terms of deposit outlined in your Data Management Plan.

4b. Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.

Guidance:
Some data formats are optimal for long-term preservation of data. For example, non-proprietary file formats, such as text ('.txt') and comma-separated ('.csv'), are considered preservation-friendly. The UK Data Service provides a useful table of file formats for various types of data. Keep in mind that preservation-friendly files converted from one format to another may lose information (e.g. converting from an uncompressed TIFF file to a compressed JPG file), so changes to file formats should be documented.
Identify steps required following project completion in order to ensure the data you are choosing to preserve or share is anonymous, error-free, and converted to recommended formats with a minimal risk of data loss.
Read more about anonymization: UBC Library or UK Data Service.

4C. If your study involves in-house or bespoke analysis pipelines, how will those pipelines be documented and will they be shared so that other researchers can reproduce your results or apply the same analyses to their data?

Guidance:
Omics studies often involve a sequence or flow graph of programmatic data analysis steps, and each step often needs to be configured with multiple parameters. Such pipelines should be described as part of the metadata of the study and/or in associated publications. But it is nevertheless difficult to accurately reconstruct such pipelines from written descriptions. Therefore, sharing of the analysis code itself is preferable wherever possible. This can be done via a PI or project website, through publisher-specific mechanisms when an article is published, through public code sharing and versioning sites such as github, or as a docker or singularity container.

5. SHARING AND REUSE

5a. What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final, and metadata). Will sharing be open or by restricted access request? If a restricted access request process is used, describe this process.

Guidance:

- Raw data are the data 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.

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

- Consider the supporting documentation for data analyses (Metadata).

5b. What type of end-user license will you use for your data?

Guidance:
Licenses determine what uses can be made of your data. Funding agencies and/or data repositories may have end-user license requirements in place; if not, they may still be able to guide you in the development of a license. Once created, please consider including a copy of your end-user license with your Data Management Plan. Note that only the intellectual property rights holder(s) can issue a license, so it is crucial to clarify who owns those rights.
There are several types of standard licenses available to researchers, such as the Creative Commons licenses and the Open Data Commons licenses. In fact, for most datasets it is easier to use a standard license rather than to devise a custom-made one. Note that even if you choose to make your data part of the public domain, it is preferable to make this explicit by using a license such as Creative Commons' CC-BY-4.0..
Read more about data licensing: UK Digital Curation Centre.

5c. What steps will be taken to help the research community know that your data exists?

Guidance:
Possibilities include: data registries, repositories, indexes, presentations, publications.
If possible, choose 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.
Contact your institution for assistance as in making your dataset visible and easily accessible.
Reused from NIH. (2009). Key Elements to Consider in Preparing a Data Sharing Plan Under NIH Extramural Support. National Institutes of Health.
Other resources can be found at https://ohri.ca/journalology/data-and-materials-sharing, and https://journalologytraining.ca/all-courses/.

6. RESPONSIBILITIES AND RESOURCES

6a. Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

Guidance:
Your data management plan has identified important data activities in your project. Identify who will be responsible – individuals, organizations, or committees (e.g., steering committee, adjudication committee, data management team) -- for carrying out these parts of your data management plan. Include names, affiliations, and roles. If applicable, identify the study sponsor, including name and contact information. The sponsor can be defined as the individual, company, institution, or organization assuming responsibility of the study. Include the sponsor's roles and responsibilities. You should also include the timeframe associated with these staff responsibilities and any training needed to prepare staff for these duties.

It is relevant to report if your study will have a Data and Safety Monitoring Committee. Identify individuals, roles and organizations, and their workflows.

6b. How will responsibilities for managing data activities be handled if substantive changes occur in the personnel overseeing the project's data, including a change of Principal Investigator?

Guidance:
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.

6b. What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

Guidance:
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 is finished. Items to consider in the latter category of expenses include the costs of curating and providing long-term access to the data. Some funding agencies state explicitly the support that they will provide to meet the cost of preparing data for deposit. This might include technical aspects of data management, training requirements, file storage & backup, and contributions of non-project staff. Find more instructions about budget estimation here: https://www.utwente.nl/en/service-portal/services/lisa/resources/files/library-public/dcc-rdm-costs-estimation.pdf

7. ETHICS AND LEGAL COMPLIANCE

7a. If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project? If applicable, what strategies will you undertake to address secondary uses of sensitive data?

Guidance:
Consider where, how, and to whom sensitive data with acknowledged long-term value should be made available, and how long it should be archived. These decisions should align with Research Ethics Board requirements. The methods used to share data will be dependent on a number of factors such as the type, size, complexity and degree of sensitivity of data. Outline problems anticipated in sharing data, along with causes and possible measures to mitigate these. Problems may include confidentiality, lack of consent agreements, or concerns about Intellectual Property Rights, among others. In some instances, an embargo period may be justified; these may be defined by a funding agency's policy on research data.
Reused from: DCC. (2013). Checklist for a Data Management Plan. v.4.0. Edinburgh: Digital Curation Centre
Restrictions can be imposed by limiting physical access to storage devices, by placing data on computers that do not have external network access (i.e. access to the Internet), through password protection, and by encrypting files. Sensitive data should never be shared via email or cloud storage services such as Dropbox.
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. The consent statement may identify certain conditions clarifying the uses of the data by other researchers. For example, it may stipulate that the data will only be shared for non-profit research purposes or that the data will not be linked with personally identified data from other sources.
Read more about data security: UK Data Service

7b. How will you manage legal, ethical, and intellectual property issues?

Guidance:
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. Research Ethics Boards are central to the research process.
Include here a description concerning ownership, licensing, and intellectual property rights of the data. Terms of reuse must be clearly stated, in line with the relevant legal and ethical requirements where applicable (e.g., subject consent, permissions, restrictions, etc.).