Data Management Plan

Example DMP for SFI

Contact person: ***PostDoc1*** *(postdoc1@example.ie)*

Based on: *Common DSW Knowledge Model, 2.4.4 (dsw:root:2.4.4)*

Project phase: *Before Submitting the Proposal*

Created by: ***Siobhan Cleary*** *(siobhan.cleary@ucd.ie)*
University College Dublin

Generated on: *12 Jul 2023*

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
There are no named versions		

Contributors

The following contributors are related to the project of this DMP:

- PostDoc1
  postdoc1@example.ie
  Roles: Contact Person, Data Manager
- PhD Student1
  phdstudent1@example.ie
  Roles: Researcher

Projects

We will be working on the following projects and for those are the data and work described in this DMP.

Investigating expression of genes in humans

Start date: *2023-05-15*

End date: *2026-05-15*

Funding: *Science Foundation Ireland: grant number not yet given (planned)*

This is an example DMP when using human RNA-Seq data in your research.

Section A: Data Collection

1. What data will you collect or create?

Instrument datasets

The following instrument datasets will be acquired in the project:

- **In house RNA-Seq dataset**

  This dataset will be collected by experts in the project, with our own equipment.

  The equipment is very well described and known.

Data formats and types

We will be using the following data formats and types:

- **FASTQ Sequence and Sequence Quality Format**

  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 360 files of average size 10 GB (i.e. approximately 3600.0 GB in total).

- **Binary Alignment Map Format**

  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 180 files of average size 15 GB (i.e. approximately 2700.0 GB in total).

- **Tab-separated values**

  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 1180 files of average size 0.005 GB (i.e. approximately 5.9 GB in total).

2. How will the data be collected or created?

Instrument datasets

- **In house RNA-Seq dataset**

  For this dataset, we are using the following instruments:

  - **NextSeq 1000**

Data storage and file conventions

We will use a filesystem with files and folders with the following folder conventions:

- There will be a **(sub)folder for each (repeated) analysis**.
- There will be a **(sub)folder for each step in the analysis workflow**.

Moreover, we have made appointments about naming the files.

We will not be storing data in an "object store" system.

We will not use a relational database system to store project data.

Section B: Documentation and Meta-data

3. What documentation and meta-data will accompany the data?

List of data to be published is given in Section E, Question 9. This also includes information about catalogs where the data can be found. Information about data types used is given in Section A, Question 1.

We will be documenting the data with DataCite metadata standard.

Section C: Ethics and Legal Compliance

4. How will you manage any ethical issues?

Data we reuse

- **The Cancer Genome Atlas** – the existing consent covers our reuse.

5. How will you manage copyright and Intellectual Property Rights (IPR) issues?

We will be working with the philosophy *as open as possible* for our data.

The data cannot become completely open immediately because of:

- legal reasons

- we have other than paper-publishing reasons: Patient Identifiable

Data that is not legally restrained will be released after a fixed time period, unconditionally.

All data will be owned by the institute.

For the reference and non-reference data sets that we reuse, conditions are as follows:

- **GRCh38.p14** – freely available for any use (public domain or CC0).
- **The Cancer Genome Atlas** – freely available with obligation to quote the source (e.g. CC-BY).

Section D: Storage and Backup

6. How will the data be stored and backed up during the research?

Storage needs are large at the beginning and will be reduced later.

All essential data is also stored elsewhere to prevent a total loss of data. We will make (automated) backups of all data stored outside of the working area.

7. How will you manage access and security?

Project members will not store data or software on computers in the lab or external hard drives connected to those computers.They will not carry data with them (e.g. on laptops, USB sticks, or other external media). All data centers where project data is stored carry sufficient certifications. All project web services are addressed via secure HTTP (https://...). Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. The risk of information leak in the project or organization is acceptably low. The possible impact to the project or organization if information is vandalised is small.

All personal data will be anonymized as early as possible.

Only all project members have read/write access to the data.

Section E: Selection and Preservation

8. Which data are of long-term value and should be retained, shared, and/or preserved?

We plan to produce the following datasets:

- **FASTQ Files** (published) – This data set will be kept available as long as technically possible. – The metadata will be available even when the data no longer exists.
- **BAM files** (published)
- **Expression Counts** (published)

9. What is the longterm preservation plan for the dataset?

- **FASTQ Files** (published)
  The distributions will be stored in:
  - Domain-specific repository: The European Genome-phenome Archive.
    We don't need to contact the repository because it is a routine for us.

  We will be adding a reference to the published data to at least one data catalogue.
- **BAM files** (published)
  The distributions will be stored in:
  - Domain-specific repository: The European Genome-phenome Archive.
    We don't need to contact the repository because it is a routine for us.
- **Expression Counts** (published)
  The distributions will be stored in:
  - Domain-specific repository: The European Genome-phenome Archive.
    We don't need to contact the repository because it is a routine for us.

None of the used repositories charge for their services.

We have a reserved budget for the time and effort it will take to prepare the data for publication.

Section F: Data Sharing

10. How will you share the data?

- **FASTQ Files**
  The distributions will be available as follows:
  - Open (shared with anyone) using a domain-specific repository: The European Genome-phenome Archive. The distribution will be available under the following license:
    - Available under some restrictions, which we will follow in our project: Sequencing data cannot be released as open access due to the ability to identify individuals using genetic variants. Therefore, this data will be released as controlled access to researchers who have approved access, under a general research use license. Re-users will be able to get access through a *Data Access Committee* for the project. The conditions will be published as part of open metadata.

  We will be adding a reference to the published data to at least one data catalogue.
- **BAM files**
  The distributions will be available as follows:
  - Open (shared with anyone) using a domain-specific repository: The European Genome-phenome Archive. The distribution will be available under the following license:
    - Available under some restrictions, which we will follow in our project: This data cannot be released as open access due to the ability to identify individuals using genetic variants. Therefore, this data will be released as controlled access to researchers who have approved access, under a general research use license. Re-users will be able to get access through a *Data Access Committee* for the project. The conditions will be published as part of open metadata.
- **Expression Counts**
  The distributions will be available as follows:
  - Open (shared with anyone) using a domain-specific repository: The European Genome-phenome Archive. The distribution will be available under the following license:
    - Freely available with obligation to quote the source (e.g. CC-BY).

Information about used repositories (i.e. where will potential users find out about the data) is provided in Section E, Question 9.

Embargo on the data is described in Section C, Question 5, and Section F, Question 11.

11. Are any restrictions on data sharing required?

Ethical and legal restrictions are documented under Section C. We have used the Data Stewardship Wizard, which made us aware of options to minimize the restrictions.

Data cannot be completely open due to legal reasons. But data that is not legally restrained will be released after a fixed time period.

A data sharing agreement will be required. People can apply to the data access committee that we will set up.

Section G: Responsibilities and Resources

12. Who will be responsible for data management?

Siobhan Cleary is responsible for implementing the DMP, and ensuring it is reviewed and revised.

PostDoc1 is responsible for maintaining the finished resource.

13. What resources will you require to deliver your plan?

To execute the DMP, additional specialist expertise is required. We will be training existing staff.

Charges applied by data repositories (if any) are mentioned already in Section E, Question 9.