Hands-On Data Management Plan Exercise

Goal

Make a Data Management Plan (DMP) for the mock Project X application by following the five DMP steps detailed below.

Discuss each step and DMP item among your group, and when you are ready, write the item on a post-it and stick it to your DMP Canvas in the appropriate section.

Keep in mind that a DMP is a living document—do not be trapped by your decisions as you progress through the steps—you can always create a new version of your DMP.

Feel free to deduce or make up any information you need that is not explicitly described in the mock Project X application, according to your own judgment of what is sensible.

Project X (Application to Fundação para a Ciência e Tecnologia)

Title of the project: Unveiling the mechanisms of Disease X

Participants:

- Prof. Coor Dinator (coor.dinator@biodata.pt) [PI & DMP Coordinator]
- Dr. Dat Manger (dat.manger@biodata.pt) [Data Manager]
- Dr. Col Hector (col.hector@biodata.pt) [Clinical Data & Sample Collector]
- Dr. R. Sercher (r.sercher@biodata.pt) [Researcher]
- Mrs. A. D’Min (a.dmin@biodata.pt) [Project Manager]

Host Institution: BioData.pt

Start date: January 1st, 2021

Duration: 36 months

Abstract:

The cause of Disease X has been recently discovered to be a virus, phage X, which infects normal gut bacteria and leads them to become virulent and cause chronic intestinal infection. Although non-fatal, this disease has been spreading rapidly in Europe, with costs in health-care reaching the tens of millions of Euros.

This project aims to uncover the mechanisms of disease X by sequencing phage X and studying the effects of its infection in human gut microbiota at the population and molecular level. We will assess which bacterial taxa are infected by phage X and what effect the infection has on the relative abundance of the various taxa, as well as what effect the infection has on the infected taxa at the gene expression level.

The project will be a key step towards improving treatment for Disease X, and potentially being able to cure it.

Research plan and method summary:

The project will be divided into four activities:

1. Sample collection
2. Phage X sequencing
3. 16S sequencing
4. Metatranscriptomics

In the sample collection activity, we will define a study group of volunteer disease X patients, numbering no less than 20, and a control group comprising their close relatives, 1-2 per patient. We will collect stool samples from each of the volunteers.

In the Phage X sequencing activity, we will carry out DNA sequencing of the stool samples and assemble the genome of Phage X. In order to facilitate the assembly while enabling the reliable identification of sequence variants, we will combine the higher quality but short read sequencing technology of the Illumina NextSeq 500 sequencer with the long read but lower quality technology of the Nanopore MinION sequencer.

In the 16S sequencing activity, we will do microbial diversity analysis of the stool samples based on DNA sequencing of the 16S rRNA gene, to assess the impact of Phage X infection of the diversity and relative abundance of gut bacteria taxa. This sequencing will be carried out in Illumina NextSeq 500 sequencers.

In the metatranscriptomics activity, we will assess the effect of Phage X infection at the gene expression level through RNA sequencing of the samples. This sequencing will be carried out in Illumina NextSeq 500 sequencers.

Expected Data & Metadata Outputs:

1. Sample Collection:
   - Patient clinical data (< 1 MB)
   - Sample identification table (< 1 MB)
2. Phage X sequencing
   - Raw FASTQ sequencing data - NextSeq (60 MB)
   - Sample preparation & sequencing metadata - NextSeq (< 1 MB)
   - Raw FASTQ sequencing data - MinION (1 GB)
   - Sample preparation & sequencing metadata - MinION (< 1 MB)
   - Assembled Phage X genome (< 1 MB)
   - Assembly metadata (< 1 MB)
3. 16S sequencing
   - Raw FASTQ sequencing data (15 GB)
   - Sample preparation & sequencing metadata - NextSeq (< 1 MB)
   - Biome tables (< 1 MB)

4. Metatranscriptomics
    - ○ Raw FASTQ sequencing data (120 GB)
    - ○ Sample preparation & sequencing metadata - NextSeq (< 1 MB)
    - ○ RNAseq count tables (< 1 MB)
    - ○ Differential expression test results (< 1 MB)

DMP Step 1 – Administrative Metadata

Available time: 15 minutes

Information to collect:

In this step you have to characterize the DMP itself as well as list the people involved in it. This information is divided into the following sections, which are further detailed in the table below:

- DMP Characterization - General information characterizing the DMP.
- Contact - A contact (person or institution) for the DMP.
- DMStaff - A listing of staff members and their roles in this DMP.

Step 1 Table (**Sections [with cardinality]** & Fields)

Mandatory	Field	Description	Example / Possible Values
DMP Characterization [1]			
✔	Title	The title of this DMP	Board Game Data Management Plan
	Description	A short free text description of the DMP	This DMP serves to characterize the functioning of the board game production unit at Happy Corp in Lisbon
✔	Language	Language of the DMP expressed using ISO 6391-1 two letter country code	en
✔	Created	The date and time of the first version of this DMP	2019-09-20T15:43:00
✔	Modified	The date and time of a modification to the DMP. This serves as the version	2019-09-23T11:14:00

✔	Ethical Issues	Indication of the presence of ethical issues in the data described in the DMP	Possible values: yes, no, unknown
	Ethical Issues Description	A description of existing ethical issues	Some of our board games might reflect past society values that may be offensive in the present
Contact [1]			
✔	Name	Name of the institution or person responsible for the DMP	Board Game Unit at Happy Corp
✔	E-mail	E-mail address for the institution or person responsible for the DMP	main.bgu@happy.corp.pt
DMP Staff [*]			
✔	Name	Name of staff member	Albert Lamorisse
✔	E-mail	E-mail address of the staff member	alamorisse@happy.corp.pt
✔	Contribution Type	The type of contribution this staff member has	Possible values: ContactPerson, Data Collector, Data Curator, DataManager, Distributor, Editor, HostingInstitution, Producer, ProjectLeader, ProjectManager, ProjectMember, RegistrationAgency, RegistrationAuthority,RelatedPerson, Researcher, ResearchGroup, RightsHolder, Sponsor, Supervisor, WorkPackageLeader, Other

DMP Step 2 – Project & Funding

Available time: 10 minutes

Information to collect:

In this step you have to characterize the project(s) covered by the DMP and their source(s) of funding. In this case there is only one project, but note that a DMP can cover several projects, or just part of a project. This information is divided into the following sections, which are further detailed in the table below:

- Project - Information regarding the project to which the DMP is associated.
- Funding - Specific information pertaining the funding to the project.

Step 2 Table (**Sections [with cardinality]** & Fields)

Mandatory	Field	Description	Example / Possible Values
Project [1+]			
✔	Title	The title of the project	Iberian Risk
	Description	A short free text description of the project	This project focuses on developing a Risk board game themed on the peninsular war of 1807-1814
✔	Project Start	The starting date of the project	2018-09-01T18:00:00
✔	Project End	The end date of the project	2019-09-22T18:00:00
Funder [*]			
✔	Funder	A unique identifier for the funder	Happy Corp
✔	Grant	A unique identifier for the grant	HC/2018/100
✔	Funding Status	An indication of the stage of the project lifecycle	Possible values: Planned, Applied, Granted, Rejected

DMP Step 3 – Dataset Characterization

Available time: 20 minutes

Information to collect:

In this step you must characterize the datasets that are encompassed by the DMP. A dataset must always include a generic characterization of its data (e.g., the type of data, if sensitive or personal data is present, the language in which the data is expressed, etc.). Additional descriptions should be given on existing security and privacy policies, technical resources used to create or process the data, or metadata standards applicable to the data. This information is divided into the following sections, which are further detailed in the table below:

- Datasets - General information about all datasets created in the context of a project.
- Security and Privacy - A characterization of any security and privacy policies associated with each of the identified datasets.
- Technical Resource - A characterization of any technical resources associated to each of the identified datasets.
- Metadata - A characterization of metadata associated to each of the identified datasets.

Step 3 Table (**Sections [with cardinality]** & Fields)

Mandatory	Field	Description	Example / Possible Values
Datasets [1+]			
✔	Title	The title of the dataset	Iberian Risk territories
	Description	A short free text description of the dataset	This dataset comprises all the information on the Risk territory cards
✔	Language	Language of the dataset expressed using ISO 6391-1 two letter country code	en
	Keyword	A set of keywords that define the dataset	Risk, territory, cards
✔	Dataset Type	The type of resources in this dataset according to a controlled vocabulary	Possible values: website, image, software, sound, book, paper, patent, report, thesis, etc
✔	Issued	The date in which this dataset was first issued	2019-09-04T20:40
	Preservation Statement	A statement about the preservation requirements of the dataset	This dataset must be preserved, to enable the creation of future editions that use some or all of these territories

✔	Sensitive Data	An indication of the presence of sensitive data in the dataset	Possible values: yes, no, unknown
✔	Personal Data	An indication of the presence of personal data in the dataset	Possible values: yes, no, unknown
	Data Quality Assurance	A description of data quality assurance policies	All territories are named according to their respective official names in the english language
Security & Privacy [*]			
✔	Title	The title of a security or privacy policy	Reproducibility
	Description	A short free text description of the security or privacy policy	The images of the risk territory cards will be kept in a password protected server
Technical Resource [*]			
✔	Technical Resource Identifier	A unique identifier for a technical resource	AutoCAD
	Description	A short free text description of the technical resource	AutoCAD was used to draw all the risk territory cards
Metadata [*]			
✔	Title	The title of the metadata policy	Portuguese Constitution
	Description	A short free text description of the metadata policy	The names of all Portuguese territories respect the district names listed in the Portuguese Constitution
✔	Language	Language of the metadata policy expressed using ISO 6391-1 two letter country code	en

DMP Step 4 – Preservation & Publication

Available time: 15 minutes

Information to collect:

In this step you are to characterize the preservation and publication policies for the datasets. For each dataset identified in step 3, you must describe its distribution plans (e.g., the file format, the data access policy, the size of the dataset, etc.). Each distribution plan should be associated with one or more licence agreements that regulate data access and usage, and with one or more data hosts (e.g. a public or private repository, a library, etc.) that will store the dataset and make it available. This information is divided into the following sections, which are further detailed in the table below:

- Distribution - Information regarding the policies on how each dataset is to be distributed.
- Host - A characterization of the data host for each distribution.
- License - A characterization of the licenses associated with each distribution policy.

Step 4 Table (**Sections [with cardinality]** & Fields)

Mandatory	Field	Description	Example / Possible Values
Distribution [1+]			
✔	Title	The title of the distribution	Risk territory cards
	Description	A short free text description of the host	The digital repository of Happy Corp
	Format	The media type, according to IANA	application/pdf
	Byte Size	The size of the dataset in bytes	1142000
	Data Access	Indicates the mode of access for the data	Possible values: open, shared, closed
	Availability	A date indicating at which the data was or will be available	2069-09-01T00:00:00
Host [1+]			
✔	Title	The title of the host	Happy Corp Digital Repository

	Description	A short free text description of the security or privacy policy	The images of the risk territory cards will be kept in a password protected server
	Support Versioning	Does the host support versioning	Possible values: yes, no, unknown
	Backup Type	The type of backup used in this host	Long term storage in tapes, short term storage in hard drives
	Backup Frequency	The frequency of backups in this host	Weekly
	Availability	The availability of data in this host in percentage	99.5
	Certification	The host repository certification standard	Possible Values: DIN31644, DINI-Zertifikat, DSA, ISO16363, ISO16919, TRAC. WDS, CoreTrustSeal, other, none
	Geo Location	Physical location of the data expressed using ISO 3166-1 country code	PT
	PID System	Persistent Identifier System in this host	Possible values: ark, doi, hdl, purl, urn, other, none
License [1+]			
✔	Licence	Type of licence under which the dataset is distributed	Possible Values: CC0, CC BY, CC BY-SA, CC BY-NC, CC BY-NC-SA, CC BY-NC-ND, other, none
✔	Start Date	The date from which the license is in vigor for the dataset (if it is future, it indicates an embargo period)	2019-09-01T00:00:00

Creative Commons License Reference Table

Licence	Can I copy & redistribute the work?	Is it required to attribute the author?	Can I use the work commercially?	Am I allowed to adapt the work?	Can I change the licence when redistributing?

CC0	✔	✘	✔	✔	✔
CC BY	✔	✔	✔	✔	✔
CC BY-SA	✔	✔	✔	✔	✘
CC BY-ND	✔	✔	✔	✘	✔
CC BY-NC	✔	✔	✘	✔	✔
CC BY-NC-SA	✔	✔	✘	✔	✘
CC BY-NC-ND	✔	✔	✘	✘	✔

DMP Step 5 – Costs

Available time: 15 minutes

Information to collect:

In this step you should look back to the previous steps, identify all costs that are associated with this DMP, and list them in the single section:

- Costs - Information on the costs associated with the DMP.

Step 5 Table (**Sections [with cardinality]** & Fields)

Mandatory	Field	Description	Example / Possible Values
Cost [*]			
✔	Title	The title of the cost	Storage
	Description	A short free text description of the cost	Storage in disk of the Iberian Risk card templates
	Value	The numeric value of the cost (under a given currency)	1000
	Currency	The currency code of the preceding value, according to ISO 4217	eur

DMP Presentation

Available time: 10 minutes

Now that you've created your DMP, we ask you to prepare a short presentation that focuses on the following points:

- Key decisions you made during the creation process, and the motivating factors for those decisions;
- Any relevant deviations you made from the guide;
- Positive and negative feedback on the Ready For BioData Management workshop.