BioData.pt elixir PORTUGAL

TRAINING DATA STEWARDS

FOR LIFE SCIENCES

Data Management Plans & Assignment 1 Feedback

Filipa Pereira, FCT-FCCN

11.04.2022

Core requirements for DMPs

1. Data description and collection or re-use of existing data
2. Documentation that will accompany the data and metadata
3. Storage and backup during the research process
4. Legal and ethical requirements
5. Data sharing and long-term preservation
6. Data management responsibilities and resources

TRAINING DATA STEWARDS
FOR LIFE SCIENCES

Based on the "Practical Guide to the International Alignment of Research Data Management", from Science Europe.

Core requirements for DMPs

1. **Data description and collection or re-use of existing data**
2. Documentation that will accompany the data and metadata
3. Storage and backup during the research process
4. Legal and ethical requirements
5. Data sharing and long-term preservation
6. Data management responsibilities and resources

1 Data description, collection and re-use of data

Data can have two generic origins/provenances:

- New data
- Existing data

1 Data description, collection and re-use of data

New data:

- **Methodologies** and **software**
- Data **description**
- State **whether data reuse was considered** and why it was ruled out.

1 Data description, collection and re-use of data

Re-use of data:

- **What** data?
- Under which **terms of use**?
- Are there **any restrictions** on the use of the data?
- How the **provenance of the data** will be documented?

1 Data description, collection and re-use of data

Type of data:

- Numeric (data base, spreadsheets);
- Textual (documents);
- Image;
- Audio;
- Video;
- Mixed;
- Other.

1 Data description, collection and re-use of data

Data format:

- The way in which the data is **encoded for storage**, often reflected by the filename extension.
  Exs: pdf, xls, doc, txt ou rdf.
- Justify the **use of certain formats.**
  Exs: decisions may be based on staff expertise within the host organisation, standards accepted by data repositories, widespread usage within the research community, etc.
- Give **preference to open and standard formats** as they facilitate sharing and long-term re-use of data.
- Several **repositories** provide lists of **‘preferred formats'.**

1 Data description, collection and re-use of data

Data volume:

- **Storage:**
  Exs:
  ☐ 0 – 10 GB
  ☐ 10 – 100 GB
  ☐100 – 1000 GB
  ☐ >1000 GB

  and/or

- **Number:**
  Exs: objects, folders, etc.

1 Data description, collection and re-use of data

Examples:

1.2.3 What types of data will the project generate/collect?

[observational (e.g., sensor data, data from surveys)]

i) Geolocalized JPEG images as the raw data. ii) Spectral content and lighting source classification will be derived from analyzing the raw data iii) Lamppost geographical coordinates (latitude + longitude) iv) Lamp types classified by users v)mobile phone model vi) grating model (i.e. "Edmunds Optics 1000 lines/mm linear diffraction grating") vii) Timestamp

1.2.4 What formats of data will the project generate/collect?

[Numerical - SPSS, Stata, Excel]

CSV file

1.2.6 What is the expected size of the data?

All the data, datasets and published articles generated from this action is estimated to consist of a maximum of 10-20GB of data.

Subjects	Volume	Data Source	Data Capture Tool	File Type	Format	Storage space
Humans	10	EPD (HiX)	Excel	Quantitative	.xlsx	0-10 GB
Human	10	MRI scans (PACS)	Research Imaging Architecture	Images	.dicom	10-100 GB

Core requirements for DMPs

1. Data description and collection or re-use of existing data
2. **Documentation that will accompany the data and metadata**
3. Storage and backup during the research process
4. Legal and ethical requirements
5. Data sharing and long-term preservation
6. Data management responsibilities and resources

2 Documentation and metadata

Documentation:

- What documentation is needed to **enable re-use**?
  Exs: Information on the methodology used to collect the data, analytical and procedural information, etc.
- How this **information will be captured** and where it **will be recorded**?
  Exs: Database with links to each item, lab notebooks, etc.
- How the **data will be organised during the project**?
  Exs: Conventions, version control and folder structures.

Consistent, well-ordered research data will be easier to find, understand, and re-use.

2 Documentation and metadata

Metadata:

- To be FAIR, data must be accompanied with **descriptive information in the form of metadata.**
- Researchers are advised to use **community metadata standards.**
  Ex: Directory of Metadata Standards.
- Depositing data in a certified or trustworthy repository will typically involve providing information about the data according to a **metadata standard scheme.**

2 Documentation and metadata

Examples:

2.1 Indicate what documentation will accompany the data.

- For thin-section data: For each thin section a detailed description of how the data was obtained and analysed will be stored together with the results of the analysis in a linked-excel spreadsheet.
- Modelling data: Detailed documentation of the model set-up will be provided as code books
- Deformation experiments: Detailed documentation of the experimental setup and execution will be supplied for each experiment (as pdf)

2.2 Indicate which metadata will be provided to help others identify and discover the data.

By depositing the data on DataverseNL, information will be automatically be provided by a metadata standard scheme and thus the data will be searchable using common tools such as google scholar. Additionally, the data will be attributed with a DOI, which will make it easily identifyable and discoverable.

Core requirements for DMPs

1. Data description and collection or re-use of existing data
2. Documentation that will accompany the data and metadata
3. **Storage and backup during the research process**
4. Legal and ethical requirements
5. Data sharing and long-term preservation
6. Data management responsibilities and resources

3 Storage and backup

Storage and backup:

- Give preference to the use of **robust**, managed storage with **automatic backup**, such as provided by IT support services of your home institution.
- Storing data on laptops, **stand-alone hard drives**, or **external storage** devices such as **USB sticks** is **not recommended**!
- Some research institutions have **networked research drives**, which offer ample storage space and data security for most purposes.

3 Storage and backup

Protection of sensitive data:

- Consider **data protection**, particularly if your data is sensitive.
- Describe the **main risks** and how these will be managed.
- Explain how the data will be **recovered in the event of an incident.**
- Explain **who will have access** to the data during the research and h**ow access to data is controlled.**
- Explain which **institutional data protection policies are in place**.

3 Storage and backup

Examples:

4.1 Describe where you will store your data and documentation during the research.

The digital files will be stored in the secured Research Folder Structure of the UMC Utrecht. Informed consent will be stored safely in a locked cabinet in a locked room in the UMC Utrecht. A project specific procedure is in place for access to the paper dossiers. Documentation of this procedure is stored in the Research Folder Structure.

4.2 Describe your backup strategy or the automated backup strategy of your storage locations.

1. All (research) data is stored on UMC Utrecht networked drives from which backups are made automatically twice a day by the division IT (dIT).
2. During data collection, automatic backups will be made in the Electronic Data Capture Tool Castor. Upon completion of data collection, all data are exported and saved in the Research Folder Structure where they are automatically backed up by the UMC Utrecht backup system.
3. Data in Research Imaging Archive (RIA) is stored in two data centers in the UMC Utrecht that are synchronized hourly. These centers are present at different locations within the UMC Utrecht. Next to this, two snapshots are daily created from the data.

Core requirements for DMPs

1. Data description and collection or re-use of existing data
2. Documentation that will accompany the data and metadata
3. Storage and backup during the research process
4. **Legal and ethical requirements**
5. Data sharing and long-term preservation
6. Data management responsibilities and resources

4 Legal and ethical requirements

Process and/or store of personal data:

- Seek advice from **specialised support staff** (DPO or equivalent) at your institution.
- Ensure compliance with personal **data protection laws** (ex: GDPR):
  - Gain **informed consent** for preservation and/or sharing of personal data.
  - Consider **anonymisation** of personal data.
  - Consider **pseudonymisation** of personal data.
  - Consider **encryption.**
  - Explain whether there is a **managed access procedure in place for authorised users of personal data.**

TRAINING DATA STEWARDS
FOR LIFE SCIENCES

4 Legal and ethical requirements

Ownership of data and intellectual property rights:

- Explain **who will be the owner of the data**.
- Indicate **whether intellectual property rights are affected**.
- Indicate whether there are **any restrictions on the re-use of third-party data.**

4 Legal and ethical requirements

Examples:

3.1 Describe which personal data you are collecting and why you need them.

Which personal data?	Why?
Patient characteristics (year of birth, sex)	To describe our study population
Medical history (history of diseases, family history date start of symptoms, specifics about referral, data send with the referral)	To specify the patients background
Hospital diagnostics (physical exam, lab resuls, radiology scans/reports, tumor pathogenicity, genetic tumor mutation, type of intervention and its details, complications)	To analyse the different diagnostic tools and treatments.
Follow-up (development of other paragangliomas/tumors or recurrence and its treatment, symtoms of enhanced lab results, treatment of these results, radiology scans, hospital visits, symptoms/physical exam/additional diagnostics and treatment during follow-up, reason lost to follow-up)	Keep monitoring the patient, follow development in diagnostics and therapy.
Questionnaires	To analyse the quality of life of the patient and keep monitoring the patient's quality of life.

4 Legal and ethical requirements

Examples:

3.2 What legal right do you have to process personal data?

- Study-specific informed consent

3.3 Describe how you manage your data to comply to the rights of study participants.

Right of Access	Research data are coded, but can be linked back to personal data, so we can generate a personal record at the moment the person requires that. This needs to be done by an authorized person.
Right of Rectification	The authorized person will give the code for which data have to be rectified.
Right of Objection	We use informed consents.
Right to be Forgotten	In the informed consent we state that the study participant can stop taking part in the research. Removal of collected data from the research database cannot be granted because this would result in a research bias. Patients can choose upon withdrawel if researchers can still use the patient's date or, that the researchers can't use this data anymore, but that the data will still be available.

Core requirements for DMPs

1. Data description and collection or re-use of existing data
2. Documentation that will accompany the data and metadata
3. Storage and backup during the research process
4. Legal and ethical requirements
5. **Data sharing and long-term preservation**
6. Data management responsibilities and resources

5 Data sharing and long-term preservation

- **Which data are of long-term value** and should be preserved?
- For **how long** will the data be available?
- **What** data will be **made available for re-use?**
- **When** will the data be **available for re-use**?

5 Data sharing and long-term preservation

- In **which repository** will the **data be archived and made available for re-use**?
- Under which **licence**?
- Strategy for **publishing the analysis software** that will be generated in the project.

5 Data sharing and long-term preservation

Examples:

7.2 Describe for how long the data and documents needed for reproducibility will be available.

Data and documentation needed to reproduce findings from this non-WMO study will be stored for at least 15 years.

7.3 Describe which archive or repository (include the link!) you will use for long-term archiving of your data and whether the repository is certified.

After finishing the project, the data package will be stored at the UMC Utrecht Research Folder Structure and is under the responsibility of the Principal Investigator of the research group. When the UMC Utrecht repository is available, the data package will be published here.

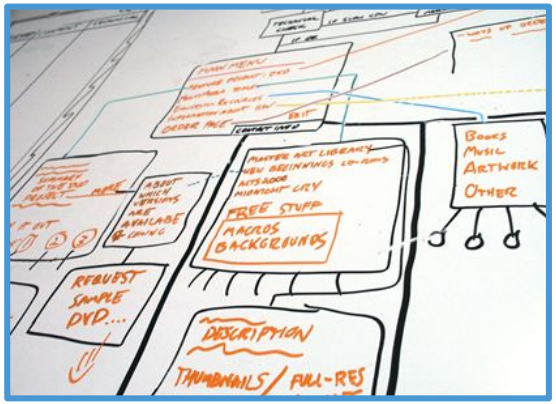
Core requirements for DMPs

1. Data description and collection or re-use of existing data
2. Documentation that will accompany the data and metadata
3. Storage and backup during the research process
4. Legal and ethical requirements
5. Data sharing and long-term preservation
6. **Data management responsibilities and resources**

6 Data management responsibilities and resources

Who?

- Outline the **roles** and **responsibilities** for data management/stewardship activities
  Exs: data capture, metadata production, data quality, storage and backup, data archiving and data sharing.
- For collaborative projects, explain the coordination of **data management responsibilities across partners.**
- Indicate who is **responsible for implementing the DMP**, and for ensuring it is **revised**, if necessary.



6 Data management responsibilities and resources

Needed resources

Describe the **resource planning** at the various levels of data management within the research project:

- Preparing data for **sharing** and **preservation**.
- Evaluation of **storage costs, hardware, HR,** resources needed to **prepare data for deposit** or to cover **any charges from data repositories.**
- **Additional resources**.

Please check:
UK Data Service - Data management costing tool and checklist
https://ukdataservice.ac.uk//app/uploads/costingtool.pdf

6 Data management responsibilities and resources

Examples:

Who will be responsible for management of the data assets during the project? Please specify their name, position, role in the project, and faculty/ institution/ group.

Name: VU at Amsterdam
Role: director of
Email: @hotmail.com
Phonenumber: 06-23925686
ORCID: yes
University: Vrije Universiteit Amsterdam
Faculty: Theology / Religion

6 Data management responsibilities and resources

Examples:

What resources will you require to deliver your plan?

It is necessary to access to the research data repository
and site of the project OSF
Moreover, during the project the following assets will be used:

- Hardware/devices: work desktop and laptop computers, personal laptop computers, USB flash drives, external disks, institutional servers, cameras, microphones, loudspeakers, smartphones, tablets.
- Software: Windows, Linux, SPSS, Excel, Word.
- Networks: entities' local networks and Wi-Fi networks with access to the Internet.
- Cloud services:
- Sites/repositories: OSF (osf.io/ugehx), research data repository INESC TEC
for only processed and anonymised data.
- Message exchanges: entities' email services, Google "GMail" emails, Skype.
- Paper transmission channels: notes related to teams, meetings, participants names, etc.

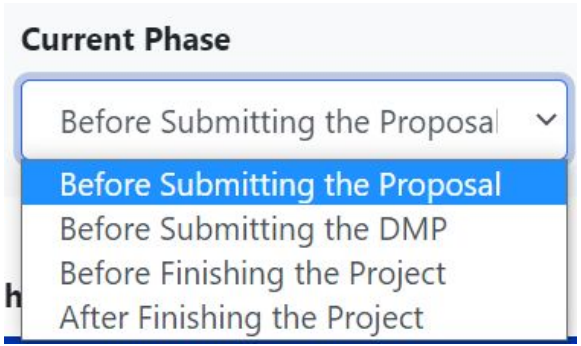
Assignment 1 Feedback

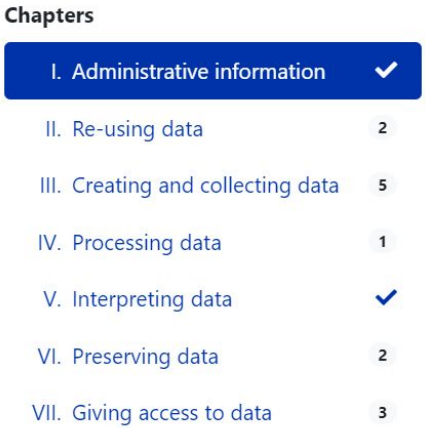
Data Management Planning

TRAINING DATA STEWARDS
FOR LIFE SCIENCES

Assignment - DMP in BioData.pt DSW

BioData.pt DSW → Science Europe DMP





TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #1

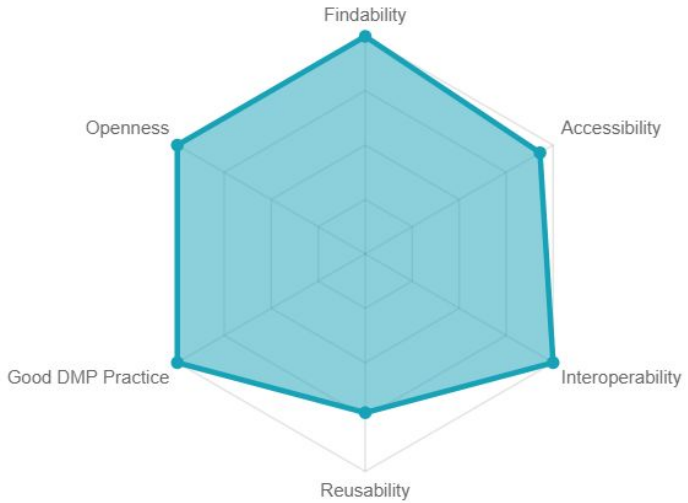
#	Researcher	Project	Phase	Model	Pre-existing data?	Creating data (format/type)
1	*Researcher Name*	*Project Title*	Before finishing the project	Science Europe	No	BMD scans and values, biometric data, exercise related data

Feedback

Strenghts	Weaknesses	Opportunities	Threats
FAIR principles followed	Could benefit from more people involved in the project, **only one person to handle multiple tasks**	Dataset **could be published**	**If results are never published** dataset will **never be openly available**
DMP well structured and well thought	Dataset **identifier not defined yet**		

DMP Inputs #1

Metric	Measure
Findability	1.00
Accessibility	0.93
Interoperability	1.00
Reusability	0.73
Good DMP Practice	1.00
Openness	1.00



TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #2

#	Researcher	Project	Phase	Model	Pre-existing data?	Creating data (format/type)
2	*Researcher Name*	*Project Title*	Before finishing the project	Science Europe	No	FASTQ Sequence and Sequence Quality Format; FASTA Sequence Format; Generic Feature Format Version 3; Rna-Seq Raw Counts.

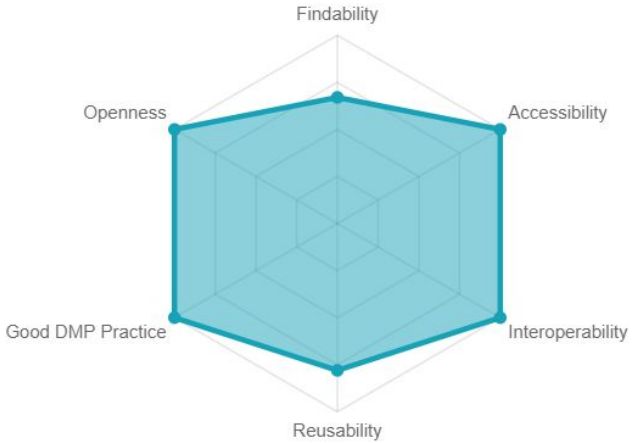
Feedback

Strenghts	Weaknesses	Opportunities	Threats
Strong support and implementation of **FAIR and CARE principles.**	No **material and methods** or **protocols** are **available to be reproduced.**	The **datasets are completely new, biologically relevant, valuable** and **unreplaceable.**	**Costs of produce and processing datasets can change**.
The **safety of the datasets** is **guaranteed during and after the project.**	The **intermediate** and **final datasets don't have description metadata contemplated**.	The **volume of data can be increased without additional costs in repository storage**. "None of the used repositories charge for their services."	Results can change the level of FAIRness of project (**Intellectual propriety with potential to the industry**).
All produced and/or collected **digital data are standardize and suitable for long term archiving.**	**No planning** regarding the **storage and processing of non-digital data** (e.g., Samples).		Implementation of the whole project **depends of the availability of samples and wet-lab experiments** (still not collected and/or performed).

TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #2

Metric	Measure
Findability	0.67
Accessibility	1.00
Interoperability	1.00
Reusability	0.78
Good DMP Practice	1.00
Openness	1.00



TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #5

#	Researcher	Project	Phase	Model	Pre-existing data?	Creating data (format/type)
5	*Researcher Name*	*Project Title*	Before submitting the proposal	Science Europe	Yes	CSV

Feedback

Strenghts	Weaknesses	Opportunities	Threats
Centralized DB that combines data from different sources	**Estimation of costs not provided**	**DB** can be integrated with **other data**	To be really useful the **DB must be kept up to date for years** (far beyond the duration of the project)
Data is already available	Findability strategy not clear (es. **metadata schemas**)	It can be a starting point for **future research projects**	
Data will be immediately **open**	The **license has not been provided**		
Integration/connection with other **well-know DBs**	**Backup strategy** not described		
Use of **CSV format**			
Funds granted			

TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #5

Metric	Measure
Findability	0.86
Accessibility	0.88
Interoperability	1.00
Reusability	0.69
Good DMP Practice	0.68
Openness	1.00

TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #6

#	Researcher	Project	Phase	Model	Pre-existing data?	Creating data (format/type)
6	*Researcher Name*	*Project Title*	Before submitting the proposal	Science Europe	Yes	Electronic document file format for long-term preservation; Tab-separated values; FASTA Sequence Format; Excel format.

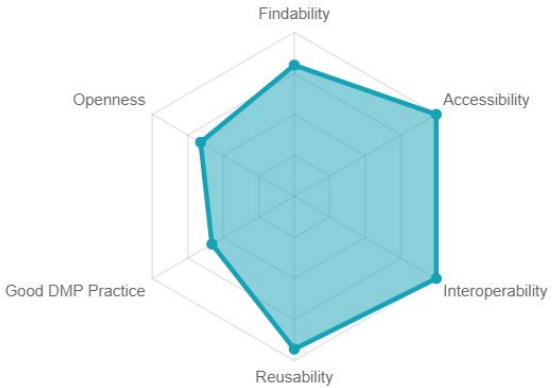
Feedback from

Strenghts	Weaknesses	Opportunities	Threats
Data **security**	No information about the **team members and their roles**	**Description of the datasets** created and/or used in the project.	The DMP is not addressing well its purpose due to **lack of some information**
Intellectual property	Not considered all data for the **space that will be required**	Definition of an **embargo period**	
Data **format and type**			

TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #6

Metric	Measure
Findability	0.80
Accessibility	1.00
Interoperability	1.00
Reusability	0.93
Good DMP Practice	0.58
Openness	0.66



TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #8

#	Researcher	Project	Phase	Model	Pre-existing data?	Creating data (format/type)
8	*Researcher Name*	*Project Title*	Before finishing the project	Horizon 2020 Horizon Europe Science Europe ma DMP	No	CSV; XTC

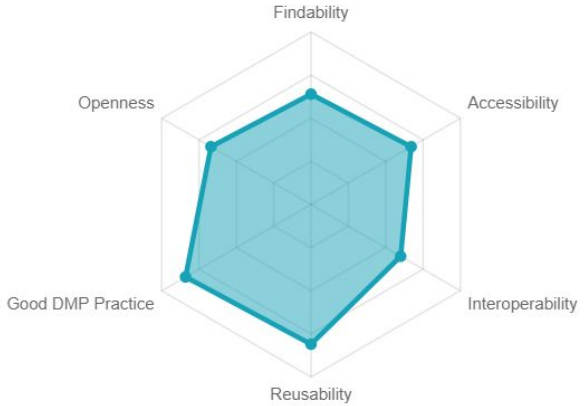
Feedback

Strenghts	Weaknesses	Opportunities	Threats
Choice of the trajectory file format used, considering the limitations of possible formats combined with storage space and portability issues known	For the experimental (wet) dataset, **only processed (treated) data will be published**, not instrument output (raw data)	For the community of biomolecular simulations **to start working on common standards for storing and sharing data**	There is **no established minimal metadata about MD data**
Data will become **completely open immediately**	The **MD dataset is in a non-standardized**, non-ascii format (though it is known to be portable...)		**MD trajectory-data is not completely interoperable**, and depending on the software producing it, require slightly different metadata description
	It is not clear whether the **input files and the topologies** will also be shared (as recommended in the RDM kit)		**The trajectory files require massive disk storage**, which poses problems in terms of backup, cold-storage, and the cost of it
	An institutional repository will be used: **no mention of reliable unique identifiers**		

TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #8

Metric	Measure
Findability	0.64
Accessibility	0.67
Interoperability	0.60
Reusability	0.81
Good DMP Practice	0.84
Openness	0.67



TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #9

#	Researcher	Project	Model	Pre-existing data?	Creating data (format/type)
9	*Researcher Name*	*Project Title*	Institution (B3iS)	No	Tables; Photos

Feedback

Strenghts	Weaknesses	Opportunities	Threats
Clear presentation of the project, administrative data and expected output data	Only gives an **example for one dataset**	**Create a general metadata template** and **state in the DMP the standards which these metadata will follow**	Data sharing strategies **might not be adjusted to the FAIR principles**
Storage costs were **estimated**	Information about **metadata is missing**		

TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #10

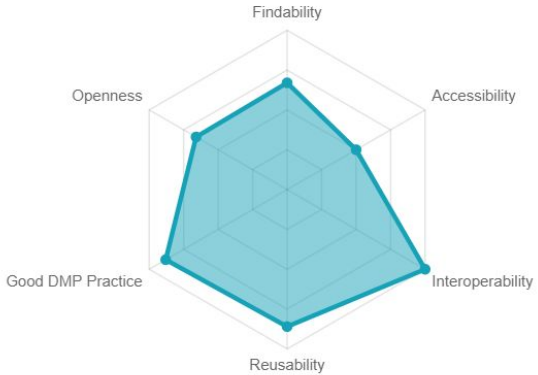
#	Researcher	Project	Phase	Model	Pre-existing data?	Creating data (format/type)
10	*Researcher Name*	*Project Title*	Before finishing the project	Science Europe	Yes	Electronic document file format for long-term preservation; FASTA Sequence Format; CSV; Mp4

Feedback

Strenghts	Weaknesses	Opportunities	Threats
Generated **types of data are explicitly split and described**	Project context is **insufficient for the uninitiated**	**Project impact is measured** – will provide feedback on its impact	Expert seminar data will be provided by multiple experts. **Metadata may be very hard to uniformize**
FAIR principles are the clear guideline for this DMP	No plans for **version-locking of the "Reference Sequence Database"**	Most (all?) important aspects of a standard DMP are covered – **the document is highly "future-proofed"**	**Mp4** is referred as a standardized format, but it **is a container format that requires more information to standardize**
	No information on **how FASTA and CSV information will be shared**		
	The **meaning of many acronyms is absent**		

DMP Inputs #10

Metric	Measure
Findability	0.67
Accessibility	0.50
Interoperability	1.00
Reusability	0.86
Good DMP Practice	0.88
Openness	0.66



TRAINING DATA STEWARDS
FOR LIFE SCIENCES

DMP Inputs #12

#	Data Curator	Project	Phase	Model	Pre-existing data?	Data format/type
12	*Researcher Name*	*Project Title*	Before submitting the proposal	Science Europe	Yes	FAIR genomes semantic model

Feedback

Strenghts	Weaknesses	Opportunities	Threats
Semantic models / ontologies / other resources are very well identified	A "Non-reference data set" has been assigned a designation but will not be used	Possibility of easily elaborating other DMPs for other projects of the institution using the same or similar tool	
Project based on FAIR principles	The data will be shared with a predefined list of people, e.g. coworkers, seems to be in conflict with other topics: - “all data will be opened immediately”, - “yes, we will be working with the philosophy 'as open as possible'	By preparing and making the DMP available, it is easier for other researchers researchers to use the data from this project	

DMP Inputs #12

Metric	Measure
Findability	1.00
Accessibility	0.75
Interoperability	1.00
Reusability	0.76
Good DMP Practice	0.91
Openness	1.00

TRAINING DATA STEWARDS
FOR LIFE SCIENCES

filipa.pereira@fccn.pt

Thank you for your attention!

Filipa Pereira		8A13-4EA4-A512		0000-0002-5732-9996

Tecnologia para o Conhecimento

TRAINING DATA STEWARDS
FOR LIFE SCIENCES