Data Management Plan

Planning

In the Data Management Plan you have to assess whether the investment to store the data balances the return on investment that the reuse of the data can give. If that assessment is positive, the data storage has to be managed. If so you have to assess whether the data are or can be made FAIR (findable, accessible, interoperable and reusable). You should consider not only the EC requirements and indications, but also your own institutes policies and recommendations. Last but not least the data management within and after the project is to be described.

Identification of RESEARCH DATASETS

The Research Data Sets are first to be identified. The idea is to take a "picture" of the research data generation. Please fill-in the following table for each dataset. There is no minimum or maximum number of datasets expected. It is advised to compile the dataset tables at WP level avoiding overlaps and ensuring information homogeneity.

DATA SET n. X– <name> – WP x – Owner(s): XXX		
1 DATA SUMMARY	**Purpose of the Data**	State the purpose of the data collection/generation, indicating the relation with the objectives of the project.
	Type and Format of data	… Describe the type of data used or generated within the project, specifying the form and format of the data: ▪ Text: field or laboratory notes, survey responses – in plain text, (txt), HTML, XLM, PDF/A … ▪ Numeric: tables, counts, measurements – in .XLSX, .CSV … ▪ Audiovisual: images, sound recordings, video – in .JPEG, .JPG, .PNG, .TIFF, AIFF, WAVE, .MP3, .MP4… ▪ For simulated data please state the model, model type and computer code… - and specify data type and format) ▪ Discipline-specific information e.g.: CIF in chemistry … (specify standard and format) ▪ Instrument specific: equipment output (specify equipment and format)
	Reused-Data	… Indicate if you re-use existing data (generated outside the project). If so, explain how.

DATA SET n. X– <name> – WP x – Owner(s): XXX		
	Data origin	… Define and describe the origin/source of your data. Data can be gathered from different sources, such as: ▪ Observational – data captured in real time- often not reproducible i.e. sensor readings, images, telemetries, sample data… ▪ Experimental – data from lab equipment, often reproducible, but with high costs - i.e. chromatograms, magnetic fields readings… ▪ Simulation – data generated by computational models where model and metadata are more important than output data - i.e. climate models, economic models, materials models,… ▪ Derived/Compiled – data coming from analyses or compilation. Reproducible but with high costs - i.e. the results of text and data mining, compiled databases ▪ Reference or Canonical – collection or conglomeration of smaller (peer-reviewed) datasets published and curated - i.e. chemical structures, gene sequence databanks, spatial data portals
	Data size	… Indicate if the dataset is: ▪ Fixed: never change after being collected or generated ▪ Growing: new data may be added, but the old data is never changed or deleted ▪ Revisable: new data may be added, and old data may be changed or deleted Then, indicate the quantity / expected size of data generated (by experiment and overall) – for example: *50MB for each experiment, overall adding up to 5GB* In case not just digital archiving is required, indicated quantities of other form of storage – for example: *2 drawers of a standard filing cabinet*
	Data Security and Storage	… Indicate how and where the data are stored and backed-up (i.e. Office computer, Hard Drive, Tape back-up system, Institute network drive, Institute Central Data storage, private Cloud storage …), briefly describing the data security policy applied.
	Data value (Long Term)	… Describe to whom the data could useful. Estimate potential value of long-term re-use of the data.

DATA SET n. X– <name> – WP x – Owner(s): XXX		
2. FAIR DATA **2.1 FAIR DATA - Making data findable**	**Discoverability of data (metadata provision)**	… Explain how data are documented and if metadata are provided, listing the information made available/discoverable. In case of materials model simulations attach a dataset specific MODA.
	Identifiability of data (refer to standard id mechanisms)	… Indicate how data are made identifiable, if a standard permanent identifier assignation scheme is used (i.e. ARK, DOI, PURL, URN, MODA …)
	Naming conventions used	… Describe the system used to name and structure electronic files and folders. Refer also to any file renaming procedure or tools used.
	Search keywords approach	… Indicate the approach to keywords generation, indexing and tagging. (For materials modelling the MODA provide this answer)
	Clear versioning approach	… Describe the versioning and traceability approach used (especially if the dataset is growing or revisable).
	Standards or procedures for metadata creation applied	… Indicate and describe the procedures and templates applied for the creation of metadata. Refer to any institute policy or recommendations by specific initiatives that are applied. In case the procedure to create metadata is not (only) manual, but automatic, refer also to any tools used for metadata creation. Some references: MODA, EMMO (European Materials Modelling Ontology), Dublin Core Metadata Initiative, DataCite Metadata Schema, Open Archives Initiative Object Reuse and Exchange, ISAtools … If there are no standards in your discipline, describe what type of metadata will be created and how.

DATA SET n. X– <name> – WP x – Owner(s): XXX		
2.2 FAIR DATA – Making data openly accessible	**Data openly available or kept close**	… Indicate ownership of the data, if it is openly available or can be made openly available. Indicate if data access is restricted, to what users, and explain the reasons.
	How data will be made available	… Indicate how you intend to make data available – i.e. through deposit in an open repository or through a platform for a specific users group or upon personal request.
	Methods or SW tools for data access	… Indicate methods and SW tools needed to access the data. Clarify if the relevant software (e.g. in open source code) is included in the data set.
	SW documentation and other information needed	… Indicate also any specific SW documentation that is needed to access the data. Indicate also any additional information that is needed to understand the data (i.e. abbreviations, supplementary notes).
	Repository for deposit of data, metadata, documentation and code	… Indicate the (open or private) repositories in which the data, metadata, documentation and code are stored and/or those in which they will be stored in the future. They might be disciplinary or institutional, open or restricted - eg. Zenodo, 4TU.Centre for Research Data, Nanomaterial Registry, Hazardous Substance Databanks… *Preference should be given to certified repositories, which support open access, where possible.*
	Access restrictions	… Indicate if there are limitations and restrictions to access the data, and if they are linked to a specific timeframe. Explain how access will be provided after these restrictions are lifted.
	Data interoperability assessment	Assess the level of interoperability of the dataset. Indicate data and metadata vocabularies, standards and methodologies followed to facilitate interoperability. Indicate if open standards are used, and (if you know) the range of utilization of proprietary SW and methodologies used to generate and manage the data.

DATA SET n. X– <name> – WP x – Owner(s): XXX		
2.3 FAIR DATA – Making data interoperable	**Standard vocabulary or mapping to commonly used ontologies**	… Refer to commonly used ontologies to map the dataset, considering also the use of existing common platforms and tools – eg: EMMO, BFO, MatONTO, Materials Ontology …
	Data licensing for wide reuse	… If applicable, define data licensing approach for the dataset wide reuse. Indicate the chosen licenses tools (eg. Creative Commons, Open Data Commons, Apache License 2.0, BSD …).
2.4 FAIR DATA – Increase data re-use (through clarifying licenses)	**Timing of data availability for re-use (incl. indications on embargo)**	… If applicable, define the timeframe for making data available for re-use. Indicate any embargo period if required.
	Data usability by Third Parties (after the end of the project)	… Indicate any limitation to the use of the data by Third Parties, after the end of the project.
	Restrictions to data re-use	… Indicate and explain any restriction to the re-use of data (i.e. confidentiality agreements, other issues).
	Quality assurance process	… Explain how quality of the data is assured, how the consistency and quality of data collection is controlled ad documented (i.e. calibration, repeat samples and measurements, standardized data capture, standardized data recording, data entry validation, peer review of data, representation with controlled vocabularies…).
	Length of time of data re-usability	… Indicate the time limit for the data re-usability, if any.

DATA SET n. X– <name> – WP x – Owner(s): XXX		
3 ALLOCATION OF RESOURCES	**Costs estimates for making data FAIR**	… Estimate the costs for making your data FAIR (*findable, accessible, interoperable and reusable*) and describe how you intend or might be able to cover these costs (i.e. institute dedicated resources, dedicated part of the project budget …).
	Data Management Responsibilities	….. Identify responsibilities for data management of this dataset (within your research group and institute, and within the project if applicable)