Inserm - Institut national de la santé et de la recherche médicale: Inserm – DMP template (english) - General informations on the data

1. INFORMATION REGARDING THE RESOURCES AND THE CONTRIBUTORS

1.1. What will be the resources (budget and time allocated) dedicated to data management to ensure that the data is FAIR: easy to Find, Accessible, Interoperable, and Reusable? *****

Recommandations:

- Explain how the resources (e.g. time) necessary to prepare the data for sharing/preservation (data curation) have been evaluated. Carefully examine and justify all of the resources necessary for disseminating the data.
- These may include storage costs, material costs, staff time, costs of preparing the data for deposit, costs of warehousing and archiving. The costs associated with data management mainly concern material resources (storage server, analysis software, warehouse storage, etc.) and human resources (e.g. hiring a data manager). At the beginning of the project, indicate the projected budget. If you do not know, you can answer this question at the end of the project.
- Indicate whether additional resources are necessary to prepare the data for deposition or to pay all costs requested by the data warehouses. If yes, specify the amount and how these costs will be covered.

Exemple de réponse:

A budget of XXX euros is considered for storing the data in an open data warehouse. The costs will be funded by the European Commission: "costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions)"

1.2. Contributors to the management of the Project data*

Recommandations:

- Describe the roles and responsibilities regarding the data management activities, e.g. data entry, metadata production, data quality, storage and backup, data archiving and sharing. Name the person(s) responsible involved, where possible.
- For collaborative projects, explain how the data management responsibilities between partners is coordinated.
- Indicate who is responsible for implementing the DMP and who ensures that it is examined and, if necessary, revised.
- Consider regular updates to the DMP.

1.3. What is the main supervisory authority of the project?

2. DESCRIPTION OF THE DATA AND COLLECTION OR REUSE OF EXISTING DATA

2.1. What is the objective of the data collection/generation? *****

Recommandations:

Explain the link between the data generated/collected and the objectives of the project

Exemple de réponse:

We will conduct experiments involving the RNA sequencing of tumor cells from patients with colorectal cancer in order to determine the molecular mechanisms responsible for the transformation of these cancers...

2.2. How many datasets/research products will you generate during this project? (In the "Dataset Specific Form" tab, you will have to specifically describe each dataset/research product) * *(the response may vary during the project)*

Recommandations:

- Research products are datasets, as well as software, workflows, and protocols produced during the project.
- A dataset is a set of raw or derived data, assembled to form a consistent whole.

One can consider that data that are managed in the same way (same methods of processing, storage, sharing, etc.) form a single dataset. The number of datasets may change during the course of the project.

2.3. How will new data be collected or produced and/or how will pre-existing data be reused? If reused, specify their origin.* *(the response may vary during the project)*

Recommandations:

- Before generating data, it is recommended to check whether it is possible to reuse datasets produced by other scientists.
- Explain what methodologies or software will be used if new data are collected or produced.
- State any restrictions on the reuse of pre-existing data. Before reusing data, you must ensure that you have the right to do so:

- By checking that they are not protected by national or international regulations, copyright, or intellectual property rights.

- By checking that the individuals are informed of the reuse of their personal data. If in doubt, you must contact XXX (define the departments that may be contacted)

- Explain how the source of the data will be documented.
- If applicable, briefly indicate the reasons why the use of existing data sources was considered but ruled out.

2.4. What data (e.g. types, formats, and volumes) will be collected or produced? * *(the response may vary during the project)*

Recommandations:

- Give details regarding the type of data: e.g. digital (databases, spreadsheets), text (documents), image, audio, video, and/or composite media.
- Detail the format of the data: how the data are encoded for storage, usually reflected by the file name extension (e.g. pdf, xls, doc, txt, or rdf).
- Justify the use of certain formats. E.g. the choice of a format can be guided by the expertise of the organization's staff, or by a preference for open formats; it can also be guided by the format standards accepted by the data warehouses, by its widespread use in a research community or by the software or equipment that will be used.
- Prefer standard and open formats as they facilitate long-term data sharing and reuse (several catalogues provide lists of these "preferred formats").
- Provide details regarding the volumes (which may be expressed as storage space required (bytes), and/or as quantities of objects, files, rows, or columns).
- Indicate the estimated volume of the data generated for the project. The volume of data generated may change during the project, give an overall estimate and mention whether this volume may change.

Exemple de réponse:

- Confocal microscopy images in .tiff and .PNG formats
- Spreadsheets in Excel format and text documents in .docx
- Flow cytometry data in .FCS format
- Akta HPLC purification profiles in .csv format
- Western blot images in .jpeg and .PNG format
- Microscopy videos in .AVI format
- Epidemiological data stored in the form of a REDCap database

3. DOCUMENTATION AND QUALITY OF THE DATA

3.1. What metadata and documentation (e.g. data collection methodology and method of organization) will accompany the data? * *(the response may vary during the project)*

Recommandations:

- Indicate what metadata will be provided to assist the research and the identification of data.
- Indicate which metadata standards will be used (e.g. DDI, TEI, EML, MARC, CMDI).
- Use scientific community standards for metadata where they exist.
  Example of response:
  The metadata are based on XXX metadata: title, author, date, contributor, description, keywords, format, type of resources, etc.
- Indicate how the data will be organized during the project, mentioning for example the naming conventions, version control, and folder structures. Data that are well classified and consistently managed will be easier to find, understand, and reuse. Specify whether a specific classification plan for the data has been implemented.
- Consider the documentation that would be necessary to enable the data to be reused. This may include information on the methodology used to collect the data, on the procedures and analysis methods used, on the definition of the variables, units of measurement, etc.
- Take into account how this information will be obtained and recorded: e.g. in a database with links to each of the files, in a "read me" text file, in file headers, in a reference book ("code book"), or in the laboratory notebooks.

Exemple de réponse:

Each unit created a classification plan at the start of the project. This classification plan is organized by data collection method (microscopy, phenotyping, sequencing, etc.), then chronologically. Raw data and processed data are stored in different folders. Below is an overview of the classification plan:

I. Collection method 1 (e.g. microscopy)

I.1 Date of first collection (e.g. 2021-01-12)

I.1.1. Processing step 1 (e.g. quality check)

I.1.2. Processing step 2 (e.g. raw data)

I.1.3. Processing step 3 (e.g. analyzed data)

I.2. Date of second collection (e.g. 2021-01-19)

II. Collection method 2 (e.g. sequencing)

3.2. What data quality control measures will be implemented? *

Recommandations:

Explain how the quality and compliance of the data collection will be monitored and documented. This involves specifying processes such as calibration, repetition of samples or measurements, standardized capture of data, validation of data entry, peer review, and representation based on controlled vocabularies.

Exemple de réponse:

To ensure data quality, various measures have been put in place:

- independent Repetition of experiments (at least three repetitions over three different days)
- Standardization of data collection (all animals raised in the same conditions, temperature control, same stimulation conditions)
- Regular review of data with the PI

4. ETHICAL AND LEGAL REQUIREMENTS, CODES OF CONDUCT

4.1. If personal data are processed, how will compliance with the provisions of the personal data and data security legislation be ensured? *

If you answer "yes" to one of the two questions below, then your project contains personal data:

- Do all or part of the data enable the identification of a human subject? (E.g. name, photo, address, etc.)

- Would all or part of the data enable the identification of a human subject if combined with other information held by you or a third party? (E.g. an identification number, a code with a correspondence table held by a third party, location data, a particular physical, physiological or genetic characteristic, elements specific to a mental, economic, cultural, or social situation, etc.)

Recommandations:

When you handle personal data, ensure that the data protection laws (e.g. GDPR) are enforced, particularly by:

- Obtaining informed consent for the preservation and/or sharing of personal data.
- Considering the anonymization of personal data for preservation and/or sharing (correctly anonymized data are no longer considered personal data).
- Considering the pseudonymization of personal data (the main difference with anonymization is that pseudonymization is reversible).
- Considering the encryption of the data, which is considered to be a particular case of pseudonymization (the encryption key must then be stored separately from the data, e.g. by a trusted third party).
- Explaining whether a specific access procedure has been put in place for users authorized to access personal data.

If your project involves personal data, you must contact the Data Protection Officer

4.2. How will other legal questions, such as ownership or intellectual property rights to the data, be addressed? What is the relevant legislation?

Recommandations:

- Explain who the owner of the data will be, i.e. who will have the right to control access to it:
- Explain the access conditions that will apply to the data. Will the data be freely accessible, or will restrictions be applied? If yes, which? Consider the use of licenses concerning access to or reuse of the data.
- Ensure that the consortium agreement covers these questions of data-access control rights for multi-partner projects and in the case of shared ownership of the data.
- Indicate whether intellectual property rights (e.g. databases directive, sui generis rights) are affected. If so, explain which ones and how this will be addressed.
- Indicate whether there are any restrictions on the reuse of data provided by third parties.

Consult Inserm Transfert regarding these questions

4.3. How will any ethical questions be taken into account, the ethical codes followed? *****

Recommandations:

- Determine whether the ethical questions will affect how the data will be stored and transferred, who will be able to see or use them, and what retention periods will be applied to them. Demonstrate that these aspects have been taken into account and planned.
- Adopt national and international codes of conduct and the institutional code of ethics and verify whether a review of practices (e.g. by an ethics committee) is required for data collection as part of the research project.

Consult the Inserm ethics /scientific integrity officer regarding these questions (Ghislaine Filliatreau)

5. STORAGE AND BACKUP DURING THE RESEARCH PROCESS

5.1. How will the data and metadata be stored and backed up throughout the research process? * *(the response may vary during the project)*

Recommandations:

- Describe where the data will be stored and backed up during the research process and how often the backup will be performed. It is recommended that the data be stored in at least two separate locations. Indicate whether your data are stored:

- On your computer,

- On a server specific to the unit,

(- On a shared storage space provided by the IT department)

Important: it is imperative that you do not store your data on an online storage space (Dropbox, Google Drive, OneDrive, etc.) as these are not secure.

- Favor the use of robust storage systems, with automatic backup, such as those provided by the IT department of the institution of origin. Storing data on laptops, external hard drives, or storage devices such as USB sticks is not recommended.

5.2. How will data security and the protection of sensitive data be ensured throughout the research process? *****

Recommandations:

- Explain how the data will be retrieved in the event of an incident.
- Explain who will have access to the data during the research process and how access to the data will be controlled, particularly in collaborative research.
- Consider the protection of the data, especially if your data are sensitive (e.g. personal data, politically sensitive information, or trade secrets). Describe the main risks and how they will be managed.
- Explain which institutional data protection policy is implemented.

Contact the Data Protection Officer

6. SHARING AND LONG-TERM STORAGE OF THE DATA

6.1. How and when will the data be shared? Are there any restrictions on data sharing or reasons to set an embargo? What is the utility of the data? *

Recommandations:

- Explain how the data can be found and shared (e.g. by depositing in a trusted data warehouse, by indexing in a catalogue, by using a secure data service, by processing data requests directly, or by the use of any other mechanism).
- Define the data preservation plan and provide information on the long-term archiving of the data. Regulatory constraints exist mainly in the case of research involving human subjects or research that uses health data. In this case, the data retention periods must be defined as soon as the protocol is defined.
- Explain when the data will be made available. Indicate the expected publication time frames. Explain whether an exclusive use of the data is claimed and, if so, for what reason and for how long. Indicate whether data sharing will be deferred or limited, e.g. for publication reasons, to protect intellectual property or patent filing.
- Indicate who will be able to use the data. If it proves necessary to restrict access for certain communities or to impose a data sharing agreement, explain how and why. Explain the measures that will be taken to overcome or minimize these restrictions.
- Indicate the utility of the data, to whom they may be useful.

6.2. How will the data to retain be chosen and for how long? *(the response may vary during the project)*

Exemple de réponse:

a) Choice:

E.g.: Datasets 1 and 2 should be retained due to the difficulties of reproducibility and the time necessary for their regeneration. Their retention is essential to ensure the reproducibility of the results presented in the publications and to be able to compare them to the data that will be generated at a later date.

b) Duration:

E.g. 1: The datasets will be retained for the maximum time allowed by the warehouse. For XXX, this corresponds to the service life of the host laboratory, which currently has an experimental program defined for at least the next 20 years.

E.g. 2: The data will be retained for an unlimited period as long as the space allocated within Inserm is available.

E.g. 3: Dataset 4 includes personal data. It will be deleted after the publication of the last article related to this project.

E.g. 4: The raw sequencing data (dataset 3) will be deleted after being uploaded to GEO or EBI, in order to gain storage space.

Recommandations:

a) Choice:

- Indicate which data must not be disclosed or which must be destroyed for contractual, legal, or regulatory reasons.
- Indicate how it will be decided which data will be kept. Describe the data that are to be preserved over the long term.

b) Duration:

- If there is a legal data retention period, cite the applicable regulations. If you consider that the data must be retained for longer than the legal period, give reasons.
- If there are no regulations but you believe that your data have long-term value, indicate how long the data will be retained.

c) Describe the foreseeable uses (and/or users) of the data within a research context.

6.3. Where will the data be preserved over the long term (e.g. data warehouse or archive)? *(the response may vary during the project)*

Recommandations:

- Indicate where the data will be deposited. Which platform or warehouse will be used to archive the datasets that are to be retained? Is this platform certified for long-term retention and management? Is this platform certified for the retention of health data?
- If no recognized warehouse is proposed, demonstrate in the data management plan that the data can be effectively managed beyond the project funding duration. It is recommended to demonstrate that the warehouse policies and deposit procedures (including metadata standards, and costs implemented) have been verified.

6.4. What methods or software tools will be necessary to access and use the data? Specify the formats chosen for archiving *(the response may vary during the project)*

Recommandations:

- Indicate whether the potential users will need specific tools to access and (re)use the data. Consider the lifetime of the software necessary to access the data.
- Indicate whether the data will be shared via a warehouse, whether access requests will be processed directly, or whether another mechanism will be used.
- Choose a format that is open and stable over time, if possible. Avoid proprietary formats or formats that depend on the technological environment.

6.5. How will the allocation of a unique and persistent identifier (like the DOI) be ensured for each dataset? *(the response may vary during the project)*

Recommandations:

- Explain how the data could be reused in other contexts. Persistent identifiers should be applied so that the data can be reliably and effectively referenced and localized. Persistent identifiers also help to count citations and reuses.
- Indicate whether it will be considered assigning a persistent identifier to the data. Typically, a trusted long-term warehouse will assign persistent identifiers. Examples of persistent identifiers: Handle, DOI (Digital Object Identifier), Ark, etc. Otherwise, indicate the URL enabling access to the dataset.

Inserm - Institut national de la santé et de la recherche médicale: Inserm – DMP template (english) - Dataset specific form

1. DESCRIPTION OF THE DATA

Description of the data

2. RENDERING THE DATA FREELY ACCESSIBLE

Will this dataset be freely available? *

- Not applicable
- No
- Yes

Location/data warehouse chosen to store and render this dataset accessible?

Will this dataset be the subject of a patent application?

Possibilities and mode of access to the dataset should restrictions apply.

What software is required to view or access the data? Do you provide documentation or open-source code for the software? *

Specify for how long the dataset will be accessible.*

3. RENDERING THE DATA FINDABLE

Is this dataset identified by a unique persistent identifier such as the Digital Object Identifier (DOI)? Otherwise, describe how this dataset and the data are identified.*

What metadata standards do you use? If you are not using a metadata standard, specify which type(s) of metadata will be created and how. *

Do you provide additional documentation to describe the data more precisely?

4. RENDERING THE DATA INTEROPERABLE

Is the data in this dataset interoperable from the technical viewpoint?

If not, what methodologies will you apply to render your data interoperable?

Specify whether you use standard vocabulary in your dataset to enable semantic interdisciplinary interoperability. If not, will you provide alignment with the most frequently used ontologies?

5. RENDERING THE DATA REUSABLE

At the end of the project, can this dataset be reused by third parties? If reuse is restricted, explain why (patent, human data, etc.) *

What license will this dataset be granted to allow for the broadest possible reuse?

Recommandations:
Examples of licenses: https://creativecommons.org/licenses/?lang=en-EN

On what date will the dataset be accessible for reuse? If applicable, specify why and for how long an embargo is necessary.

Specify for how long the dataset will be reusable. *

6. SECURITY OF THE DATA

Must this dataset be kept confidential during your project? If yes, can you specify to whom it can be disseminated? *

During the project, is this dataset stored securely? *

Does the data warehouse chosen to retain this dataset after the project implement a security policy for its information system?

What security measures are implemented for the collection and exchange of data?