Institut Pasteur: Institut Pasteur - DMP template (ENG) - General information on data

1. Overview of the data

What is the purpose of the data collection/generation?

Recommandations:

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

Exemple de réponse:

We will acquire microscope images of Yellow fever virus infections to perform Super-resolution analysis of the cellular infection. Microscopy images, videos from live infection as well as WB and virus titer images are required for total understanding and completion of the project.

How many dataset(s) will you generate during this project?

Exemple de réponse:

3 datasets

Recommandations:

A dataset is a set of raw or derived data that form a coherent whole. Data that are managed in the same way (same method of processing, storage, sharing...) can be considered to form a single dataset.

The number of datasets may change over the course of the project.

Duplicate the "Dataset" part as many times as necessary: if you have three datasets, you have to complete it three times. To know how to do it in DMP OPIDoR, please see our practical guide (part 4b in particular)

What is the nature and format of generated/collected data?

Exemple de réponse:

- Microscopy images from confocal microscopy in .tiff and .PNG formats
- Spreadsheets in Excel format as well as .docx text documents.
- Flow cytometry data in .FCS format.
- HPLC chromatogram profiles from Akta purification in .csv format.
- WB images in .jpeg and .PNG format
- Live microscopy videos in .AVI format
- Epidemiological data stored as a REDCap database

Recommandations:

Specify here the nature of data (whether they are generated or reused) as well as their format(s): genomic, proteomic, glycomic, lipidomic, transcriptomic, molecular structure data, molecular interaction data, cellular, physiological, phenotypic, clinical, epidemiological data…

To help you, a document that shows examples of data types and formats in the biomedical field is available here.

Give the expected volume of generated data for this project

Recommandations:

At the beginning of the project, indicate the estimated volume. If unknown, you can answer this question at the end of the project.

Exemple de réponse:

3 To

Will you reuse existing data? If yes, specify their origin.

- Not applicable
- No
- Yes

Exemple de réponse:

Ex 1: No old data will be used in this project. All deliverables from the project will come from original data acquired during this period.

Ex 2: Epidemiological data from 2 cohorts in Cambodia and Papua New Guinea will be reused. These data were generated under the NIH-funded project X.

Recommandations:

Before generating data, it is recommended to check if it is not possible to reuse datasets produced by other scientists.

For help on the different sources to look for data, see our practical sheet.

Before reusing data, you have to ensure that you have the right to reuse these data:
- by verifying that they are not protected by national or international regulations, a copyright or another intellectual property right
- by verifying that people are informed of the reuse of their personal data

To check that the reuse is done in compliance with the legal framework, consult this flowchart

In case of doubt, you should contact the Clinical Research Coordination Office in case of data related to human beings (crt-guichetunique@pasteur.fr) or the Legal Affairs Department for other types of data (equipedj@pasteur.fr).

2. Resources needed for data management

What hardware resources do you need to manage your data?

Exemple de réponse:

Additional storage space will be necessary. Moreover, a database will have to be set up to manage the spectrometry data during the project.

Recommandations:

Hardware resources may be necessary for data collection, storage, analysis and transfer. For instance, storage servers, computers, tablets, phones, security screens…

See our practical sheet for a summary of the different tools and infrastructures available at the Institut Pasteur for data management, storage and sharing.

Who is in charge of data management during the research project?

Recommandations:

Depending on the complexity of your project, you can :
- either indicate the role of each person individually (as in the example)
- or indicate more generally which research team is responsible for managing which type of data.

Exemple de réponse:

- A is responsible for data collection, processing and analysis
- B is responsible for the generation of the metadata and documentation related to the data
- C is responsible for data storage
- D is responsible for data archiving and sharing

What training or support do you think is necessary to help you manage your data?

Recommandations:

Indicate if training is needed (DMP writing, metadata generation, data generation...).

Also indicate any documents or information materials that would be useful to manage your data, ensure the quality of data or data management...

Exemple de réponse:

Ex 1: No specific training will be necessary to curate and manage the data generated in this project.

Ex 2: The project manager would like legal and organizational advice on the following topics: personal data and reuse licenses. The team (5 people) will also need training on technical issues: metadata, metadata standards and archiving.

What budget do you have for managing your data? How do you intend to cover these costs?

Recommandations:

The costs related to data management mainly concern hardware resources (storage server, analysis software, storage in a repository...) and human resources (hiring a data manager for example). At the beginning of the project, indicate the estimated budget. If you don't know it, you can answer this question at the end of the project.

Exemple de réponse:

Ex 1: A budget of XXX euros has been planned for the recruitment of a data manager. The costs will be covered by the European Commission.

Ex 2: The host institute and host lab have a dedicated budget to provide storage and maintenance of the data internally.

3. Legal, ethical and security aspects

Does your project include personal data?

- No
- Not applicable
- Yes

Exemple de réponse:

Yes, the project includes personal data. We took legal steps to manage this type of data.

Recommandations:

If you answer "yes" to one of the 2 questions below, then your project includes personal data:
- Does all or part of the data identify a human person?
(Ex: name, photo, address ...)
- Could all or part of the data identify a human person if they were associated with other information held by you or a third party?
(Ex: an identifier number, a code with a table of correspondence held by a third party, location data, a particular physical, physiological or genetic characteristic, elements specific to a psychic, economic, cultural or social situation ....)

If your project includes personal data, you must contact the Data Protection Officer (dpo@pasteur.fr) or the Clinical Research Coordination Office (crt-guichetunique@pasteur.fr).

You do not have to detail the steps taken in the answer of this question. Simply indicate that the measures to manage this type of data have been completed. If relevant, include references to ethics deliverables and ethics chapter of your research project.

Does your project include other data subject to a contractual, regulatory or legal obligation? If so, what type?

- Yes
- No
- Not applicable

Exemple de réponse:

Ex 1: The project includes data related to a contract with an industrial company, and that cannot be made freely accessible.

Ex 2: A patent application will be considered during the project. The data cannot be disseminated before the patent application is filed.

Recommandations:

To help you determine whether your project includes data subject to a regulatory, contractual or legal obligation, see our two flowcharts:

- Flowchart - Legal issues related to the reuse of research data
- Flowchart – Legal issues related to research data dissemination

Is there any data that should be kept confidential during your project? If so, please specify what types of data are concerned and to whom it can be made accessible

Exemple de réponse:

Ex 1: Raw data are personal data that must remain confidential and only accessible to project researchers at the Institut Pasteur. Pseudonymized intermediate results will be more widely accessible to project researchers (multi-partner project).

Ex 2: The data related to XXX will be the subject of a patent application. These data will remain confidential until the patent application is filed. They will be accessible only to the members of the project.

Recommandations:

Specify in particular if certain types of data must remain confidential because they are subject to a contractual, regulatory or legal obligation. Indicate if the access to these data must be restricted (to the members of the Institut Pasteur and Orex, to the researchers of the project, to certain people of the project...).

In particular, if you plan to protect an invention and file a patent application, the data must remain confidential before the filing of the patent application. For any additional information, do not hesitate to contact the Patent and Inventions Department: sbi@pasteur.fr

What security measures are implemented for data storage during the project?

Exemple de réponse:

Appropriate security measures will be implemented for clinical data: data stored on secure internal servers at the Institut Pasteur during the project, in a folder accessible only to project researchers.

Recommandations:

The measures to be implemented to secure the data are dependent on the level of sensitivity of the data. For any help on this subject, do not hesitate to contact rssi@pasteur.fr.

To determine the security measures to be implemented, refer to the data classification guideline (only in French)

What security measures are implemented for data collection and exchange?

Exemple de réponse:

Ex 1: Data exchange is done via Drive Pasteur, a secure and encrypted space for storing and sharing files.

Ex 2: Our project does not include data collection or exchange. No security measures are necessary.

Recommandations:

See this document (in French) for a summary of the means of data transfer to and from the Institut Pasteur's IT system

4. Data management during the project

What is the storage location of your data during the project?

Exemple de réponse:

Ex 1: As the project only includes partners from the Institut Pasteur Paris, a common project folder has been created on the storage servers of the Institut Pasteur (Gaïa project space).

Ex 2: Each partner stores the data it produces on its own servers. A Drive Pasteur folder has been created to allow data exchange between partners.

Recommandations:

Indicate if your data are stored on:
- your computer
- a server from your research unit
- a shared storage space provided by IT

Be careful, you absolutely must not store your data on a storage space on the internet (e.g., Dropbox, Google Drive, OneDrive) because these spaces are not secure.

Do you use a file classification scheme to manage your data files? Briefly indicate how it is organized.

- No
- Not applicable
- Yes

Exemple de réponse:

Yes, a file classification scheme has been created for the common storage space of all project partners. It is organized by data collection method (microscopy, phenotyping, sequencing...) and then chronologically. Raw data and processed data are stored in different folders. Below is an overview of the classification scheme:
I. Collection method 1 (e.g. microscopy)
I.1 Date of first collection (e.g. 2021-01-12)
I.1.1. Processing step 1 (ex: 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)

Recommandations:

The filing classification scheme refers to the folder tree structure set up to classify data files. The purpose of this organization is to allow any collaborator to easily and quickly locate and retrieve the data he or she needs.

See our practical sheet to help you set up a filing classification scheme.

For research projects on human subjects, we advise you to follow the file classification scheme provided by the Clinical Research Coordination Office. To use this file plan, please contact crt-guichetunique@pasteur.fr

What naming conventions do you use for your data? What rules do you use for clear versioning?

Exemple de réponse:

Each file is named with: date and name of the gene/protein. All info is separated by underscore. Processed and analyzed files will follow date of analysis plus version of the document. Text documents follow date and version as well as authors initials throughout revision.

Recommandations:

See the page written by the Archives division

The rules for a good name are:
- a short name: 30/40 characters maximum
- a meaningful name: subject_doctype_date_version
- an interoperable name: no space (underscore only), no punctuation, no special character, date written as follows: AAAAMMJJ

What measures are in place to ensure the quality of the data?

Recommandations:

Explain how the quality of data collection will be monitored and documented. This includes processes such as calibration, repeat samples or measurements, standardized data capture, data entry validation or peer review.

Exemple de réponse:

In order to guarantee the quality of the data, various measures have been implemented:
- Independent repetition of the experiments (minimum of three repetitions on three different days)
- Standardization of data collection (all animals raised under the same conditions, temperature control, same stimulation conditions)
- Regular review of data with PI

5. Data selection and long term preservation

Are your data subject to preservation regulations? If yes, which ones?

- Not applicable
- Yes
- No

Exemple de réponse:

Ex 1: No specific regulatory constraints

Ex 2: The preservation constraints were defined at the time of protocol design. Data including personal data will be deleted after the publication of the last article related to this project.

Recommandations:

Regulatory constraints essentially exist in the case of research on human subjects or using health data. In this case, data retention periods must be defined at the time of protocol design.

To know the preservation constraints of data from human research, see our fact sheet.

Which datasets are of long-term value and should be preserved? What are the datasets to destroy?

Exemple de réponse:

Ex1: Datasets 1 and 2 should be preserved due to difficulties in reproducibility and time consuming in regenerating them. Their preservation is essential to ensure the reproducibility of the results presented in publications and to be able to compare them with data that will be generated later.

Ex 2: Raw sequencing data (dataset 3) will be deleted after their deposit on GenBank, in order to save storage space.

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

On which platform(s) or in which repository(s) will the datasets to be preserved be archived in the long term (after the end of the project)?

Exemple de réponse:

Ex 1: By the end of the project, dataset 1 will be transferred to the ZENODO repository, which ensures sustainable archiving of the final research data.

Ex 2: Dataset 2 contains sensitive data, it cannot be made available on a repository external to the Institut Pasteur. It will therefore be preserved on Institut Pasteur servers.

Recommandations:

Non-sensitive data may be stored for the long-term on a data repository and/or on Institut Pasteur servers. Sensitive data must be preserved on Institut Pasteur's secure servers.

See our practical sheet for advice on how to organize space on the Institut Pasteur servers.

Specify the formats chosen for archiving.

Exemple de réponse:

PDF, TIFF, PNG and CSV formats will be used.

Recommandations:

Choose open and stable-over-time formats whether possible. Avoid proprietary formats or formats that depend on the technological environment.

See our practical sheet for an explanation of the difference between open and closed formats.

How long will the data be preserved?

Exemple de réponse:

Ex 1: Datasets will be retained for the maximum duration allowed by the repository. For Zenodo, this corresponds to the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least.

Ex 2: The data will be kept for an unlimited period of time as long as the space allocated within the Institut Pasteur is available.

Recommandations:

If there is a legal obligation to preserve data, you should cite the applicable regulations.
If you consider that the data should be preserved for a longer period than the legal period, you should justify it.
If there is no regulation but you think your data have a long-term value, state how long the data will be preserved.

What is the expected volume of archived data?

Exemple de réponse:

2 To

Recommandations:

At the beginning of the project, indicate the estimated volume. If unknown, you can answer this question at the end of the project.

If a long term preservation is needed, how do you intend to cover these costs?

Exemple de réponse:

The costs of long term preservation will be covered by the Institut Pasteur.

Recommandations:

Long-term preservation of data after the end of the project may incur costs (mainly storage space) that must be anticipated during the project.

Institut Pasteur: Institut Pasteur - DMP template (ENG) - Datasets

1. Data description

Title of the dataset

Who is the provider or producer of the data?

Recommandations:

This question is important in the case of collaborative projects. You must indicate here which partner will produce or provide the dataset.

Exemple de réponse:

This dataset is generated by partner 1 of the project

What are the nature and format of the data in this dataset?

Exemple de réponse:

This dataset contains epidemiological data stored in a REDCap database, Excel and CSV files.

Recommandations:

To help you, a document that shows examples of data types and formats in the biomedical field is available here.

Describe in more detail the data in this dataset

Exemple de réponse:

This dataset includes brain images (immunofluorescence and RNAscope) of mouse animal models for different deafness genes.

Describe the method of data collection and/or generation

Exemple de réponse:

Data are generated by confocal microscopy and analyzed by the ImageJ software and NeuroInfo softwares. GraphPad will be used for statistical analysis. Excel will be used for data management, graph representations and data analysis.

Recommandations:

Indicate how the data are generated or collected: machine-generated data, survey, observation, simulation, analysis... If some data are reused, i.e. not generated or collected during the project, indicate their source (other laboratory, online database...).

Describe your dataset with keywords

Exemple de réponse:

Drosophila larva, behavioral classification, sequences, competitive interactions, neuron substrates

Recommandations:

We recommend that you describe the dataset with at least 3 keywords. A precisely described dataset will be more easily found and therefore reused.

Indicate the URL or the persistent identifier to access your dataset

Recommandations:

Some data repositories assign persistent identifiers to datasets. If so, indicate that identifier here. Otherwise, specify the URL to access the dataset.

Examples of persistent identifiers: Handle, DOI (Digital Object Identifier), Ark...

Exemple de réponse:

Ex 1 : not available at this stage of the project, the identifier will be indicated upon publication of the dataset.

Ex 2 : https://doi.org/10.17867/10000105

What is the expected volume of data in this dataset?

Exemple de réponse:

1 To

2. Making data openly accessible

Will this dataset be freely accessible?

- No
- Not applicable
- Yes

Recommandations:

Indicate if the finalized dataset will be freely available to the scientific community (at the end of the project, at the time of publication of the associated article, etc.).

If this is not the case (dataset accessible but with some restrictions, dataset not accessible...), explain why in the next question.

Exemple de réponse:

Ex 1: This dataset will be made freely available at the time of pre-publication of the associated article.

Ex 2: No, this dataset will not be freely available because it contains personal data.

If this dataset cannot be freely disseminated, explain why.

Exemple de réponse:

Ex 1: This dataset contains non-anonymized and non-pseudonymized personal data. Therefore, it cannot be made freely available to the public.

Ex 2: This dataset was produced in collaboration with a private company. The contract with this company states that the data cannot be made public.

Ex 3: This dataset will be the subject of a patent application. Therefore, it cannot be made freely available to the public.

Recommandations:

Some research data can not be made public because it is data subject to a regulatory, contractual or legal obligation.

See the flowchart "Legal issues related to research data dissemination" for more information

Which data repository did you choose to store and make accessible this dataset?

Exemple de réponse:

Ex 1: This dataset will be made freely available via the PRIDE/Zenodo/ENA/... repository.

Ex 2: This dataset will be deposited on the European Genome-phenome Archive (EGA), a repository for controlled access to human data.

Ex 3: Due to legal constraints, this dataset will not be deposited on a data repository, it will be kept on the servers of the Institut Pasteur.

Recommandations:

Data repositories are the best solution for sharing data with a wide audience. In addition, some repositories allow you to control access to data: sharing data via a repository does not mean that the data will be accessible to everyone without restriction.

To help you find the repository best suited for your needs, see our practical sheet

Specify how access to this dataset will be provided in case of restriction

Exemple de réponse:

Ex 1: All requests for access to data deposited in the European Genome-phenome Archive (EGA) will be verified by the Data Access Committee of the project. For any access request, contact xxx@pasteur.fr

Ex 2: The dataset is deposited on Zenodo but with restricted access. The access conditions are specified on Zenodo: the person who wants to download the dataset must first explain how he/she intends to use it. Based on this justification, a decision will be made to grant or deny access.

Ex 3: The dataset is stored on the servers of the Institut Pasteur and is accessible upon request at xxx@pasteur.fr

Recommandations:

If the dataset is stored on a repository but not freely accessible, indicate how the dataset can be accessed: access upon request, subject to approval by a scientific committee....

If the dataset is stored on the Institut Pasteur's servers, indicate who to contact to request access to the data, what are the conditions of access, etc.

What software is necessary to read or access the data? Do you provide the documentation or the open source code of the software?

Recommandations:

Indicate which software you use to display, read or analyze the data.

If you have developed specific software, indicate where the source code is stored or how to access it.

Exemple de réponse:

Ex 1: Data access requires a software developed by our unit. To make our data accessible, we provide the open source code of this software.

Ex 2: Regularly used software: ImageJ, GraphPad

3. Making data findable

Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object Identifiers)? If not, describe how data and this dataset are identified.

- Yes
- No
- Not applicable

Recommandations:

The type of identifier depends mainly on the repository you choose to deposit your data. Most repositories assign unique identifiers to datasets, but they are not necessarily persistent.

To check if the chosen repository assigns persistent identifiers, search for the repository in Re3data and see if the 4th icon is blue. Under the 'standards' tab, the type of identifier is listed first.

Exemple de réponse:

Ex 1: Yes, this dataset is identified by a DOI provided by the Zenodo repository.

Ex 2: This dataset is identified by an Accession Number (unique but not persistent identifier) provided by the Sequence Read Archive (SRA) repository.

Ex 3: Not applicable, this dataset is not deposited on a data repository, so it does not have an identifier

Which metadata standards do you use? If you don't use metadata standards, outline what type(s) of metadata will be created and how.

Exemple de réponse:

Ex 1: We plan to deposit this dataset on the PRIDE repository, a repository using the "Minimum Information About a Proteomics Experiment (MIAPE)" metadata standard. We will therefore describe this dataset following this standard, as soon as the data are generated.

Ex 2: We have not yet determined the repository in which the data will be deposited. We will follow the metadata standard or the rules of the repository chosen to store the data.

Ex 3: Given that no disciplinary standard exists in our field and that this dataset will not be deposited in a repository, we have defined our own metadata that will be collected in an Excel file and associated with the data: XXXX

Recommandations:

Indicate what metadata is associated with the data so that the data are accurately described and therefore reusable. Metadata is structured information that describes the data. It can be general (e.g. author, format, date of creation...) or more scientific (e.g. organism, sample, allele, pathology...).

For more information about metadata and metadata standards, see this document.

Is this dataset described by keywords in order to make it easily findable?

- Not applicable
- Yes
- No

Exemple de réponse:

Yes, this dataset is described with 3 keywords minimum

Do you provide a supplementary documentation in order to describe more precisely your data?

- Not applicable
- No
- Yes

Recommandations:

Documentation is text that describes the data, contextualizes it and provides all the information necessary to understand it. It may be a README file associated with each dataset, the research protocol, or any other document helpful in understanding the data.

Exemple de réponse:

Yes, a README file is associated with the dataset to contextualize it and summarize the analyses performed. This file is written by the experimenter at the time of data generation and will be shared along with the data files.

4. Making data interoperable

Are the data of this dataset technically interoperable?

- Yes
- No
- Not applicable

Exemple de réponse:

Yes, the microscopy photographs are in PNG format. Tables accompanying the photographs are in CSV format. PNG and CSV formats are open formats and therefore interoperable.

Recommandations:

Data is interoperable if it can be easily combined with other data. From a technical point of view, this essentially depends on the format in which it is saved. It is recommended to use an open, widely available format that can be used by many software programs.

See our practical sheet for an explanation of the difference between open and closed formats

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

Exemple de réponse:

Our data are in a format only readable by a software developed by our service. However, we provide the source code of the software needed to access the data (additional documentation).

Recommandations:

If your data is in a proprietary format, you can:
- if possible transform it into an open and interoperable format
- otherwise indicate in the metadata associated with the dataset the name and version of the software needed to read the data.

If you have developed a software to produce or analyze the data, it is recommended that you share it.

Specify whether you will be using standard vocabulary for your dataset, to allow semantic inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies?

Recommandations:

An ontology defines a common vocabulary for researchers who need to share information in a field. It includes machine-readable definitions of the basic concepts in the field and their relationships.

If you know the data repository in which you are going to deposit the dataset, you can find out which ontology is used to describe the data and indicate it in your answer.

Exemple de réponse:

Ex 1: As our project involves medical products for human use, we used the MedDRA (Medical Dictionary for Regulatory Activities) to describe our data.

Ex 2: We did not use a specific ontology but we use a uniform vocabulary throughout the dataset. The vocabulary is based on commonly used terms in life sciences, specific terms are specified in the documentation associated with the dataset.

Ex 3: The dataset will be deposited in the SRA database (Sequence Read Archive) which allows the alignment with several ontologies : Disease Ontology, Cell Ontology, Uberon, Experimental Factor Ontology, Cellosaurus

5. Increase data reuse

At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.

Exemple de réponse:

Ex 1: Once published, this dataset may be reused by the scientific community, with the restriction that the data may not be used for commercial purposes (a CC-BY-NC license will be associated with the dataset).

Ex 2: The dataset will be the subject of a patent application. Thus, the data cannot be reused without the agreement of the patent owner: any reuse requires a license agreement.
As an exception, the data can only be reused to experimentally verify that the patent works (research exemption).

Recommandations:

If a dataset that may be of interest to the public is made freely accessible, don't hesitate to contact the Department of Communications and Fundraising for promoting the Institut Pasteur's research.

Reuse of datasets that have been the subject of a patent application is restricted : even if the data are made public after filing the patent application, they cannot be reused by third parties without a license agreement. For any further information, please do not hesitate to contact the Patents and Inventions Department: sbi@pasteur.fr

What license will be assigned to your dataset to permit the widest reuse possible?

Recommandations:

A public copyright license is a legal instrument that allows the data owner to grant users certain rights to use the data in advance. A license may also include restrictions on use (e.g. no commercial use). For international use, we recommend Creative Commons licenses.

Check out this online tool to help you choose your Creative Commons license.

Exemple de réponse:

Ex 1: The Zenodo repository chosen for the publication of this dataset allows to choose the license to be associated to the data. Consequently, this dataset will be made freely available under the CC-BY-NC license (https://creativecommons.org/licenses/by-nc/4.0/).

Ex 2: This dataset will be deposited on the OpenNeuro repository which makes the data available under CC0 license (imposed license).

Ex 3: This dataset will not be made freely available, therefore no license is associated with it. Access to the dataset will be provided upon request and the conditions of use will be defined at the time of sharing.

When will the dataset be available for reuse? If applicable, specify why and for what period an embargo is needed.

Exemple de réponse:

Data will be available for reuse within 1 year, after manuscript submission and publication.

Recommandations:

You can choose not to allow the reuse of your data for a certain period of time (embargo). For example, if you want to file a patent or if you want to conduct further research with these data.

Specify how long the dataset will remain reusable

Exemple de réponse:

Data will be stored during x years in a repository allowing their reuse.