

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

1. Information on the Data Management Plan (DMP)

Author of the DMP

Exemple de réponse:

First name, Last name, e-mail

Date of the first version of the DMP

Exemple de réponse:

February 1, 2017

Current version of the DMP and date

Exemple de réponse:

V2, July 21, 2017

Location of storage of the DMP

Exemple de réponse:

Initial and intermediate versions will be stored on an internal server and published on the project website. The final version will be stored on ZENODO and published on the project website.

2. Information on the project

Coordinator of the project

Exemple de réponse:

First name, Last name, e-mail

Organization and unit of the coordinator

Exemple de réponse:

Institut Pasteur, Bacteria-cells interaction unit

Start date of the project

Exemple de réponse:

02/01/2017

End date of the project

Exemple de réponse:

01/31/2021

3. Overview of the data

What is the purpose of the data collection/generation?

Exemple de réponse:

We will sequence the complete genome of 24 strains of *Mucor circinelloides* in order to try to understand the transmission of this species in the burn treatment department.

Recommandations:

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

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

Exemple de réponse:

6 datasets

Recommandations:

A dataset can be described as the aggregation of raw or derived data that show a certain "unity" and form a coherent whole. The number of datasets can be filled at the end of the project.

By using DMP OPIDOR, you will not be allowed to duplicate the "dataset" section to describe separately each of your datasets. If you want to do this, we advise you to use the DMP template in [Redcap](#).

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

Exemple de réponse:

Generated data will mainly consist of images (confocal microscopy and electron microscopy) in png format.

Recommandations:

Specify here the nature of data (whether they are generated or reused): genetic sequences, phylogenetic trees, diffraction data, audio files, images, tabular data, web data...

To help you, a document that shows examples of data types that can be generated at the Institut Pasteur is available [here](#).

Give the expected volume of generated data for this project

Exemple de réponse:

3 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.

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

Recommandations:

Before reusing data, you have to contact:

- the Center for translational Science in case of data related to human beings: crt-opendesk@pasteur.fr
- the Legal Affairs Department for other types of data

Indeed, 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

Underline the potential reuse of data: who will it be useful to?

Recommandations:

In case of research involving human beings, this question must be included in the protocol. If needed, don't hesitate to contact your CRT project manager or the Open Desk of the CRT: crt-opendesk@pasteur.fr

4. 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...

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

Exemple de réponse:

In the research team X (repeat for each research team):

- 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?

Exemple de réponse:

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.

Recommandations:

Indicate if training is needed (DMP writing, metadata generation, data generation...)
How many days? How many people?

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

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

Exemple de réponse:

A budget of XXX euros is provided for the storage of data in an open data repository. The costs will be covered 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)"

Recommandations:

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

5. Legal and ethical aspects

Does your project include personal data?

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 Center for Translational Science (crt-opendesk@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?

Exemple de réponse:

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

Recommandations:

Below is a list of data that your project may include:

- top secret classified data
- data that could be detrimental to the nation's scientific and technological potential
- data protected by professional secrecy
- data that will be the subject of a patent application
- personal data
- data that come from a significant database extraction
- data coming from statistical services, like INSEE
- data protected by a copyright
- data related to a contract with an industrial company

For more details and to know what to do if your project includes data subject to a contractual, regulatory or legal obligation, see [this flowchart](#)

6. Data management during the project

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

Exemple de réponse:

Data are stored on a shared storage space provided by the IT Department.

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 filing scheme to manage your data files?

Exemple de réponse:

Yes, each unit created a classification plan at the beginning of the project. This chronological classification follows the different stages of the project. Raw data and processed data will be stored in different files.

Recommandations:

For research projects on human subjects, we advise you to follow the filing scheme provided by the CRT. To use this filing scheme, please contact crt-clincore@pasteur.fr
See the [document written by the Archives division](#)

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

Exemple de réponse:

Each file is named as follows: topic_doctype_team_date_version.

The different versions are named as follows: V01, V02, DV (draft version), FV (final version)

Recommandations:

See the [document 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:

To know what measures to take, contact qualite@pasteur.fr

How do you make sure that the data in your research project are properly managed? How often do you check it and what do you do in case of discrepancy?

Exemple de réponse:

At the beginning of the project, then regularly during the project (every 6 months), the project coordinator checks that the answers to the checklist provided by the Institut Pasteur (attached) are positive. In case of negative answers, improvement actions will be implemented quickly.

Recommandations:

A checklist is available [here](#) to verify that your project data are properly managed.

For any advice or additional information, you can contact the quality department: qualite@pasteur.fr

7. Data selection and long term preservation

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

Recommandations:

To know the regulation, contact the Legal Affairs Department.

For research on human subjects or using health data, contact the CRT (crt-opendesk@pasteur.fr).

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

Exemple de réponse:

These data are not reproducible and should be preserved.

Recommandations:

To help you answer all the questions related to data archiving, see [the document written by CeRIS](#).

Are datasets that need to be preserved for the long term archived in a data repository certified for

long-term preservation and management? Which one?

Exemple de réponse:

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

Recommandations:

Please note that the data in question here are the final data that can be put online. During the research, you should preferably store your data in the secure tools provided by the IT Department.

In France, CINES is the only digital archiving platform for Higher Education and Research, but there are other international archiving platforms. To archive your data on a platform external to the Institut Pasteur, it will be necessary to establish a contract.

Specify the formats chosen for archiving.

Exemple de réponse:

XML, CSV, PDF/A, RDF, etc...

Recommandations:

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

See the [document written by the Archives division](#)

How long will the data be preserved?

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, you should indicate it.

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.

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

1. Data description

ID and name of the dataset

Which repository did you chose to store the data of your dataset and make them accessible?

Exemple de réponse:

Internal repository, external repository (GenBank, RCSB Protein Data Bank, Zenodo...)

Recommandations:

H2020 projects holders must deposit their research data (those needed to validate the results presented in scientific publications) in a research data repository.

The Institut Pasteur is currently developing a data repository called Padawan (Pasteur Data Warehouse solution). If you want to deposit data in this repository, contact the IT department: informatique@pasteur.fr

If the internal repository is not suitable, CeRIS provides you a [document](#) to help you in your choice.

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

Exemple de réponse:

Tabulated data in CSV format, pictures in PNG format, Neuroimaging data in NIfTI format, molecular structures in FCS format, diffraction data in ASCII format.

Recommandations:

To help you, a document is available [here](#) to show you examples of data types that can be generated at the Institut Pasteur.

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

See the [document written by the Archives division](#)

Describe in more detail the data in this dataset

Exemple de réponse:

This dataset includes 20 biological samples (cow and sheep) that were analysed 4 times by mass spectrometry as well as control samples.

Describe the method of data collection and/or generation

Exemple de réponse:

Data are generated by a mass spectrometer and then analysed with the software MassChroQ.

Recommandations:

Indicate how the data are generated or collected: machine-generated data, survey, observation, simulation, analysis... Specify if the data are generated during the project or reused.

What methods or software tools are needed to access the data? Do you provide the documentation or the open source code of the software?

Exemple de réponse:

Access to data requires software developed by our unit. To make our data accessible, we provide the open source code of this software.

Indicate the URL or the persistent identifier to access your dataset

Recommandations:

Some data repositories and some publishers 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...

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?

Exemple de réponse:

Protein structures will all be freely available as they will be stored in RCSB Protein Data Bank under a CC0 license.

Which repository did you chose to store the data of your dataset and make them accessible?

Exemple de réponse:

Internal repository, external repository (GenBank, RCSB Protein Data Bank, Zenodo...)

Recommandations:

H2020 projects holders must deposit their research data (those needed to validate the results presented in scientific publications) in a research data repository.

The Institut Pasteur is currently developing a data repository called Padawan (Pasteur Data Warehouse solution). If you want to deposit data in this repository, contact the IT department: informatique@pasteur.fr

If the internal repository is not suitable, CeRIS provides you a [document](#) to help you in your choice.

Will this dataset be the subject of a patent application? If yes, this dataset has to be kept confidential.

Exemple de réponse:

This dataset will be the subject of a patent application. Data from this dataset will be kept confidential during 18 months after the filing of the patent application.

Recommandations:

If you plan to protect an invention and file a patent application, make sure to respect the following:

- Do not publish your data until you have checked the patentability of your invention with with the Patent and Inventions Department.
- Do not mention or deposit data on your project's website as this would compromise the patentability of your invention.
- The data will not be published until at least 18 months after filing the patent application.
- Mark your data as confidential (see the Classification of Information Directive, being validated)

For any additional information, do not hesitate to contact the Patent and Inventions Department: sbi@pasteur.fr

Declaration of Invention forms are available [here](#).

If this dataset has to be kept closed for other reasons, explain why.

Exemple de réponse:

Example #1: this dataset contains personal data that are neither anonymised nor pseudonymised. This dataset cannot, therefore, be made public.

Example #2: this dataset was produced in collaboration with a private company. The contract with this company provides that the data cannot be made public.

Recommandations:

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

To help you determine which data should not be published, see [this flowchart](#).

You can also see the [INRA guide](#), which mainly concerns public research organizations.

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

Exemple de réponse:

Data from this dataset could be detrimental to the nation's scientific and technological potential. Access to the data requires authorization from the Institut Pasteur, as the data owner. After authorization, login details will be provided to the requester to access the data.

Recommandations:

To know if there are access restrictions to your data, contact the information classification expert at rssi@pasteur.fr

What methods or software tools are needed to access the data? Do you provide the documentation or the open source code of the software?

Exemple de réponse:

Access to data requires software developed by our unit To make our data accessible, we provide the open source code of this software.

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.

Exemple de réponse:

Yes, each dataset is identified by a DOI. Data themselves are not identified by a DOI but with a clear name: `topic_doctype_team_date_version`

Recommandations:

Examples of persistent identifiers: Handle system, DOI, Ark.

The choice of the persistent identifier generally depends on the repository.

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:

Metadata are based on ZENODO's metadata, including the title, creator, date, contributor, description, keywords, format, resource type, etc.

Recommandations:

Metadata is defined as the data providing information about the data. Unlike documentation, metadata is structured and readable by a machine and by humans.

When describing your data, you should give preference to your discipline's standards.

For more information about metadata and metadata standards, see [this document](#).

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

- 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?

Exemple de réponse:

Yes, a file is available for each data to sum up the analysis that was performed.

Recommandations:

The documentation is only human-readable, while the metadata must be machine-readable.

4. Making data interoperable

Are the data of this dataset interoperable?

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 recommended by the French reference document on interoperability ("Référentiel Général d'Interopérabilité").

Recommandations:

Interoperability is the ability of a system to operate with other existing or future systems without any restriction of access or implementation.

A data format is interoperable if:

- it is open
- it is accessible, widely distributed, and many software can execute it.
- there are tools that can migrate it to an other format
- it does not depend on the technological or economical environment.

For more information about interoperability and recommended formats, see the [reference document on interoperability](#).

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).

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

Exemple de réponse:

Example #1: We describe the data in the Crystallographic Information File (CIF) format. It's a standard for the archiving and distribution of crystallographic data:

<http://www.iucr.org/resources/cif>.

Example #2: Our metadata are specific to our project but we will bring them into line with Dublin Core and EML (Ecological Markup Language)

Example #3: As our project involves medical products for human use, we used the MedDRA

(Medical Dictionary for Regulatory Activities) to describe our data.

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.

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:

Example #1: Data can be reused internally (by other Institut Pasteur's units), by third-party institutes and by industrial companies.

Example #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 18 months 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?

Exemple de réponse:

Example #1: CC-BY license (<https://creativecommons.org/licenses/by/2.0/>)

Example #2: Open Data Commons Attribution License (<http://opendatacommons.org/licenses/by/{version}>)

Example #3: Etalab license

Recommandations:

To learn how to apply a licence to your research data, and which licence would be most suitable, see the [Digital Curation Center guide](#).

If necessary, do not hesitate to contact the Legal Affairs Department.

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

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.

6. Data security

Has your project been the subject of a risk analysis validated by the Chief information security officer (CISO)? If yes, ignore the 2 following questions.

- Yes
- No

(Optional) How necessary is the data availability during the project?

Exemple de réponse:

The need for availability is expressed on a four-level scale:

- Low: data access can be restored within a week
- Medium: access to data can be restored within about two days
- Strong: Data access can be restored within a day
- Critical: Data access must be restored as soon as possible

Recommandations:

The idea is to understand your need for data availability in order to provide solutions to meet your needs.

This need is expressed in relation to an incident occurring on the platform hosting the data: how long can the data be inaccessible without having a significant impact on your project?

(Optional) Does this dataset have to remain confidential during your project? If so, can you specify to whom it can be made accessible?

Exemple de réponse:

Raw data must only remain accessible to Institut Pasteur researchers. Intermediate results will only be available to researchers involved in the project (multi-partner project).

Recommandations:

Specify the data that are not subject to regulations and yet must remain confidential for the duration of the research.

Indicate whether access to some data should be restricted (to members of the Institut Pasteur and Orex, to the project's researchers, to certain people related to the project...) or if the data are public and do not need to be secured.

During the project (before storage of data in a repository), is the dataset safely stored?

Recommandations:

! - security must meet a legitimate need. If sensitive data, such as personal data, must be protected, there is no need to protect public data.

Thank you for describing all the measures taken to protect your data. Data are secure if:

- They are stored on a secure space that complies with the Institut Pasteur's IT security policy
- Access to your data is restricted to the people concerned
- Data stored on mobile devices (workstations, USB key ...) are encrypted
- You regularly make backup copies of your data. These copies should not be stored in the same room as the original.

For any questions about data security, do not hesitate to contact rsi@pasteur.fr

Has the data repository chosen to store the dataset after the project implemented a security policy regarding its information system?

Exemple de réponse:

We ensured that our data would be well secured in the repository selected. Indeed, we signed a Security Insurance Plan with the organization hosting our data.

Recommandations:

This question concerns all the projects. By default, projects whose data are stored in the DSI Datacenter are compliant. For projects whose all or part of the data are hosted externally, additional documents may be required. You have to:

- sign with the third party contractual clauses on data security
- OR sign with the third party a Security Insurance Plan
- OR make sure that the third party has an ISO27001 certification for at least 3 years
- OR make sure that the third party has implemented a security policy regarding its information system

What security measures are in place for data collection and exchange?

Exemple de réponse:

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

Recommandations:

Data exchanges are secure if:

- a secure exchange platform has been implemented
- transmitted data are encrypted (https, fts, encrypted attachment or use of pgp)

To help you answer this question, don't hesitate to contact rsi@pasteur.fr