Guidelines for Data Management Plan

within the framework of Horizon Europe: RIA, IA, CSA, EIC Pathfinder, Cofund

université PARIS-SACLAY

Guidelines for Data Management Plan

(within the framework of Horizon Europe: RIA, IA, CSA, EIC Pathfinder, Cofund, etc.)

11 May 2023
Version 1

This guide is created by the following members: **Mireille Brenel, Cédric Mercier, Zhengdong Shi, Stela Suhan, Samrit Mainali** and **Mohamed Diallo** of the service “Données et appels à projets” from the DiBISO (Direction des Bibliothèques, de l’Information et de la Science Ouverte) of Université Paris-Saclay.

It provides suggestions for drafting the pre-DMP (Data Management Plan) during a call for projects and thus, it equally serves as a DMP guide for the winners during the project.

For all questions concerning open science, research data and DMP (Data Management Plan), please contact : donnees-recherche@universite-paris-saclay.fr

Designer graphique : **Azaée Legrix**

This guide is reviewed by the members of 'Comité d'Europe' and 'Atelier de la donnée' of Université Paris-Saclay.

AgroParisTech : **Gaëlle Jaouen** (ingénieure de recherche, gaelle.jaouen@agroparistech.fr)
Eva Legras (science ouverte – bibliométrie, eva.legras@agroparistech.fr)
Elodie Popenda (Responsable du Pôle Europe, elodie.popenda@agroparistech.fr)

CNRS : **Dr. Etienne Snoeck** (Directeur de Recherche au CNRS, Direction Europe de la Recherche et Coopération Internationale (DERCI), etienne.snoeck@cnrs.fr)

ENS UP Saclay : **Virginia Branco** (virginia.branco@ens-paris-saclay.fr)
Agnes Nikiema : (agnes.nikiema@ens-cachan.fr)

Université d'Evry : **Alicia Ribeiro** (Chargée des systèmes d'information documentaire / Données de la recherche / Service des Thèses, alicia.ribeiro@univ-evry.fr)

UVSQ : **Michel De Moura** (Valorisation de la production scientifique / Données de la recherche, michel.de-moura@uvsq.fr)

Pôle Europe/UP Saclay : **Monica Henao** (Responsable du Pôle Europe, monica.henao@universite-paris-saclay.fr)

Finally, thanks also to **Pedro Santiago**, deputy director of the department 'International and European Affairs' and **Tania di Gioia**, Director of the department "Recherche et la Valorisation" (DiReV) for their valuable suggestions.

Table of contents

Introduction

What is a Data Management Plan?

« The Data Management Plan is a living summary-document that provides assistance with organising and planning all the phases of the data lifecycle. It explains, for each dataset, how it will be managed, from creation (or collection) to sharing and archiving. The Data Management Plan is based on the data lifecycle (image below), which refers to the different stages of data processing during a research project »
Introduction to data management plans of Université Paris-Saclay.

DMP Contents

The Data Management Plan serves as a document to well-organise different stages of your research project, it is thus to be developed even before the project starts. This document should be changed and updated regularly, adapting to the progress of your project, by specifying and elaborating different aspects. It defines different management methods, in order to adhere to the FAIR principles (Findable, Accessible, Interoperable and Reusable).

1. Before the project - Preliminary DMP for the grant proposals

Guidance : In this section, the question you must ask yourself is how you are going to manage data during the project. It is a kind of a ***preliminary DMP*** that you need to fill, especially to respect the FAIR principles : « findability, accessibility, interoperability and reusability ».

[Individual answer according to your project]

For example, questions in a RIA and IA application form.
Research data management and management of other research outputs.

- Research **data management and management of other research outputs:** Applicants generating/collecting data and/or other research outputs (except for publications) during the project must provide maximum 1 page on how the data/ research outputs will be managed in line with the FAIR principles (Findable, Accessible, Interoperable, Reusable), addressing the following (the description should be specific to your project): *[1 page]*

  Types of data/research outputs (e.g. experimental, observational, images, text, numerical) and their estimated size; if applicable, combination with, and provenance of, existing data.

  Findability of data/research outputs: Types of persistent and unique identifiers (e.g. digital object identifiers) and trusted repositories that will be used.

  Accessibility of data/research outputs: IPR considerations and timeline for open access (if open access not provided, explain why); provisions for access to restricted data for verification purposes.

  Interoperability of data/research outputs: Standards, formats and vocabularies for data and metadata.

  Reusability of data/research outputs: Licenses for data sharing and re-use (e.g. Creative Commons, Open Data Commons); availability of tools/software/models for data generation and validation/interpretation /re-use.

  Curation and storage/preservation costs; person/team responsible for data management and quality assurance.

  ⚠ *Proposals selected for funding under Horizon Europe will need to develop a detailed data management plan (DMP) for making their data/research outputs findable, accessible, interoperable and reusable (FAIR) as a deliverable by month 6 and revised towards the end of a project's lifetime.*

  ⚠ *For guidance on open science practices and research data management, please refer to the relevant section of the HE Programme Guide on the Funding & Tenders Portal.*

Screenshot from a document "Standard Application Form (HE RIA, IA)"

Types of data/research outputs: e.g. experimental, observational, images, text, numerical and their estimated size; if applicable, combination with, and provenance of, existing data.

Guidance : It is recommended to describe the data format(s) and type(s) used in your project (open or non-proprietary data formats are recommended). You can consult the data formats and types recommended by the UK data Service. The list of Recommended file formats from the Cornell University Library is also very helpful.

Generic text suggestion to adapt according to your research project:
Recommended open formats (.rtf for textual data, .csv for spreadsheets, .tif for images, .mpeg for digital videos, etc.) will be used for data/research outputs.

Findability of data/research outputs: Types of persistent and unique identifiers (e.g. digital object identifiers (DOI)) and trusted repositories that will be used.

Guidance : The most common persistent identifiers (PID) are: DOI, Handle, ARK, PURL, URN, ORCID, ISNI...
Generally, data repositories assign a persistent or unique identifier to datasets. For example, Zenodo assigns a DOI to each dataset for free.

Generic text suggestion to adapt according to your research project:
Data and metadata will be ensured by the use of persistent identifiers (for example, DOI, Handle, ARK, PULR, URN...) and trusted repositories.

Accessibility of data/research outputs: IPR (Intellectual Property Rights) considerations and timeline for open access (if open access is not provided, explain why); provisions for access to restricted data for verification purposes.

Guidance : The European Commission has adopted the principle « as open as possible, as closed as necessary » for data sharing (could be consulted via the official report of study about « Open Science and Intellectual Property Rights »). In France, the guide « Ouverture des données de la recherche : guide d'analyse du cadre juridique en France » also addresses the dataset IP issues.

Generic text suggestion to adapt according to your research project:
Ex 1: The data will follow the principle “as open as possible, as closed as necessary” adopted by the European Commission about data sharing. IPR issues will be taken into consideration and timeline for open access (if open access is not provided, the reason shall be well-explained) will be properly defined.

Ex 2: IPR issues and timeline for open access will be taken into consideration. Reasons will be given if open access is not provided or is restricted.

Interoperability of data/research outputs: Standards, formats and vocabularies for data and metadata.

Guidance : Data is interoperable if it can be easily combined with other data. In order to achieve this :

- 1 It is recommended to use an open and non - proprietary format for the reuse of data. (You can also find this information in the section «Types of data/research outputs »)

-2 If possible, use a list of common vocabularies within the scientific community such as standard vocabularies, ontologies or thesaurus. Metadata is the information that describes the data. They can have either generic standards (Dublin core, MARC...) or disciplinary (DDI, EML, Darwin Core...).

Helpful resources about the metadata standards:
http://rd-alliance.github.io/metadata-directory/standards/
https://www.dcc.ac.uk/guidance/standards/metadata
https://doranum.fr/metadonnees-standards-formats/

Generic text suggestion to adapt according to your research project:
In order to ensure easy combination with other data, open and non-proprietary format and common vocabulary within the scientific community will be used. Metadata will be machine-readable and standardised (Dublin Core, DataCite, DDI, etc)

Reusability of data/research outputs: Licenses for data sharing and re-use (e.g. Creative Commons, Open Data Commons); availability of tools/software/models for data generation and validation/interpretation/re-use.

Guidance : A license provides the legal condition to share and re-use your data. Helpful tool to choose a licence: License Selector (ufal.github.io)

Generic text suggestion to adapt according to your research project:
Data and research outputs will be made freely available under the latest available version of the Creative Commons Attribution International License (CC BY); Metadata on the other hand, will be attributed with Creative Commons Public Domain Dedication (CC0) (or a licence with equivalent rights).

Curation and storage/preservation costs : person/team responsible for data management and quality assurance.

Guidance:
It is important to take into consideration and justify the resources required to implement the plan (for example, storage costs, equipment, staff time, data documentation, data curation, data preservation, etc).
Below are some of the helpful tools:
https://www.ukdataservice.ac.uk/manage-data/plan/costing).
how-to-comply-to-h2020-mandates-rdm-costs (openaire.eu)

Generic text suggestion to adapt according to your research project:
Data curation and storage/preservation costs and person/months involvement for data management and quality assurance will be carefully defined and justified to ensure the implementation of good data practices.

2. During the project – Horizon Europe Data Management Plan Template with annotations

Once your project is selected by the European Commission, you need to submit ***the first version*** of the DMP ***by month 6*** of the project[1], ***the second version*** by ***the middle*** of your project, and ***the last version*** by the end.

The sections of the template and the questions hereafter are taken from the Horizon Europe FAIR Data Management Plan (DMP) template[2].The use of the template is recommended by the EU commission but voluntary.

General Project Information

Action Number : [*insert project reference*]
For example : 101017572.

Action Acronym : [*insert acronym*]
For example : EUGLOH

Action title : [*insert project title*]
For example : European University Alliance for Global Health

Date : [*insert date*]
For example : 21/09/2022

DMP version : [*insert DMP Version*]
For example : version 1.

1 Note: ***For ERC projects***, it is sufficient to provide ***a single version of the DMP*** as a deliverable 6 months after the start of the project, but the implementation of this plan should be monitored.
2 https://dmp.opidor.fr/public_templates?page=1&search=horizon+europe

1. Data summary

a) Will you re-use any existing data and what will you re-use it for?

State the reasons if re-use of any existing data has been considered but discarded.

[Individual answer according to your project]

Guidance: Before you decide to reproduce/regenerate data, it is recommended to verify whether the datasets, already produced by scientists elsewhere, are reusable to your work. The website stated below might be helpful to look for datasets on various sources:
https://doi.org/10.18167/coopist/0071

The other issue to address is the verification of the right to reuse data coming from third party sources.

To do so:
- Verify that there are not additional national /international regulations to follow, whether or not there is a copyright attached with a dataset or any other intellectual property right.
- Verify that in case of reuse of data, the people behind the datasets are informed of the use of their personal data (wherever applicable).

Some examples of answers:

Ex 1: Not any pre-existing data will be used in this project. We shall generate our own data that guarantee the necessary deliverables.

Ex 2: Epidemiological data from Covid-19 outbreak will be re-used in our project. A static copy of these datasets, uploaded in a 'Figshare' will be re-used. A live version of these datasets, downloadable from GitHub, shall also be deployed for our project.

b) What types and formats of data will the project generate / collect or re-use?

[Individual answer according to your project]

Guidance: What are the different data types that your scientific project will generate (if not re-used) at the end, what are the formats etc.
Data format: Keeping in mind the interoperability, the formats are to be as sustainable and open as possible.
The website below contains adequate information on open and closed formats:
https://doranum.fr/stockage-archivage/quiz-format-ouvert-ou-ferme_10_13143_mcwq-qs64/

Similarly, you could browse through the website of CINES (link here: https://facile.cines.fr/) to verify whether or not the format you choose is sustainable and could be archived for long-term use. The website below lists several formats that enable easier integration, re-use and distribution: https://www.esri.com/en-us/arcgis/products/arcgis-data-interoperability/supported-formats

Some examples of answers:
- Live images from satellite imaging in .tiff and .PNG formats
- Technical reports in .docx format
- The measurement of cell temperature variation in the form of tabular data with .csv format.

c) What is the purpose of the data generation or re-use and its relation to the objectives of the project?

[Individual answer according to your project]

Guidance: Explain the correlation between the data generated (or collected) and the objectives of the project.

Some examples of answers:
Via the Tunnelling Electron Microscopy (TEM), we will generate a new set of crystallographic structures for the molecules we are interested in. The idea at the end will be to opt for next generation molecules to store energy more efficiently in the long-run.

Via the mass spectrometry, we will acquire the knowledge of different chemical compounds supposed to exist (but not fully mastered yet) in the interstellar medium and their structure. This will help us to better understand different Polyaromatic Hydro-Carbons that are still missing in the universe.

d) What is the expected size of the data that you intend to generate or re-use?

[Individual answer according to your project]

Guidance: Indicate the expected volume of data that will be re-used (or generated) throughout the project. This is a good opportunity to search for alternatives (if necessary) to store your data in a sustainable way. The volume could be indicated in units such as MB/GB/TB. Specify equally the amount of data (generated or re-used), their access (who, how) and preservation (where, how).

Some examples of answers:
5 To.
2 MB of raw data from the previous study (specify) will be re-used.

e) What is the origin/provenance of the data, either generated or re-used?

[Individual answer according to your project]

Guidance: Specify how the data for your project are generated (simulation, machine-generated) or collected (via observation, experiments, surveys). In cases where the data are re-used, indicate the origin (author, another research lab, online databases, commercial sources etc).

Some examples of answers:
The data are obtained via computational multifocal microscopy
The data (clinical trial, for instance) are generated in a hospital (specify the name and place) while using a NMR facility

f) To whom might the data be useful ('data utility'), outside your project?

[Individual answer according to your project]

Guidance: Specify the community (stakeholders, general public etc.) who could ultimately serve your data for their own research activities.

Some examples of answers:
The clinical data obtained will be useful for researchers and scientists to better understand the rate of change of antibody levels in human beings over time once injected with the vaccine.
The data could equally serve pharmaceutical companies to have efficient vaccine that could prolong the efficiency of single-shot vaccines and reduce the necessity of using them several times over a given period.

2. FAIR data - Making data findable, including provisions for metadata

a) Will data be identified by a persistent identifier?

Guidance: Persistent and unique identifiers (PIDs) are essential to ensure the fact that your research outputs are easily findable by other people. They can be assigned both to digital objects (data, publications etc.) or non-digital ones (researchers, research organizations, grants organisations etc.)

Examples of persistent identifiers (PID) for digital objects: DOI, Handle, ARK, PURL...
Examples of persistent identifiers (PID) for non-digital objects: ORCID, ROR etc.

Data repositories (whether generalist or disciplinary) usually assign a PID to the datasets deposited on their platform. For scientists based in French research institutions, it is possible to deposit your research data in a national portal (Recherche Data Gouv) with a possible attribution of DOI. However, keep in mind that the repository is general. To find out whether the repository you have chosen provides a PID or not, you can browse through the Re3data portal (if it's the case, you will see a blue icon with a symbol).
In case where the data is stored locally (data size too large or sensitive data, for instance), it is equally possible to subscribe to PID OPIDoR to assign DOIs.

Some examples of answers:

Yes, each of the datasets will be identified by a unique identifier via a DOI provided automatically by Zenodo for dataset 1 and an accession number (generated by the ENA repository)

No, the datasets are not deposited in any of the repositories (specify the reason) and are therefore not assigned any persistent identifiers.

b) Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

[Individual answer according to your project]

Guidance:
Metadata (or the underlying data) is the information about your data generated during the project, but is not the data itself. It can be rather general (author information, format, date of creation etc.) or more scientific (research organisation, funding source, data version, codes etc.) The standardization of metadata is done according to generic standards (Dublin Core, EAD, MARC, MOCS etc.) or disciplinary ones (EML, DCC etc.) Different repositories have their own set of standards for metadata to facilitate the description of data.

Some of the metadata standards can be found in the websites below:
http://rd-alliance.github.io/metadata-directory/standards/
https://www.dcc.ac.uk/guidance/standards/metadata

Some examples of answers:
Metadata are based on the standards set by 'The Cambridge Structural Database' (handled by DCC, Data Curation Center).
The metadata standard will be defined by the chosen repository (this will be decided with the advancement of our project).

c) Will search keywords be provided in the metadata to optimise the possibility for discovery and then potential re-use?

Guidance: The best way to describe and help the findability of the datasets is via the key words chosen. It is recommended to use the controlled vocabularies to favour the visibility within the limited search results and within their domain. To learn more about this, you can refer to the following websites:
- https://fairsharing.org/standards/
- https://bartoc.org/

Some examples of answers:
Yes, each dataset is described with 3 keywords minimum.

d) Will metadata be offered in such a way that it can be harvested and indexed?

Guidance: The data deposited in a repository and described within the workable framework of metadata standards will make them machine-executable and are therefore indexed for long-term search preservation and offer easiness in searching. Note that not all metadata may be indexed or made available for searching. For example, a file with formats such as .csv or .txt, that describe your metadata, is human-readable but not indexable. Thus, it is not harvestable.

Some examples of answers:

Yes, the data will be deposited in repositories (provide the names whenever possible) that allow metadata to be harvested and indexed for long-term use.
No, the data will not be stored in any repositories. The metadata will simply be described by a .csv file. It will thus be human readable but neither harvestable nor indexable.

3. FAIR data - Making data accessible: Repository

a) Will the data be deposited in a trusted repository?

Guidance : A repository is an online-based platform where researchers can deposit their data (but not limited to) for mid (about 10-20 years) to long-term (more than 20 years, generally speaking) archival. Its primary objective is to open up and share the data. They could be institutional, general or disciplinary. You are recommended to prefer disciplinary over generalist one wherever possible.

Trustworthy digital repositories are the ones certified by several international certification standards. In the framework of European Commission, they are the following:
- Certified repositories (certification done by the standards such as CoreTrustSeal, Nestor Seal, ISO16363)
- General repositories like Zenodo, that aim for certification in the near future
- Domain based repositories that are commonly used and endorsed by the research communities.
- Institutional repositories with essential characteristics of trusted repositories.

The website below provides an overview regarding the criteria for a trusted repository within the framework of European Commission:
https://www.openaire.eu/find-trustworthy-data-repository

Please note that the websites (personal or institutional), blogging site, publisher's websites, cloud-based storage solutions (like dropbox, google drive etc.) are not considered repositories. Similarly, the platforms like

ResearchGate or Academia are also not considered as repositories

Some examples of answers:
The datasets will be deposited to the 'HEPData'.
The data are sensitive; they will thus not be deposited anywhere.

b) Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Guidance: Specify whether you have found a repository that falls within the specific security measures defined for your project. For example, whether the access can be limited among the partners of the consortium during the project before opening up the data for the public.

Examples of answers:
The partners of the consortium currently use GitHub to share the data within them. The developed code at the end of the project will be publicly available through APGL-3.0 public license.
The data not being confidential will be deposited in Zenodo with an immediate possibility for sharing and re-use.

c) Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Guidance: Most of the repositories automatically assign the unique identifiers. Some of the most commonly used are DOI, Handle etc. Other types of archival identifiers like ARK (Archival Resource Key) or SWHID (Software Heritage ID) are also often used.

Some examples of answers:
The datasets will be uploaded to Zenodo, which will then generate a DOI for each of our dataset.
The datasets will be deposited in our institutional repository with a proper ID generated automatically during submission.

4. FAIR data - Making data accessible: Data

a) Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

[Individual answer according to your project]

Guidance: The principle put forward by the European Commission regarding the data publishing is the following:
"As open as possible, as closed as necessary": The data should be open by default; however, there may be legitimate reasons which could highlight why the data publishing is not possible (this has to be justified through some solid reasons like legal, ethical or contractual obligations, for instance).

More information on this issue is provided in the following website (under the section 'Limited exceptions to these guidelines'):
https://open-research-europe.ec.europa.eu/for-authors/data-guidelines#addstatement

Some examples of answers:
The generated protein structures will be freely stored and made publicly available in RSCB Protein Data Bank under CC0 license.
The datasets are subject to patent. They will thus be confidential before the patent is granted.
The datasets obtained are clinical and will not be anonymized or pseudonymized. They will therefore not be published.
The dataset 4 is a joint collaboration with an industrial partner, not bound to comply with the open access principles. The data therefore will not be made public.

b) If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

[Individual answer according to your project]

Guidance: The notion of embargo is to be considered when different options of "early sharing" are not immediately implementable. However, note that once the publication(s) related to your work is (are) out in the market, you are obliged to share openly your data as well (see Horizon Europe Programme guide for more details). If a patent is thought of for your project, it should be clearly specified and the data could thus remain closed until the patent has been accepted.

Some examples of answers:
We envisage keeping data within our reach until the patent has been filled.
The data will be made immediately available along with the publication.

c) Will the data be accessible through a free and standardized access protocol?

[Individual answer according to your project]

Guidance: Specify how one could access the data that are deposited in a trusted repository. Are there any additional requirements for the users to download data?

Some examples of answers:
The datasets will be stored in the French national repository Recherche Data Gouv. The access to these data will be guaranteed without the need of creating a user account. The detailed and complete metadata for each dataset deposited in 'Recherche Data Gouv' will be made publicly available with the principle of Open Access.

d) If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

[Individual answer according to your project]

Guidance: Describe in detail how the access will be handled, who will be the person in charge of managing the access, by what means the data will be limited within the collaborators during the project.

Examples of answers:
Not applicable.
The data will be deposited on 'Zenodo'. However, the access will be restricted. One who wishes to re-use/download the datasets should contact the person in charge (provide the name) with the intention of data handling. Whether the access is to be granted or not depends on the way how the third party wants to use our data.

e) How will the identity of the person accessing the data be asertained?

[Individual answer according to your project]

Guidance: State in this part how the identity of the person wishing to access the data will be verified.

Ex1: The data will be made accessible via the open access policy, wherever possible. Regarding the cases where this is not possible, the readers will be offered a possibility of requesting the access to the data owner via the email address provided along with the metadata.

Ex2: The data during the project will be limited within the members of the consortium; the access can be guaranted via the institutional login during this period. Once the project achieves its end, the data will then be accessible to the public.

f) Is there a need for a data access committee (e.g. to evaluate/ approve access requests to personal/sensitive data)?

Guidance:
Specify how the request from a third party to use personal/sensitive data will be treated.

Contact your institution's legal office to check if this is necessary.

5. FAIR data - Making data accessible: Metadata

a) Will metadata be made openly available and licensed under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

[Individual answer according to your project]

Guidance: Licensing policy: Mention clearly how the research data will be made accessible, state the licensing and attribution terms as required by the funding agency. The easiest way is to put your data under the latest version of Creative Commons Attribution International Public License (CC BY) or equivalent depending on your need(s). Below is an official text from the French government regarding the use of licenses.

https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000034502557

Metadata requirements:
Keep in mind that the metadata is to be openly accessible under the licensing terms guaranteed by CC0 or equivalent.

The metadata should correspond to what is required by the FAIR principles. While interoperability is to be guaranteed, keep in mind to follow the standardized format, make them machine-actionable and provide as much information (for instance, date of dataset deposited, version of dataset deposited, information on the authors, embargo, if any, funding agencies, name of the project, acronym, licensing terms etc.) as possible to increase the visibility of your research outputs.

In case where the dataset itself is closed but without any compelling reasons regarding the closeness for the data, it is recommended that the metadata be open with licenses such as CC0 or equivalent (see AMGA, Annexe 5).

Some examples of answers:

Yes, the chosen repository for our datasets offers the possibility of making the metadata publicly available, under CC0 license.
No, since the data are not deposited in a repository, it will not be possible to make the associated metadata openly available.

b) How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

[Individual answer according to your project]

Guidance: Maintaining the datasets in a given server for a long-term conservation (more than 10 years) is a costly process and has to be justified. In this case, it is therefore necessary to look for alternative solutions that provide sustainability over a longer period.

Some examples of answers:
Ex 1: Data will be deposited in our institutional repository (give the name) for a guaranteed limit of five years. They will undergo a thorough revision afterwards to decide whether they are still pertinent to be present on the server.
Ex 2: The data will be archived in a certified platform of CINES for the long run.

c) Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)? [Individual answer according to your project]

Guidance: State clearly what software will be used to read, analyze and visualize the data. In case where your work leads to the development of a specific software/code, specify its location and conditions to access.

In general, it is recommended to create a 'Readme' file (https://recherche.data.gouv.fr/fr/actualite/bien-decrire-ses-donnees-pour-les-valoriser-un-modele-de-readme-a-votre-disposition) to provide valuable information such as authors, licensing terms, acknowledgements regarding the previous and current contributors etc.

Some examples of answers:

Our unit has developed a specific software to visualize the data. The open-source code of this software will be made accessible through GitHub.

6. FAIR data - Making data interoperable

a) What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

[Individual answer according to your project]

Guidance: Interoperability refers to the fact that the data coming from various sources can be integrated easily without losing the essence. Technically speaking, it is maintained by using the same formats for all the files to be integrated. You can see the list of Recommended file formats from the Cornell University Library to familiarize yourself with different formats.

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. Wherever possible, it is suggested to put your data in a non-proprietary format to facilitate the interoperability. When a proprietary format is used and the conversion is not possible, specify which software to use to read the data. In the link below, you could see whether the format you have chosen is open or not (https://doranum.fr/stockage-archivage/quiz-format-ouvert-ou-ferme/)

From a semantic point of view, a dataset is considered to be interoperable if the same vocabularies are used between the different scientific commu-

nities. Once the repository is decided, for the sake of visibility, you can use standard vocabularies or ontologies to describe the data.

Some examples of answers:

The data are readable only by the software developed by our team. The source code of this software is freely available in GitHub (provide the link).

We have our datasets described in text format. This is openly available.

The microscopy images are in PNG format. The tabular data associated with these images are in CSV format. Both of these are open and therefore easily interoperable.

We used MedDRA (Medical Dictionary for Regulatory Activities) to describe the produced data.

b) In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

[Individual answer according to your project]

Guidance: The question is very precise and pin-point. It is possible that it may be inconsistent with your project.

Some examples of answers:

The sets of vocabularies were generated through a software: TyDi (Terminology Design Interface), installed in our data archival platform.
The vocabularies used are common in life sciences and the same terminologies are used throughout the different datasets produced.

c) Will your data include qualified references to other data ? (e.g. other data from your project, or datasets from previous research)

[Individual answer according to your project]

Guidance: Indicate, whether within the framework of your project, you will use other datasets that were previously produced elsewhere (this is why it is important to assign a unique identifier to your datasets that facilitate the findability of what you are searching for). It is possible to include additional links (publications, documentation etc.) to your metadata in a given repository

Some examples of answers:

A link to the associated datasets will be provided, and linked with our publications.

7. FAIR data - Increase data re-use

a) How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

[Individual answer according to your project]

Guidance: The documentation of your work is extremely important as it is the way to contextualize and describe the data. The documentation provides complete information about your data. This could be done for instance via a 'readme' file for each dataset or protocol produced (or used) during the research. A data dictionary describing the variables, units, etc. can also be considered. It is important to note that good management practices (adopting a convention when it comes to name a file, for instance) are indeed essential for documentation of datasets. Other sets of rules may include homogeneous practices when it comes to archiving and data sharing. Some of the helpful resources are listed below:

4TU: Guidelines for creating a README file
Doranum : "Comment bien nommer ses fichiers ?"

Some examples of answers:

For each of the datasets produced, we will provide the readers with an additional 'readme' file detailing all the analyses. Our publications and the datasets will be linked with each other to further enhance the complete reading and better understanding for the readers.

b) Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

[Individual answer according to your project]

Guidance: Licensing is a legal way to allow the third party (the readers, for instance) certain rights in advance so that they can use the data without any problems. The commonly used one is Creative Commons (CC) licenses that allow the data re-use and sharing with proper citation credits for the authors. It is also possible to add even more restrictions on a chosen license (NC, for Non-Commercial use, for instance). As per the regulation set by the European Commission, the data and the associated research products are to be made publicly available under the latest version of CC BY licences or any other licences with equivalent rights.

In France, Etalab, a specific open licence from the government, provides more insight on different modalities of re-use while guaranteeing the authorship of one's work (this license however, is close to CC-BY than to CC0).

Helpful resource to choose a license: Public license selector

Some examples of answers:

The crystallographic images will be freely available under CC-BY-NC license. This is to prevent the non-commercial use of our results.

The protein structure deposited in RSCB Protein Data Bank will be available under CC0 license.

c) Will the data produced in the project be usable by third par-

ties, in particular after the end of the project?

[Individual answer according to your project]

Guidance: Specify if the data generated will be kept under some restrictions regarding the opening (this is usually the case when one has clinical data, data coming from third party commercial source, question of intellectual property etc.) and explain why, if the case prevails.

Specificity for patent data:

The datasets that are subject to patents are strictly prohibited for re-use, even after they are made public. The third party can however, use those data with proper licensing agreement.

Some examples of answers:

The datasets will be made available under CC-BY-NC license. Once public, the third party are free to use them (with proper citation and absence of commercial use).

The data will be submitted for a patent. However, they can be provided to the third party through specific licensing agreements once the patent has been filed. Meanwhile, these data can be re-used to prove that the patent works (research exemption).

d) Will the provenance of the data be thoroughly documented using the appropriate standards?

[Individual answer according to your project]

Guidance: Provenance of data means the source: who generated (or collected) it? Whether it has been published elsewhere? How has the processing been done? Does it contain data from the third party that was transformed or completed? For the data deposited in a repository, the provenance is to be documented during the time of deposit. It is also important to know the identifier(s) like DOI for the dataset(s) that were re-used in your work.
Some examples of answers:

The origin of data and the associated metadata will be succinctly documented so as to address the questions such as who (when, how) collected the data, how were they transformed to a usable form for your work (if it was not case) etc.

e) Describe all relevant data quality assurance processes.

[Individual answer according to your project]

Guidance: In this section, you are required to state clearly how the monitoring and documentation of data collection will be done. It could be, for instance, via experimental observations, repeated calibration and measures, data entry validation, peer review, and so on. Some of the projects may require quality standards (ISO) which should also be cited.

The other thing that could be presented are the control procedures put in place in order to control and verify the results obtained, regular use of laboratory notebooks to note down all the activities and procedures followed for quality check.

Some examples of answers:

To guarantee the quality of the data, various measures have been implemented:
- The experiments were repeated several times by several PhD students and researchers of the team under different conditions
- The data collected were standardized so as not to have any biases (same breeding condition for mammals, same experimental setup with same conditions for regular verification, etc.)
- Regular meeting with the Principal Investigator (PI) for a review of data.

f) Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

[Individual answer according to your project]

Guidance: See below the official text from the EC:

"Results to which access can be given in the form of scientific publications, data or other engineered results and processes such as software, algorithms, protocols, models, workflows and electronic notebooks". (AMGA, Annex 5)

The fact that you have to provide DMP compulsorily for your project means that the research outputs will be well-maintained via the documentation. Moreover, it also addresses all the questions related to production and re-use of data.

8. Other research outputs

a) In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

[Individual answer according to your project]

b) Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

[Individual answer according to your project]

9. Allocation of resources

a) What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?

[Individual answer according to your project]

Guidance: Please calculate the expected costs (software, human resources) for the research data management. This could include the

costs necessary for maintaining the server (if it's institutional, for instance), staff time and salary, necessary equipment for data visualization and analysis, cost for filing patent or depositing your data online in a given repository. The cost required for technical expertise and training for the personnel to maintain the data should also be reported. See the website below for more:

https://www.ukdataservice.ac.uk/manage-data/plan/costing
how-to-comply-to-h2020-mandates-rdm-costs (openaire.eu)

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

[Individual answer according to your project]

Guidance: Specify the options that you have in your mind for data management and their publication at the end of your project.

Here are some helpful resources:
https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/costing/

Data management costing tool and checklist
(ukdataservice.ac.uk)how-to-comply-to-h2020-mandates-rdm-costs
(openaire.eu)

c) Who will be responsible for data management in your project?

Some examples of answers:

Guidance: As metadata ensures the quality of your research output, it is important to name a person in-charge to maintain them. After having generated a huge volume of data, they ensure that all datasets are properly documented with metadata according to the principles detailed in the previous sections of this DMP.

Some examples of answers:
The principal investigator (PI) will be responsible for this part of the project.
We will collectively manage the Data Management with the support of all the researchers involved
Our data management will be handled by our European Project Manager (give his/her name) while consulting and regrouping the suggestions coming from all the partners involved in the study.

d) How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

[Individual answer according to your project]

Guidance: Explain in detail, along with the scientific need(s), the monetary and material necessities to preserve your data for long-term. Explain, how long will they be preserved, what will be preserved and how will you maintain the quality of datasets preserved. For instance, in France, the national organism CINES offers preservation service for your data on long-run (more than 30 years, for example).

Some examples of answers:
"The long-term preservation solution will be kept in place. Our collaborators are developing a special platform that could house our data for 10 additional years after the end of the project".

10. Data security

a) What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

Guidance: Data security includes data recovery as well as secure storage and transfer of sensitive data. Explain how the security of your data will be managed, who will be given the access and how it will be controlled (for persons out of collaboration during the project, for instance).

What about the protection of sensitive data (industrial secrets, for instance)? What are the risks associated with them? What are the guidelines from your institution regarding the data protection policy? Are you implementing them?

Some examples of answers:
The third party will not have access to our clinical data before our project. It involves a wide range of anonymization so as to preserve the potential medical history of the patients involved in our study.

b) Will the data be safely stored in trusted repositories for long term preservation and curation?

[Individual answer according to your project]

11. Ethical aspects

a) Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Guidance: Explain about the judicial and moral aspects related to data sharing policies. Further study on this issue can be done on How to complete your ethics self-assessment.

Anonymization of personal data: Be sure not to include any sensitive data while submitting the datasets. Remember to completely anonymize/pseudonymize them as per the well-defined standards so as to avoid any legal or ethical obstacles.
Be sure to have the consent of the people involved in your study. State whether the patients in your study (or people involved in your sample census, for instance) agree to have their personal information associated with the metadata.

Personal rights: As the online submission of your data is done by a single person, be sure to have a collective consent of all the collaborating scientists so that they agree to have their names and affi-

liations displayed. Be sure also to confirm that you are not violating any personal rights while omitting the name(s) of any participating scientist(s)[3].

Text Suggestions:
The contents uploaded in Recherche Data Gouv are done once the anonymization (performed via a commercial software) of our collected data is finished. This is in-line with the good scientific practices recommended by the European Commission.

b) Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?

[Individual answer according to your project]

12. Other issues

a) Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

[Individual answer according to your project]

If yes, please describe the procedures in brief.

3 For better understanding, see the following link: https://api-depositonce.tu-berlin.de/server/api/core/bitstreams/9101c798-7e39-470a-97bb-1f2cd1b1ff3d/content

Annex:

The creation of this annotated Horizon Europe DMP Guide was inspired by the following documents:

« Horizon Europe Programme Guide », 60p, version 3.0, April 2023. (OS: “Open science in Horizon Europe”, pp.40-56.)

« EU Grants. AGA – Annotated Model Grant Agreement. EU Funding Programmes 2021-2027 », 328p, version 1.0 - DRAFT », 01 April 2023. (OS: “Annex 5 HE Communication, Dissemination, Open Science and Visibility”, pp.277-290.)

Kuberek, Monika, Guidance for Creating a Data Management Plan in Horizon 2020 Projects, Technische Universität Berlin, 23 October 2019. CC-BY, 4.0.
DOI: http://dx.doi.org/10.14279/depositonce-7199

England, Jonathan, Malaguarnera, Giulia, & Príncipe, Pedro. (2022, June 14). OpenAIRE Webinar: Horizon Europe Open Science requirements in practice. Zenodo. https://doi.org/10.5281/zenodo.6641829

Horizon Europe Data Management Plan Template, version 1.0, 05 May 2021.

DMP Opidor (https://dmp.opidor.fr/plans)

“Horizon Europe requirements for research data”, https://www.openaire.eu/horizon-europe-os-requirements-in-practice-webinar-results-key-messages

“RDM in Horizon Europe Proposals”: https://www.openaire.eu/rdm-in-horizon-europe-proposals

“How to identify and assess Research Data Management (RDM) costs”: https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs

“What will it cost to manage and share my data?”: https://www.openaire.eu/rdm-researcher-costs-infographic/view-document

* http://rd-alliance.github.io/metadata-directory/
* https://www.dcc.ac.uk/guidance/standards/metadata

* https://doranum.fr/metadonnees-standards-formats/
* https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

"Science ouverte et montage de projets de recherche"
AgroParisTech, 2023

List of glossaries:

Horizon Europe Program: European Commission's research and innovation funding programme until 2027. For detailed information, see: https://www.horizon-europe.gouv.fr/

Data management plan: A living summary-document that provides assistance with organising and planning all the phases of the data lifecycle. "Data management plans (DMPs) are a cornerstone for responsible management of research outputs, notably data and are mandatory in Horizon Europe for projects generating and/or reusing data (on requirements and the frequency of DMPs as deliverables consult the AGA article 17)".

Research Data: "Factual records (figures, texts, images, sounds, videos...), which are used as primary sources for scientific research and are generally recognised by the scientific community as necessary to validate research results." (Source: Definition of OCDE (Organisation for Economic Co-operation and Development, 2007.)

Dataset: A data set (or dataset) is a collection of data in a well-structured and organised form.

Data Repository: A digital space where digital objects can be deposited or hosted.

Metadata: Metadata represents all the necessary information about data. It enriches the data with information so that it is easier to find, use, and manage.

Metadata standards: "A metadata standard is a requirement which is intended to establish a common understanding of the meaning or semantics of the data, to ensure correct and proper use and interpretation of the data by its owners and users. To achieve this common understanding, a number of characteristics, or attributes of the data have to be defined, also known as metadata" (https://en.wikipedia.org/wiki/Metadata_standard)

Machine-Readable: “The capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention.” (Details on: GO-FAIR)

FAIR principles: Findable, Accessible, Interoperable and Reusable.
Findable aims to simplify how data is found by humans and computer systems, by requiring a description and indexation of (meta)data.
Accessible promotes sustainable conservation of (meta)data and makes accessing or downloading them easier, by specifying how they can be accessed (open or limited access) or used (under licence).
Interoperable can be broken down into being downloadable, usable, intelligible, and combinable with other data, by humans and machines.
Reusable highlights the characteristics that make the data reusable for future research or other purposes (teaching, innovation, reproduction/transparency of science). Its primary aim is to make all results verifiable.

PIDs (Persistent and Unique Identifiers): A Persistent Identifier (PI or PID) is a long-lasting reference to a document, file, web page, or other object. (Wikipedia).

For numerical object : DOI (Digital object identifiers), Handle, ARK (Archival Resource Key), PURL (Persistent Uniform Resource Locator), URN (Uniform Resource Name), SWHID (Software Heritage ID, a unique persistent identifier for software deposited in the Software Heritage open archive), etc.

For contributors: ORCID, ISNI (International Standard Name Identifier), IdHAL, etc.

For organisations: ROR (Research Organization Registry); a registry of open persistent identifiers for research organizations.

Ontologies: “Can be roughly described as a vocabulary with hierarchies, meaningful relations among concepts, and their constraints”. (GO FAIR, more context).

Thesaurus: “synonym dictionary or dictionary of synonyms, is a

reference work which arranges words by their meanings,[1][2] sometimes as a hierarchy of broader and narrower terms, sometimes simply as lists of synonyms and antonyms" (Wikipedia)

Recherche data gouv (https://recherche.data.gouv.fr/fr): French national repository for sharing and opening research data. Université Paris-Saclay has its own institutional space for the deposit of research data: https://entrepot.recherche.data.gouv.fr/dataverse/upsaclay

CC-BY: Creative Commons Attribution International License (CC BY).

CC0: Public Domain Dedication (CC0)

GitHub: An internet service for software development and hosting.

Zenodo: General open repository developed under the European OpenAIRE program and operated by CERN, in which researchers can deposit papers, data, publication, reports, software, etc.

université-paris-saclay.fr
Bâtiment Breguet - 3 rue Joliot Curie
91190 Gif-sur-Yvette

université
PARIS-SACLAY

CentraleSupélec

école
normale
supérieure
paris—saclay