



## **D4.7 Guidelines for Data Management Plan implementation**

### **WP4 Joint Integrative Projects**

Responsible Partner: Sciansano

Contributing partners: SVA



## GENERAL INFORMATION

<b>European Joint Programme full title</b>	Promoting One Health in Europe through joint actions on foodborne zoonoses, antimicrobial resistance and emerging microbiological hazards
<b>European Joint Programme acronym</b>	One Health EJP
<b>Funding</b>	This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 773830.
<b>Grant Agreement</b>	Grant agreement n° 773830
<b>Starting Date</b>	01/01/2018
<b>Duration</b>	60 Months

## DOCUMENT MANAGEMENT

<b>Deliverable</b>	D4.7 Guidelines for Data Management Plan Implementation
<b>WP and Task</b>	WP4; Task 5-2
<b>Leader</b>	Valérie De Waele (Sciensano)
<b>Other contributors</b>	Mickele A.D. Francisco (external expert), Marc Dispas (Sciensano), Mickaël Cargnel (Sciensano), Maria-Eleni Filippitzi (Sciensano) and Ann Lindberg (SVA)
<b>Due month of the deliverable</b>	M10
<b>Actual submission month</b>	M12
<b>Type</b> <i>R: Document, report</i> <i>DEC: Websites, patent fillings, videos, etc.</i> <i>OTHER</i>	R and DEC
<b>Dissemination level</b> <i>PU: Public</i> <i>CO: confidential, only for members of the consortium (including the Commission Services)</i>	PU



<b>HISTORY OF CHANGES</b>			
<b>Version</b>	<b>Publication date</b>	<b>Authors</b>	<b>Change</b>
1.0	19.12.2018	Valérie De Waele (Sciensano)	▪ Initial version
			▪



## Table of contents

1	General introduction .....	7
2	Introduction to Data Management Plan .....	7
2.1	What do we mean by “data”? .....	7
2.2	What do we mean by data management plan (DMP)? .....	8
2.3	What may your data management plan describe? .....	9
2.4	H2020 requirements .....	9
2.5	Data life cycle and DMP components .....	9
2.6	FAIR principle.....	10
2.6.1	Findable .....	10
2.6.2	Accessible .....	11
2.6.3	Interoperable.....	11
2.6.4	Re-usable .....	12
3	Usefulness of DMP .....	13
3.1	Facilitate project management .....	13
3.2	Saving time and resources in the long run .....	13
3.3	Preventing duplication of effort (data interoperable and reusable) .....	13
3.4	Ensuring research accuracy .....	14
3.5	Ensuring research integrity.....	14
3.6	Ensuring research reproducibility (data findable and accessible).....	14
4	Steps guidance to develop DMP.....	15
4.1	Guidance.....	15
4.2	How-to.....	15
4.3	Data planning checklist.....	16
4.4	Development of the first version of the project DMP.....	16
4.4.1	Steps toward FAIRer data .....	17
4.4.2	How FAIR are your data? .....	17
4.4.3	Things to remember .....	18
4.4.4	Further to the FAIR principles, DMPs should also address: .....	18
4.5	Review of the project DMP .....	18
5	Detailed DMP components .....	19
5.1	Data Collection & Data Format .....	19
5.1.1	Project Information .....	19
5.1.2	Data Collection .....	19



5.1.3	Data formats.....	22
5.1.4	Data formats transformation .....	24
5.1.5	Summary.....	25
5.2	Allocation of resources (Responsibilities and resources).....	26
5.2.1	The cost of data management.....	26
5.2.2	How to calculate costs?.....	26
5.3	Sharing, and preservation .....	27
5.3.1	Sharing your research data.....	27
5.3.2	Data sharing: Benefits .....	27
5.3.3	Data sharing: Barriers.....	28
5.3.4	Access, Sharing and Privacy Examples .....	29
5.4	Making Data “Findable”, Including Provision for Metadata .....	30
5.4.1	Data ID, data citation and impact.....	31
5.4.2	Data paper .....	32
5.4.3	Organizing data, naming convention and versioning.....	32
5.4.4	Documentation.....	35
5.4.5	Metadata .....	41
5.5	Making data openly “Accessible” .....	42
5.5.1	How open is open? .....	43
5.5.2	Security.....	44
5.5.3	Storage.....	46
5.5.4	Backup .....	48
5.5.5	Data repository.....	51
5.5.6	Data protection .....	53
5.5.7	Ethical review of research projects .....	55
5.6	Making data “Interoperable” .....	57
5.6.1	The different levels of interoperability .....	57
5.7	Increase data “Reuse”, through clarifying licenses .....	58
5.7.1	Open data licensing.....	59
5.7.2	Creative Commons 4.0 licenses explained .....	60
5.7.3	Licensing Example.....	61
5.7.4	Policies and provision for reuse & re-distribution examples .....	61
6	Annexes .....	62
6.1	How to use the DMPonline.be tool? .....	62



6.2	How to use the OHEJP community repository on Zenodo? .....	66
6.3	Dublin Core Metadata Element Set.....	69
6.4	Data Management Plan Checklist Sample.....	72
6.5	Core Vocabularies sample .....	76
6.6	A Data Management Plan Example from DMPTool Public Plan.....	77
7	Further support in developing your DMP .....	79



## 1 General introduction

An **overarching data management plan (DMP)** (doi: 10.5281/zenodo.2541570) defines the strategy on how One Health EJP (OHEJP) data are managed under conditions that conform with the requirements of Horizon 2020. As the OHEJP is a co-funded program, agreements between partners and stakeholders are required to collect/process/use data. It must be acknowledged that the source of co-funding may have priority in some decisions regarding data management, i.e. that it may dictate where and how the programme output, including data, should be deposited and named. Consequently, the principles provided by the OHEJP overarching DMP are meant to complement any requirements from individual funders, while still ensuring that the data are FAIR, as far as possible.

Due to the heterogeneity of the data that will be collected, processed or generated within OHEJP, and due to the level of detail needed, each joint research project (JRP) and joint integrative project (JIP) have to develop **project specific DMP's**, using as baseline the overarching DMP and guide towards DMP implementation via training materials: **webinar** held on the 19<sup>th</sup> of December 2018 and available online (video: doi: 10.5281/zenodo.2564974; slides: doi: 10.5281/zenodo.2565750); **workshop** planned for the Annual Scientific Meeting in May 2019, and present **guide**.

**The present guidance document is meant to introduce researchers to DMP and to provide direction to implement it. Different details of information regarding DMP and FAIR principles (i.e. Findable, Accessible, Interoperable, and Re-useable) can be found in the document. Step by step guide tour is also provide in the following sections:**

- [Steps guidance to develop DMP](#)
- [How to use the DMPonline.be tool?](#)
- [How to use the OHEJP community repository on Zenodo?](#)

**It is recommended to use the DMPonline.be tool to develop your project DMP and to deposit it on the community OHEJP repository set up on OpenAIRE platform: <https://www.zenodo.org/communities/ohejp/?page=1&size=20>.**

## 2 Introduction to Data Management Plan

A data management plan (DMP) is a written document that describes the data you expect to acquire or generate during the course of a research project. It describes how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.

You may have already considered some, or all of these issues with regard to your research project, but writing them down helps you formalize the process, identify weaknesses in your plan, and provide you with a record of what you intend or intended to do.

A data management plan will help you in the management of research data generated in your project. Put simply, a data management plan, describes the collected or generated data that in the course of your work and what happens to this data during its life cycle (storage, publication, citation, long-term availability, anonymity, deletion, etc.).

The goal of a data management plan is to meet the requirements of good scientific practice and to allow for reproducibility of research results. It is recommended to start early with the preparations for the handling of research data and to update the procedures during the project.

### 2.1 What do we mean by “data”?

By data, we mean the content generated in your project such as, raw data, publications, documentations (reports, lab procedure), research data, events (workshops, webinars) or anything that can be used to support the integrity of the results and reproduce the research. However, the word



data means different things to different people in different contexts. Different disciplines have and use discipline-specific language around the subject 'research data'. Some people refer to everything digital as data. Others refer to both analogue and digital materials as data.

For our purposes, research data is regarded as the data created in a digital form (born digital) or converted to a digital form (digitized).

### What are research data?

Unlike other types of information, research data are collected, observed or created for the purposes of analysis to produce and validate original research results.

Data may be viewed as the lowest level of abstraction from which information and knowledge are derived. Another way of looking at it is to view "data" as the level at which measures were originally collected, e.g. Individual responses to a survey or census, hourly measures of temperature, wind speed and wind direction, etc.

Research data can also be regarded as **situational** in that the same digital information or materials may be data for some research questions but not others. Likewise, the same information may be research data for a person at one time point, but not data at another time point, depending on whether that person uses that information or material for analysis.

For example:

- A photographic image of an old municipal building in a historical archive is an archived image in an image bank. However, when used by a researcher to study the history of a city, the photographic image becomes data for that researcher.
- CCTV footage may be archived (or destroyed) by a security firm. However when used by a researcher to study human behavior or 21st century surveillance methods, the video footage becomes data for that researcher.

Thus, research data are very much, about when they are used, what they constitute and the purpose for which they are to be used. Data can also be created by researchers for one purpose and used later by another set of researchers for a completely different research agenda.

## 2.2 What do we mean by data management plan (DMP)?

A data management plan (DMP) is a formal document you develop at the start of your research project, which outlines all aspects of your data (e.g. what you will do with your data during and after your research project). Data Management Plans are a key element of good data management (European Commission, 2016). Data management refers to all aspects of creating, housing, delivering, maintaining, and archiving and preserving data. It is one of the essential areas of responsible conduct of research.

Information regarding your data management needs to be easily found and understood, not least if you are working on a project that runs over several years and involves a large team of people. In order to simplify data management, a DMP can be created early in the research process. A DMP is a formal document that provides a framework for how to handle the data material during and after the research project. It is a "living" document that changes together with the needs of a project and its participants and should be updated throughout the project, in order to make sure that it tracks such changes over time and that it reflects the current state of your project.

**Important!** Data management plans must be continuously maintained and kept up-to-date throughout the course of the research project.





A lot of diversity exists in DMPs because they are always built around the particular needs of the data collected within your project. Sometimes there are particular requirements that have to be answered in the DMP from stakeholders such as your funders or your institution.

### 2.3 What may your data management plan describe?

- What research data you will be creating or collecting?
- Who will be responsible for each aspect of the management plan you are developing?
- What policies (funding, institutional, and legal) will apply to your data?
- How will the data be organized (folder structures, file naming conventions, file versioning)?
- How will the data be documented during the collection and analysis phase of your research?
- What data management practices (backups, storage, access control, archiving) will you use to store & secure your data?
- What facilities and what equipment will be required (hard-disk space, backup server, and repository)?
- Who will have ownership and access rights to your data?
- How the data will be preserved and made available in the long term, once your research is completed?

### 2.4 H2020 requirements

Good data management is the basis of successful research. Managing your data effectively across the data lifecycle is very important for the success of your research project.

In Horizon 2020, the Commission has extended the pilot for open access to research data (ORD pilot). The pilot aims to improve and maximize access to and reuse of research data generated by Horizon 2020 projects, taking into account

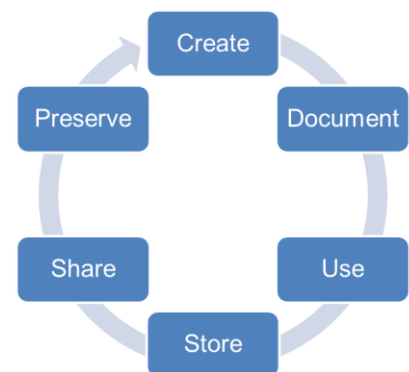
- The need to balance openness and protection of scientific information
- Commercialization and IPR
- Privacy concerns
- Security
- Data management and preservation questions
- **Open access to scientific publications**, which is an obligation,
- **Open access to research data**, where opt-outs are possible
- **Research data management**

Participating projects will be required to develop a Data Management Plan (DMP), in which they will **specify what data will be open**: detailing what data the project will generate, whether and how it will be exploited or made accessible for verification and reuse, and how it will be curated and preserved.

### 2.5 Data life cycle and DMP components

Activities involved in research data management

- **Creating data**: Think carefully about choosing the type and format of data you will create and how your data will be generated.
- **Documenting data**: Providing information to users (and yourself later) to understand your data. Think carefully:
  - Is the file structure/naming understandable to others?
  - Which data will be kept? And which data can be discarded?
- **Accessing/using data**: How will you organize your research data during the life of your project?
- **Storage and backup**: how do you plan to store and save your data safely and securely during your project?





- **Sharing data:** Making your data publicly available (where possible) at the end of your project.
- **Investigate:** Are you expected/allowed to share your data? What factors might restrict you being able to share your data?
- **Preserving data:** How will you preserve your data after the end of your project?

## 2.6 FAIR principle

The attention of researchers is increasingly directed to the phase of the research lifecycle where data are published, shared, discovered and reused. One of the perceived ways to achieve optimal reuse is to make data FAIR (Findable, Accessible, Interoperable and Reusable) ([doi:10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18)).

The FAIR guiding principles consist of 15 facets, which describe a continuum of increasing reusability. Importantly, data should not only be FAIR for humans but also for machines, allowing for instance, for automated search and access to data. Funders like the European Commission have drafted [Guidelines on FAIR Data Management for the H2020 program](#).

Good data management is one way to support the FAIR principles.



Because humans increasingly rely on computational support to deal with data, as a result of the increase in volume, complexity, and creation speed of data, the authors of the FAIR principle intended to provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets.

The principles emphasize machine-actionability (e.g., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention).

### 2.6.1 Findable

To be **Findable** any Data Object should be uniquely and persistently identifiable

- The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.
- [F1. \(Meta\)data are assigned a globally unique and persistent identifier](#)
- [F2. Data are described with rich metadata](#)





- [F3. Metadata clearly and explicitly include the identifier of the data they describe](#)
- [F4. \(Meta\)data are registered or indexed in a searchable resource](#)
- Principle F1 is arguably the most important because it will be hard to achieve other aspects of FAIR without a globally unique and persistent identifiers. Hence, compliance with F1 will already take you a long way towards publishing FAIR data.
- The above principles mean also that the same Data Object should be re-findable at any point in time, thus Data Objects should be persistent, with emphasis on their metadata.
- A Data Object should minimally contain basic machine-readable metadata that allows it to be distinguished from other Data Objects.
- Identifiers for any concept used in Data Objects should therefore be Unique and Persistent

The focus here is on the **Persistent & Unique**. As an example, the Digital Object Identifier (DOI) is a well-known identifier. Having a persistent identifier or **PID** is an important aspect in making sure your data meet the requirements of Findability and Accessibility. Cf. [A persistent identifier \(PID\) to your dataset](#)

For detailed information, please refer to the [Making Data Findable, Including Provision for Metadata](#) section

## 2.6.2 Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorization.

- [A1. \(Meta\)data are retrievable by their identifier using a standardized communications protocol](#)
  - [A1.1 The protocol is open, free, and universally implementable](#)
  - [A1.2 The protocol allows for an authentication and authorization procedure, where necessary](#)
- [A2. Metadata are accessible, even when the data are no longer available](#)

Data is **Accessible** in that it can always be obtained by machines and humans

- Upon appropriate authorization
- Through a well-defined protocol
- Thus, machines and humans alike will be able to judge the actual accessibility of each Data Object.



In other words, FAIR data retrieval should be mediated without specialized tools or communication methods. Usually, this is provided by the repository where you deposit your data, so it is not something that you have to worry about.

For detailed information on how to achieve accessibility, please refer to the [Making data openly accessible](#) section.

## 2.6.3 Interoperable

Data usually needs to be integrated with other data. In addition, the data needs to interoperate with applications or workflows for analysis, storage, and processing.



- [11. \(Meta\)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.](#)
- [12. \(Meta\)data use vocabularies that follow FAIR principles](#)
- [13. \(Meta\)data include qualified references to other \(meta\)data](#)



For Data Objects to be **Interoperable**:

- Data should be readable by machines without the need for specialized or ad hoc algorithms, translators, or mappings.
- (Meta) data formats utilize shared vocabularies and/or ontologies.
- As an example, *X is regulator of Y* is a much more qualified reference than *X is associated with Y*

For detailed information, please refer to the [Making data interoperable](#) section.

## 2.6.4 Re-usable

The ultimate goal of FAIR is to optimize the reuse of data.

To achieve this, metadata and data should be well described so that they can be replicated and/or combined in different settings.

- [R1. Meta\(data\) are richly described with a plurality of accurate and relevant attributes](#)
- [R1.1. \(Meta\)data are released with a clear and accessible data usage license](#)
- [R1.2. \(Meta\)data are associated with detailed provenance](#)
- [R1.3. \(Meta\)data meet domain-relevant community standards](#)



The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource (the infrastructure component).

For Data Objects to be **Re-usable** additional criteria are:

- In order to allow the user to decide if the data is actually useful in a particular context, the data publisher should provide not just metadata that allows discovery, but also metadata that richly describes the context under which the data was generated. e.g:
  - Describe the scope of your data: for what purpose was it generated/collected?
  - Mention any particularities or limitations about the data that other users should be aware of.
- R2 is about *legal* interoperability. What usage rights do you attach to your data? This should be described clearly. e.g: Commonly used licenses like MIT or Creative Commons can be linked to your data. Cf. [Increase data reuse, through clarifying licenses](#)
- For others to reuse your data, they should know where the data came from, who to cite and/or how you wish to be acknowledged. Include a description of the workflow that led to your data, e.g:
  - Who generated or collected it?
  - How has it been processed?
  - Has it been published before?
- R4: It is easier to reuse data sets if they are similar: same type of data, data organized in a standardized way, well-established and sustainable file formats, documentation (metadata) following a common template and using common vocabulary.

For detailed information on achieving reusability, please refer to the [Sharing, and preservation](#) section.

### 3 Usefulness of DMP

Developing a data management plan may seem daunting. However, it is a vital step in your research process that you cannot afford to skip. It helps you ensure your research data are accurate, complete, reliable, and secure both during and after you complete your research.



To make your research as time-efficient, reproducible and safe as possible, it is important that your data management is well thought through, structured, and documented. A good data management strategy takes into account technical, organizational, structural, legal, sustainability and the ethical aspects. The time invested in setting up a good data management strategy pays off when time comes to reproduce your analysis and

results. You will be able to easily find and understand your data, increase your data's reuse potential and comply with funder mandates at the same time.

#### 3.1 Facilitate project management

Many organizations use different sources of information for planning, trends analysis, and managing performance. Within an organization, different scientists may even use different sources of information to perform the same task if there is no data management process and they are unaware of the correct information source to use.

The value of the information is only as good as the information source, as the old idea of, garbage in garbage out. Data entry errors, conclusion errors and processing inefficiencies, are all risks for institutions that do not have a strong data management plan and system.

DMP helps:

- Clarification of resource requirement
- Flexibility
- Increase decisions accuracy

#### 3.2 Saving time and resources in the long run

Good data management will make your institution more productive. Good DMP makes it easier for scientist to find and understand information that they need to do their job. It also provides the structure for information to be easily shared with others, to be stored for future reference and for easy retrieval.

- Increase productivity

#### 3.3 Preventing duplication of effort (data interoperable and reusable)

Another benefit of proper data management is to avoid unnecessary duplication. By storing and making all data easily referable, organizations ensures not having employees conducting the same research, analysis or work that has already been completed by another colleague.

- Increase cost-efficiency
- Prevent duplication of effort by enabling others to use your data



- Facilitate the analysis of change by providing data, with which, data at other points in time can be compared

### 3.4 Ensuring research accuracy

Ensure your research data and records are accurate, complete, authentic and reliable both during and after you complete your research.

### 3.5 Ensuring research integrity

Successful research is based on competence. However, in practice this alone does not suffice, instead an environment of trust is needed for it to unfold. On the other hand, trust can only develop where action is determined by integrity.

In addition, there are multiple risks if your data is not managed properly or if your information falls into the hands of the wrong people. A strong data management system will greatly reduce the risk of this ever happening to your organization.

The primary reasons of bad data and data loss, is that there is no data management system or plan in place, or the plan or system is of poor quality. Most institutions are reactive to an issue, instead of being proactive, which in the long run, costs them significantly more.

With a data management system and plan in place that all your partners know, can greatly reduce the risk of losing vital information. With a data management plan, things will be put in place to ensure that important information is backed up and is retrievable from a secondary source, if the primary source ever becomes non-accessible.

- Enhancing data security
- Minimizing the risk of data loss



### 3.6 Ensuring research reproducibility (data findable and accessible)

Make your research more visible and with greater impact. Planning helps you to achieve these benefits. It is ultimately most useful to you. Making a plan helps you save time and effort and makes the research process easier. By considering, what data will be created and how, you can check if you have the necessary support in place. Planning also enables you to make sound decisions, bearing in mind the wider context and consequences of different options. Publishers and research funders may require that you share your data, so it is worth investing time to plan for effective data management. Several funders ask for data plans as part of grant proposals.



## 4 Steps guidance to develop DMP

Research data management refers to how you handle, organize, and structure your research data throughout the research process. Data management plans differ according to your research area. They do not come in one-size-fits-all. They should be simple, clear and regularly updated throughout the course of your research.

Here are some tips to get you started with a data management plan:

- Use the Digital Curation Centre's **DMPonline** tool (Cf. [DMPonline.be](https://dmponline.be)).
- Gain an understanding of data management terminology and issues.
- Gain an understanding of your project.
- Check if your group's Institute/Research Centre has a data management plan.
- Check for the availability of data papers that you can use (Cf. [Data paper](#))
- Talk to your colleagues, research coordinator, IT Officer, your institution's Records Management Section to find out about data authorship, Intellectual Property (IP), licensing, storage and backup facilities and policies.
- Use the research data planning checklist provided as an example (Cf. [Data Management Plan Checklist Sample](#))
- Use this document to develop your own data management plan. Do not worry about filling in all the sections. Focus on what is important to you or your research area.
- Do not spend too much time.
- Keep it practical and simple.
- Do not hesitate to ask for help when you need it! Check, what help is available from your institution.

**Important!** Data management plans are live documents and therefore never finished. Review your data management plan regularly through the course of your research.

### 4.1 Guidance

- Involve all project partners
- Guidance from [OHEJP](#): Webinar (video: DOI: 10.5281/zenodo.2564974; slides: DOI: 10.5281/zenodo.2565750), guidance document, forum, workshop planned for the Annual Scientific Meeting held in May 2019 at Dublin
- Guidance from institutional DMP and/or quality system departments
- Guidance from European portal: [Guidelines on data management in Horizon 2020 research projects](#)
- Useful links: [DMPonline](#)

### 4.2 How-to

It is recommended to refer to the Digital Curation Centre [DMPonline](#) tool, which offers DMP templates that match the demands and suggestions of the Guidelines on Data Management in Horizon 2020. A DMP in DMPonline can be saved anytime. It can also be shared and downloaded in various formats. Follow the procedure explained in Annex "[How to use DMPonline.be tool](#) ?".

- The DMP leaders of your project need to ask to the overarching DMP team to create a DMP template on [DMPonline](#) tool
- You need to create your [ORCID](#) to obtain access to the project DMP template
- You then be able to login to [DMPonline](#) tool
- By default for OHEJP projects, the research funder selected is the European Commission (Horizon 2020)
- It is recommended to check the box for additional DCC guidance



- Answer the list of questions provided by the tool, and used the DCC guidance and/or this present document to guide you in answering the different points
- When your DMP is developed, you can publish it on the OHEJP community of the platform Zenodo (see Annex "[How to use the OHEJP community repository on Zenodo?](#)"). A link between the published DMP and the OHEJP website will be made.

### 4.3 Data planning checklist

The planning process for data management begins with a data-planning checklist. A checklist later assists you in the development of your data management plan.

For inspirational purpose, you can also find a detailed checklist example from the Digital Curation Centre at [www.dcc.ac.uk/sites/default/files/documents/resource/DMP\\_Checklist\\_2013.pdf](http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf) or the example provided in Annex (Cf. [Data Management Plan Checklist Sample](#))

Your checklist might include, considering some, or all of the following questions. You can add to or subtract from this suggested list depending on the nature of your project.

- What data will you collect or create, how will it be created, and for what purpose?
- How will you manage any ethical issues?
- How will you manage copyright and Intellectual Property Rights issues?
- What file formats will be used?
  - Are they non-proprietary, transparent and sustainable?
  - What directory and file naming conventions will you use? (Cf. [Organizing data, naming convention and versioning](#))
  - Are there any formal standards that you will adopt?
  - What documentation and metadata will accompany the data? (Cf. [Making Data Findable](#))
- How will the data be stored and backed up during the research? (Cf. [Storage & Security](#))
  - How will you manage access and security?
  - Who will be responsible for data management?
- Are there existing procedures that you will base your approach on? For example, are there institutional data protection or security policies to follow, department or group data management guidelines, or Research Data Management policies defined by your institution or funder that must be considered?
- What is the long-term preservation plan for the dataset? For example, which data should be retained, shared, and/or preserved?
- How will you share the data, and are any restrictions on data sharing required? (Cf. [Sharing, and preservation](#))
- What resources will you require to deliver your plan? For example, are there tools or software needed to create, process, or visualize the data?

### 4.4 Development of the first version of the project DMP

- Begins with your initial considerations regarding what will be necessary for using or collecting your particular type of data.
- Includes measures for maintaining the integrity of the data, making sure that they are not lost due to technical mishaps, and that the right people can access the data at the appropriate time.
- Looks forward to the future, making it clear that you should provide detailed and structured documentation to be able to share your data with other colleagues and prepare them for long-term availability.





#### 4.4.1 Steps toward FAIRer data

You should take appropriate measures to prepare your data for optimal reuse from the start. Submitting your dataset into a data repository is one of the ways to make sure your data will not become “unusable” (Cf. [Sharing, and preservation](#)).

To achieve FAIRness, data objects should at least have:

- **A persistent identifier (PID) for the data object as a whole** (Cf. [A persistent identifier \(PID\) to your dataset](#))

Persistent identifiers like DOIs form the solution to link rot. Link rot is the process by which hyperlinks stop referring to the original source in time because it was moved. Without a PID, the data object simply will not be findable let alone reusable (Cf. [Data ID, data citation and impact](#))

- **A sufficient set of metadata**

A sufficient and standardized set of metadata (elements that describe the data. Cf. [Metadata](#)) will enhance findability, interoperability and reusability. The quality of the descriptive information regarding the data has a profound impact on their reusability. So the more documentation of the data’s context, the better. As a minimum, there should be sufficient amount of metadata to make the data findable but also understandable and reusable by other researchers.

Examples of metadata elements that may be used to describe data: (Cf. [Dublin Core Metadata Element Set](#))

- Title of the project
- Names of researchers involved
- Abstract or summary about the project and the data
- Research topic, or subject of research
- Temporal coverage / information
- Spatial coordinates / Location information
- Instrumentation used
- Access or rights policies/restrictions
- Hardware and/or software used

- **A clear license** (Cf. [Increase data reuse, through clarifying licenses](#))

The motto is ‘**As Open as Possible, as Closed as Necessary**’. Researchers (and computers) who find a dataset should immediately know what they are allowed to do with it. Stating clear reuse rights is like having a warm 'Welcome' on the doormat of your dataset.

In the interest of FAIR data, researchers are advised to deposit their data in a research data archive, along with all the documentation needed to make sure they can be reused, with both the explicit goal and the necessary expertise to store data sustainably and maintain their usability.

Making data FAIR is a joint responsibility of researchers and data repositories. In a comprehensive document, the Swiss National Science Foundation explains ([Explanation of the FAIR data principles](#)) how the responsibilities of both are distinct.

#### 4.4.2 How FAIR are your data?

One way to make sure your data is FAIR is to apply and regularly check the FAIR principle against your data, to determine how Findable, Accessible, Interoperable and Reusable it is.



### 4.4.3 Things to remember

- FAIR is a set of principles, not a standard.
- Does following the FAIR principles mean that your data has to be shared openly with everyone? **NO**.
- Data can be FAIR but not open. For example, data could meet the FAIR principles, but be private or only shared under certain restrictions.
- Open data may not be FAIR. For example, publically available data may lack sufficient documentation to meet the FAIR principles, such as licensing for clear reuse.
- If you are in receipt of H2020 funding the EC requires a Data Management Plan (DMP). The FAIR principles can help you understand how to practically describe, create, store, share, manage and preserve your data in your DMP.

### 4.4.4 Further to the FAIR principles, DMPs should also address:

#### 4.4.4.1 Allocation of Resources

- What are the costs for making data FAIR in your project?
- How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
- Who will be responsible for data management in your project?
- Are the resources for long-term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

#### 4.4.4.2 Data Security

- What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
- Is the data safely stored in certified repositories for long-term preservation and curation?

#### 4.4.4.3 Ethical Aspects

- Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review.
- Is informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?

#### 4.4.4.4 Other Issues

- Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

## 4.5 Review of the project DMP

**A DMP is a “living” document that changes together with the needs of a project and its participants.** It is updated throughout the project to make sure that it tracks such changes over time and that it reflects the current state of your project.

The DMP should be updated as a minimum in time with the periodic evaluation/assessment of the project. If there are no other periodic reviews foreseen within the grant agreement, then such an update needs to be made in time for the final review at the latest. Furthermore, the consortium can define a timetable for review in the DMP itself.

New versions of the DMP should be created whenever important changes to the project occur due to inclusion of new data sets, changes in consortium policies or external factors.



## 5 Detailed DMP components

### 5.1 Data Collection & Data Format

#### 5.1.1 Project Information

<b>Administrative information</b>	<ul style="list-style-type: none"> <li>• Name &amp; ID of the project</li> <li>• Project description</li> <li>• Funding body</li> <li>• Principal investigator name and possibly IF</li> <li>• Project data contact</li> <li>• Related policies</li> <li>• Date of first version</li> <li>• Date of last update</li> </ul>
-----------------------------------	--

#### 5.1.2 Data Collection

This part of DMP should cover the type of data collected, their format, collection methods and any external references.

<b>Data Collection</b>	<ul style="list-style-type: none"> <li>• It may be important to manage the following research records both during &amp; beyond the life of a project</li> <li>• Existing dataset to be reused</li> <li>• Methods by which data will be collected or created</li> <li>• Structures, naming &amp; versioning system for folders and files</li> <li>• Quality assurance processes</li> <li>• Data collection or creation methods</li> <li>• Data description (type, format, volume)</li> <li>• Use of secondary data</li> </ul>
------------------------	--

##### 5.1.2.1 Data classification (collection methods)

Classification should apply to your field of work. The list below presents only some general guidance.

<b>Observational</b>	<i>Data captured in real time, usually unique and irreplaceable e.g. foodborne surveillance</i>
<b>Experimental</b>	Data from experimental results, e.g. from lab equipment, often reproducible, but can be expensive.
<b>Simulation</b>	Data generated from test models where model and metadata may be more important than output data from the model
<b>Derived or compiled</b>	Resulting from processing or combining 'raw' data, often reproducible but expensive e.g. compiled databases, text mining, aggregate census data
<b>Reference or canonical</b>	A (static or organic) conglomeration or collection of smaller (peer reviewed) datasets, most probably published and curated e.g. gene databanks, crystallographic databases

##### 5.1.2.2 New (primary) or reused (secondary) data (and origin)

As the name suggests, primary data is one that is collected for the first time by the researcher while secondary data is the data already collected or produced by others. The most important difference is that primary data is factual and original whereas secondary data is just the analysis and interpretation of the primary data.



### 5.1.2.3 Data domains and categories

Data Domains are not physical repositories or databases. Instead, they are “logical” categories or groupings of data deemed important and necessary. Data Domains include both internally generated data as well as externally acquired data. It is imperative that these strategic categories of data are identified, defined and inventoried to ensure their proper maintenance and use.

<b>Domain</b>	<p><i>A logical representation of a category of data that has been designated and named e.g.</i></p> <ul style="list-style-type: none"> <li>• <i>Animal health data</i></li> <li>• <i>Food data</i></li> <li>• <i>Public health data</i></li> <li>• <i>Environmental data</i></li> </ul>
<b>Categories</b>	<ul style="list-style-type: none"> <li>• <b>Publications</b> are where results and knowledge are shared and taught to others</li> <li>• <b>Data paper</b> (Cf. <a href="#">Data paper</a>)</li> <li>• <b>Records</b> (a version of data) e.g.             <ul style="list-style-type: none"> <li>○ Master Record</li> <li>○ Copy of data</li> </ul> </li> <li>• <b>Research data:</b> unlike other types of information, research data is collected, observed or created to produce original research results for the purpose of analysis. It is a recorded factual material, which is commonly retained by and accepted, in the scientific community, as necessary to validate research findings. Although, the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created.</li> </ul>

### 5.1.2.4 Data types: data objects and data documentation

<ul style="list-style-type: none"> <li>• <i>Questionnaire data</i></li> <li>• <i>Clinical data</i></li> <li>• <i>Biological data</i></li> <li>• <i>Molecular data</i></li> <li>• <i>Modelling data</i></li> <li>• <i>Other</i></li> </ul>
---

A good documentation for your research data will facilitate its use by others. DMP should indicate the type of documentation you would provide with the data.

<b>DATA</b>	<b>OBJECTS</b>
<b>E.g.</b>	Georeferenced satellite image, Audio interview file, Nuclear magnetic resonance image (NMRI), Thin layer chromatogram, Electron Microscopy Micrographs
<b>Research data</b>	<ul style="list-style-type: none"> <li>• Documents (text, Word), spreadsheets</li> <li>• Laboratory notebooks, field notebooks, diaries</li> <li>• Questionnaires, transcripts, codebooks</li> <li>• Audio tapes, video-tapes (ex: Audio interview file)</li> <li>• Photographs, films (ex: Nuclear magnetic resonance image (NMRI))</li> <li>• Test responses</li> <li>• Slides, artefacts, specimens, samples</li> <li>• Collection of digital objects acquired and generated during the process of research</li> </ul>



	<ul style="list-style-type: none"> <li>• Statistical or other data files</li> <li>• Database contents (video, audio, text, images)</li> <li>• Models, algorithms, scripts</li> <li>• Contents of an application (input, output, log files for analysis software, simulation software, schemas)</li> <li>• Methodologies and workflows</li> <li>• Standard operating procedures and protocol</li> </ul>
<b>DATA</b>	<b>RECORDS</b>
<b>E.g.</b>	Technical Appendix about the methods used in the project, Email correspondence about the project, etc.
<b>Research data</b>	<p>It may be important to manage the following research records both during and beyond the life of a project:</p> <ul style="list-style-type: none"> <li>• Correspondence (electronic mail and paper-based correspondence)</li> <li>• Project files</li> <li>• Grant applications</li> <li>• Ethics applications</li> <li>• Technical reports</li> <li>• Technical appendices</li> <li>• Research reports</li> <li>• Research publications</li> <li>• Master lists</li> <li>• Signed consent forms</li> <li>• Social media communications such as blogs, wikis, tweets etc.</li> </ul>

#### 5.1.2.5 Data type example

<ul style="list-style-type: none"> <li>• <i>“The associated data types will be captured using X survey software and analyzed using X data analytics tool.”</i></li> <li>• <i>“Over the course of the project, data will be generated from sensors and recorded in X format.”</i></li> <li>• <i>“This project will produce public-use nationally representative survey data for Belgium covering Belgian’ social backgrounds, enduring political predispositions, social and political values, perceptions and evaluations of groups and candidates, opinions on questions of public policy, and participation in political life.”</i></li> <li>• <i>“This project will generate data designed to study the prevalence and correlates of DSM III-R psychiatric disorders and patterns and correlates of service utilization for these disorders in a nationally representative sample of over 8000 respondents. The sensitive nature of these data will require that the data be released through a restricted use contract.”</i></li> <li>• <i>“Few datasets exist that focus on this population in the Europe and how their attitudes toward assimilation differ from those of others. The primary resource on this population, [give dataset title here], is inadequate because...”</i></li> <li>• <i>“Data have been collected on this topic previously (for example: [add example(s)]). The data collected as part of this project reflect the current time and historical context. It is possible that several of these datasets, including the data collected here, could be combined to better understand how social processes have unfolded over time.”</i></li> <li>• <i>“For quantitative data files, the repository ensures that missing data codes are defined, that actual data values fall within the range of expected values and that the data are free from wild codes. Processed data files are reviewed by a supervisory staff member before release.”</i></li> </ul>
--

### 5.1.3 Data formats

Specify what file formats will be used for data collection and processing. Make use of commonly used data formats. If special data formats are used, specify how to read them.

<b>FORMAT</b>	<b>FILE TYPE EXAMPLES</b>
<b>Text</b>	Plain text file (TXT), Flat files (EMBL), MS Word (DOC, DOCX), Portable Document Format (PDF), Rich Text Format (RTF), Hyper-Text Markup Language (HTML), Extensible Markup Language (XML)
<b>Numerical</b>	SPSS, Stata, MS Excel, SAS, Flat files, fixed field format files, delimited files, Hierarchical files
<b>Multimedia</b>	JPEG, TIFF, GIF, Dicom, MPEG, Quicktime, Bitmap, PNG
<b>Models</b>	3D, Statistical, Similitude, Macroeconomic, Causal
<b>Software</b>	Java, C, Perl, Python, Ruby, PHP
<b>Discipline Specific</b>	<u>Chemistry</u> : Crystallographic Information File (CIF), <u>Meteorology</u> : GRIdded Binary (GRIB)
<b>Instrument Specific</b>	Carl Zeiss Digital Microscopic Image Format (ZVI)

#### 5.1.3.1 Introduction to file formats

A file format is a way of encoding information within a computer file. A program or application must be able to recognize the file format in order to access data within the file. For example, a web browser is able to process and display a file in the HTML (hypertext markup language) file format so that it appears as a Web page.

File format is often indicated as part of the file name by an extension, or suffix. Conventionally, the extension follows a dot in the filename and contains three or four letters that identify the format (.jpg or .jpeg).

The software in which the file was created must usually open the file in proprietary formats. Someone without a license to the software may not be able to open the file at all.

Open formats, in which the software company or collective, publishes the format rather than keeps it proprietary, are more likely to be readable by more than one application. Adobe PDF is an example of an open format that may be viewed in a number of applications, not just Adobe products.



#### 5.1.3.2 Text vs binary formats

File types are either text or binary encoding:



- Text files are machine-readable through a character-encoding standard such as ASCII or Unicode. Well-known file extensions of 'plain text' are **.txt**, **.csv**, **.asc**, **.html** and **.xml**.
- Binary files can only be read by applicable software, and may be proprietary. Only binary formats can be executed. Some files may contain both binary and text (such as Rich Text Format - **.rtf** files).

A great advantage of creating or saving research data in a text format, where possible, is that such files can be read in by a plain text editor like *Windows Notepad* and are human readable. They can be read by any operating system, and by a wide range of applications. Therefore, text files are the most unlikely to become obsolete over time, and are a good format for sharing and long-term preservation.

Most software applications offer export or exchange formats that allow a text-formatted file to be created for importing into another program. A typical example is Microsoft Excel, which through the Save As command, can save spreadsheet data in comma delimited format (**.csv** or comma separated values).

### 5.1.3.3 File format obsolescence & standards

File formats that are non-proprietary (e.g. open source) or in widespread use, will tend to retain the best chance of being readable in the future. Proprietary formats, especially non-standard formats, used only by a specific software program or specific software version, are likely to present problems for future use. Rapid changes in technology and the market mean that file formats can become obsolete quickly - often a software application is unable to read a file created by an earlier version of itself. The implications for research data management depend on how long data need to be retained for your (or others') future use.

Data formats that conform to an agreed international standard are less likely to become obsolete, because a variety of software applications should be able to read them. However, there are likely to be trade-offs in terms of software functionality, for example loss of formatting or macros.

Sometimes a *de facto* standard is used. For example, **PDF**, an openly published portable document format invented by Adobe, has become a *de facto* standard for publishing documents on the World Wide Web in a way that retains the original layout, fonts and text formatting.

### 5.1.3.4 File format migration

At some point during your research, you may need to convert or migrate your data files from one format to another. This may be due to a new computer, new software, sharing with someone who has different software, working on shared platforms, or simply in order to ensure that your data can be used in the future.

Some "*lossiness*" (that is the loss of information and/or quality in the original data) may occur when migrating from one file format to another. It is important for you to understand what is at risk for the type of data you are working with.

**Potential risks** for loss or corruption on conversion or migration to new media

- Word-processed files: fonts, text formatting, headers, footers, footnotes, links to other documents...
- Numeric files: special characters (such as tabs), end of line (returns), last characters in rows (due to row size limitations), and especially blanks used as a missing data code, last rows (due to row number limitations).
- Database files: as above but also relations among items in a table and among tables.
- Image files: loss of layers, color fidelity, resolution etc.
- Multimedia: as above, but attention to frame rates, sound quality, codecs and wrappers are needed.

File sizes may change and even become surprisingly large.



It is worth briefing yourself on the format you are converting from (and to) before you begin.

Check the integrity of converted files as thoroughly as possible immediately afterwards, e.g. by counting rows and columns, testing functionality, testing export, etc. as well as simply 'eyeballing' the data to check it looks as it should.

#### 5.1.3.5 *Time to reflect*

In thinking about your planned research, consider what software dependencies you have and what risks are involved with these dependencies.

Can you think of a way to reduce the risk to the longevity of your research outputs by using or saving to another format? Can you envisage a reason you might have to migrate your work to another file format before your research is finished?

If you are using proprietary software for your research, how likely is it that the company will change hands sometime soon?

Is there an option to export a generic, non-proprietary format, such as Open Document Formats (ODF), .csv format for tabular data files, text, html/xml/sgml or pdf format for word processed files, tiff or jpeg formats for raster files etc.?

Should that be included in your data management plan?

#### 5.1.3.6 *Standard Data example*

- “Research data will be stored using X file formats. Related files in different formats will be linked by file naming conventions, e.g. “
- “Digital video data files generated will be processed and submitted to the [repository] in MPEG-4 (.mp4) format.”
- “Data will be stored in a CVS system and checked in and out for purposes of versioning. Variables will use a standardized naming convention consisting of a prefix, root, and suffix system. Separate files will be managed for the two kinds of records produced: one file for respondents and another file for children with merging routines specified.”
- “Data will conform to best practices and standards from the X community.”
- “Internal calibration (for geophysical data), instrument calibrations, duplicate samples and field blanks (for hydro chemical data) will be recorded and tested against collected/recorded data to ensure their validity. Qualitative descriptions (lithological data) will be validated through comparative descriptions of collected materials.”

### 5.1.4 Data formats transformation

#### 5.1.4.1 *Data compression*

At some point, you may choose to compress your data files for the purpose of local or networked storage, transportation or transmission. This is known as bit-rate reduction, which involves encoding information using fewer bits than the original representation. Zip (.zip) is a *de facto* standard compression format that is used on Windows, Macintosh, Linux and UNIX platforms. Zip is a “*lossless*” type of compression, which means the file should be identical to the original once unzipped.

There are also “*lossy*” types of compression associated with some multimedia file formats, which may result in some distortion or loss of quality/fidelity when played. *Lossiness* can be one trade-off of compression. Another is the processing time it takes to compress/decompress before or during use, or the amount of computing resource this takes, in the case of very large files or shared servers.





#### 5.1.4.2 Data normalization

Depending on the research field you are working in, normalization may have different meanings. Below are two distinct meanings, each of which may be relevant to your own research data management practice.

- **Statistical normalization:** using a formula or algorithm to transform variables measured on different scales into a common scale so that they can be compared or analyzed in a chosen statistical model. A typical example is computing the logarithm of the variables to make a skewed distribution normal.
- **Database normalization:** following a set of relational database design rules that make the database more robust by eliminating duplication and inconsistency. For example, breaking up large tables into smaller groups of tables, and linking fields between tables through a “key” or common ID. By reducing complexity, the chance of anomalies occurring in the data is reduced and the database becomes more flexible with regard to how it can be used.

#### 5.1.4.3 Data transformations

There are a number of reasons you might need to transform your data during your research project or afterwards. Unlike the earlier discussion about migrating your file format, data transformation involves computing new values from old in the actual data content. Of course, there are implications for decisions about which form of the data to keep for the long term, which form to share with other researchers, and how to document all changes to the data.

Some transformations are purely statistical, and are used to prepare the data to fit a model. An example was given earlier of one kind of transformation.

Another reason for data transformation may be to visualize the data effectively. A simple example is converting data where there is a numerator and a denominator, from ratios to percentages in order to display on a bar chart or pie graph.

A number of techniques may be used to transform confidential or sensitive data so that they may be shared with other researchers. These include:

- **Aggregation:** the combination of related categories, usually within a common branch of a hierarchy, to provide information at a broader level to that at which detailed observations are taken.”
- **Anonymization:** cases are stripped of revealing identifiers such as name and address. Pseudo anonymization is a common technique for protecting identities in qualitative data.
- **Perturbation:** a deliberate distortion is introduced at the level of tabular data cells. E.g., Population Census data are sometimes released with perturbations as a trade-off for geographical detail.

#### 5.1.5 Summary

A file format encodes information in a computer file, enabling another program to access data within it. **HTML** and **PDF** are two examples of commonly used file formats and may be identified by the filename suffixes **.html** and **.pdf** respectively. Cf. [File format obsolescence & standards](#)

Files are based on either text or binary encoding. The former is both machine and human readable and the latter only readable by means of appropriate software. Thus, text files are less likely to become obsolete. Examples of file name extensions for text files are **.txt** and **.csv**. Cf. [Text vs binary formats](#)

If you need to convert or migrate your data files from one format to another, you need to be aware of the potential risk of the loss or corruption of your data and take appropriate steps to avoid/minimize it. Cf. [File format migration](#)



When compressing your data files for the purpose of storage, transportation or transmission, you encode the information using fewer bits than the original representation. Commonly used compression programs are Zip, GNU Zip (**.gzip** or **.tar.gz**) and Stuffit. Cf. [Data compression](#)

In your research, you may also use the process of data normalization. Two meanings of this that may be of relevance to you, statistical normalization and database normalization. Cf. [Data normalization](#)

You may also need to compute new values from old in your data, a process called data transformation, which may also be a necessary prelude to analyzing your data. Three techniques classified as data transformation are: aggregation (combining data into larger units), anonymization (removing information identifying human subjects) and perturbation (distortion). Cf. [Data transformations](#)

## 5.2 Allocation of resources (Responsibilities and resources)

### 5.2.1 The cost of data management

Under H2020, you are encouraged to deposit your data in a research data repository where they will be findable and accessible for others. Data management and sharing activities need to be costed into research, in terms of the time and resources needed. By planning early, costs can be significantly reduced.

### 5.2.2 How to calculate costs?

We cannot predict your costs for you. The costs for data management and storage vary and depend on your project and the volume, the domain, level of documentation and preservation of your data.

Remember that when your research project nears the end you do not want these additional data management activities to compete with delivery of your planned outputs, writing of publications and the timely delivery of your project. At this later stage, the costs of preparing data for sharing may be significantly higher.

Writing a data management plan in itself will cost you about a few hours to a few days, depending on the complexity of your project. It is a well spent time, because early planning of data management (especially when preparing for a funding application) can significantly reduce the costs.

#### 5.2.2.1 Cost Example

“Staff time has been allocated in the proposed budget to cover the costs of preparing data and documentation for archiving. The [repository] has estimated their additional cost to archive the data at [insert euro amount]. This fee appears in the budget for this application as well.”

#### 5.2.2.2 Costing for roles & responsibilities (when applicable):

Responsibilities & Resources

- Named person responsible for implementation of the DMP
- Named person responsible for each data management activity
- Hardware and software required (any that is additional to existing institutional provision)
- Additional specialist expertise or training required
- Charges to be applied by data repositories
- Person responsible for DMP implementation
- Hardware required

#### 5.2.2.3 Roles & Responsibilities example

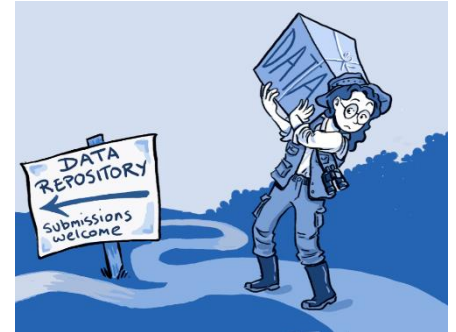
- “The project will assign a qualified data manager certified in disclosure risk management to act as steward for the data while they are being collected, processed, and analyzed.”



- “All research data collected as part of this project is owned by the Institution. The Principal Investigator (PI) of this project will take responsibility for the collection, management, and sharing of the research data.”
- “Day-to-day quality assessment will be the responsibility of the Lab Director who in turn is overseen by the Project Director.”

### 5.3 Sharing, and preservation

Research data remain a valuable resource, even after the life of the research project for which they were collected and used. Sharing research data enables future researchers to open up new lines of enquiry without the duplication of effort involved in collecting the data again. The decision to share your data will in turn require the consideration of a number of issues relating to their subsequent discovery, access and future use. Understanding the issues involved in data sharing will make you a more knowledgeable consumer of secondary sources of data.



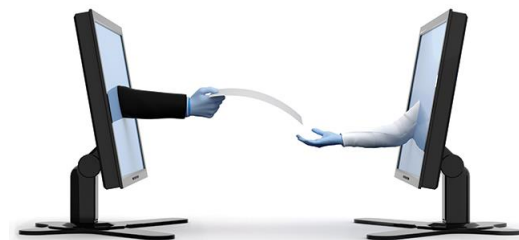
#### 5.3.1 Sharing your research data

Many research funders view research data as a public good that should be openly available to the academic community and preserved for future reuse. In principle, publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner. Where it is possible or permissible to share your research data there are significant benefits to be gained.

However, there are numbers of reason why you may not wish or be able to share your research data. Not all of the reasons may be overcome. However, working through these issues and challenges when developing your data management plan at the beginning of your project will enable you to determine whether you can or should share your data.

#### 5.3.2 Data sharing: Benefits

Researchers devote a large amount of physical and intellectual effort to collect, manage, collate, and analyze their data before publishing their results. Many of these datasets have significant value beyond the usage for the original research, and sharing the data can be seen as beneficial in a number of ways.



##### 5.3.2.1 Scientific integrity

Publishing research data and citing its location in published research papers allows other to replicate, validate or build upon your results thus improving the scientific record by encouraging scientific enquiry and debate. Openly sharing research data also encourages the improvement and validation of research methods and minimizes the need for data re-collection.

##### 5.3.2.2 Funder requirements

Many funding bodies and research councils have adopted research data sharing policies and mandate or encourage researchers to share data and outputs to avoid duplication of effort and reduce data collection costs.

##### 5.3.2.3 Impact

Others who reuse your data and cite it in their own research help to raise interest in your research and increase your impact within your field and beyond.



In their paper “The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data”, Pienta, Alter and Lyle (2010, [doi.org/10.3886/ICPSR29941.v1](https://doi.org/10.3886/ICPSR29941.v1)) examined the extent to which social science research data are shared and assessed whether data sharing affects research productivity tied to the research data themselves. They found that data sharing resulted in a marked increase of both primary and secondary publications.

#### 5.3.2.4 Collaboration

Data sharing may lead to new collaborations between data users and data creators. Sharing data can often lead to improvements, such as corrections in the documentation, or the combination or comparison with other datasets, leading to new information.

#### 5.3.2.5 Innovation

Data created for one research purpose may be re-invented or re-interpreted for future unrelated research. Data sharing and reuse across borders and disciplines can also promote innovation by potential new data users.

Many see 'Open data' as an engine for innovation. Open data advocates are interested in harnessing new technologies for combining, analyzing and visualizing data, often in web environments, to create new forms of data or knowledge.

*“The coolest thing to do with your data will be thought of by someone else.”*  
- Rufus Pollock, Cambridge University and Open Knowledge Foundation

#### 5.3.2.6 Preservation for your own future use

Some research data will be unique and cannot be replaced if destroyed or lost. By preparing your data for sharing with others, you will benefit by being able to identify, retrieve, and understand the data yourself after you have lost familiarity with it, perhaps several years hence.

#### 5.3.2.7 Teaching

Your data may be useful for students or new generations of researchers to learn how to collect and analyze similar types of data. Attempting to reproduce published results by analyzing an existing dataset is an excellent way for students to learn methods.

#### 5.3.2.8 Public record

There is a growing movement for making publicly funded research including resultant research data available to the public, in line with the [OECD](#) principles and guidelines for access to research data from public funding. Thus sharing research data maximizes transparency and accountability.

### 5.3.3 Data sharing: Barriers

Not all data can or should be shared. The following are some of the perceived barriers to data sharing, with some suggested solutions.

#### 5.3.3.1 Financial

It may well be that your data has a financial value attached. Thus, it may not be initially advisable to share your data, even where a license or terms and conditions for use are attached. You may, therefore, need to consult appropriately (e.g. with your institutional research office) within your institution on this issue.



### 5.3.3.2 Confidentiality & Sensitive data

The sensitivity of personal information of human subjects will be readily understood and cannot be overstated. In addition, it will be subject to your own written consent and the institutional ethics codes that govern your research activity. However, it is feasible to anonymize such data so that it may be shared safely and freely with others, without violating any of the above.

### 5.3.3.3 Ownership

You will also need to give serious thought to the matter of just who actually owns your data. For example, it may be owned, in whole or part, by other authors or commercial entities. In some cases, you may not be permitted to share it. At the very least, you will need to consult with collaborators or commercial partners to ensure that you have the necessary permissions.

## 5.3.4 Access, Sharing and Privacy Examples

- “Data will be posted on a website within three months of the grant closing. Data will be contributed to X public database. Data will be submitted to supplementary materials sections of peer-reviewed journals.”
- “Data will be available and cited in publication. Researchers will be able to contact the Principal Investigators (PI) for access to data. Data will be maintained in an open XML format to enable open reuse of the data.”
- “Our project will generate a large volume of data, some of which may not be appropriate for sharing since it involves a small sample that is not representative. The investigators will work with staff of the [repository] to determine what to archive and how long the deposited data should be retained.”
- “X and third party copyright will be protected. The PI will be responsible for ensuring that all project members are aware of the ownership of data and who may access them under which conditions. Online access to the data will be password protected.”
- “This project will generate data linked to administrative records, so the data will be distributed through a restricted data use agreement managed by [repository]. Through this mechanism, users will apply to use these files, create data security plans, and agree to other access controls.”
- “The principal investigators on the project and their institutions will hold the intellectual property rights for the research data they generate but will grant redistribution rights to [repository] for purposes of data sharing.”
- “Our research group has been trained in human subjects’ protection and only trained project staff operating under the IRB approval for the project will have access to the confidential individually identifiable data, and all data will be aggregated or anonymized for publication.”
- “The following language will be used in the informed consent: The information in this study will only be used in ways that will not reveal who you are. You will not be identified in any publication from this study or in any data files shared with other researchers. Your participation in this study is confidential. Federal or state laws may require us to show information to university or government officials [or sponsors], who are responsible for monitoring the safety of this study.”
- “For this project, the principal investigators will request expedited IRB review compliant with procedures established by the [University] campus IRB. Research activities envisioned present no more than minimal risk to human subjects.”
- “During data analysis, the data will be accessible only by certified members of the project team. The research project will remove any direct identifiers in the data before deposit with [repository].”



## 5.4 Making Data “Findable”, Including Provision for Metadata

Make your data findable by ensuring it:

- Has a persistent identifier (Cf. [A persistent identifier \(PID\) to your dataset](#))
- Has rich metadata (Cf. [Metadata](#))
- Is searchable and discoverable online
- Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
- What naming conventions do you follow? (Cf. [Naming conventions for research data files](#))
- Will search keywords be provided to optimize possibilities for reuse?
- Do you provide clear version numbers? (Cf. [Versioning](#))
- What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

<b>Documentation &amp; metadata</b>	<i>A list of the needed information in order for the data to be read and interpreted in the future</i>
	<ul style="list-style-type: none"> <li>• How you plan to collect or create this documentation and metadata</li> <li>• The metadata standards you will use</li> </ul>
	Questions to be asked
	<ul style="list-style-type: none"> <li>• How documentation is to be collected</li> <li>• How metadata is to be created</li> <li>• Metadata standards</li> <li>• Information needed for reuse of your data</li> </ul>
	Some examples of data documentation:
	<ul style="list-style-type: none"> <li>• Laboratory notebooks (Cf. <a href="#">Documentation example Electronic Laboratory Notebooks</a>) &amp; experimental protocols</li> <li>• Questionnaires, codebooks, data dictionaries</li> <li>• Software syntax &amp; output files</li> <li>• Information about equipment settings &amp; instrument calibration</li> <li>• Database schema</li> <li>• Methodology reports</li> <li>• Provenance information about sources of derived data</li> </ul>
	Add detailed description for collections or files. E.g. what is in a file? Where did it come from? How could it be retrieved if needed?
<b>Data sharing</b>	<ul style="list-style-type: none"> <li>• Step to be taken to maximize the data’s discoverability</li> <li>• Any conditions or restrictions on sharing the data and whether these will be set out in a data sharing agreement</li> <li>• Mechanism for sharing. E.g. via a repository, direct correspondence or other arrangement</li> <li>• Timing of publication</li> <li>• Timing of data publication</li> <li>• Arrangements if any to obtain a persistent identifier for the data</li> <li>• Discoverability</li> <li>• Persistence identifier process</li> <li>• Condition on sharing</li> </ul>



## 5.4.1 Data ID, data citation and impact

### 5.4.1.1 Standard identification mechanism

Data citation is the practice of providing a reference to data in the same way as researchers routinely provide a bibliographic reference to other scholarly resources.

A wide range of research outputs such as data sets, software, blog posts, presentations, tweets, etc. may determine the impact of your research. Being able to cite such research outputs is important for building a culture where all types of research outputs count.

### 5.4.1.2 Citing your data

Persistent identifiers ensure future access to unique published digital objects, such as a text or data set. Persistent identifiers are assigned to data sets by digital archives.

For data products to be uniquely identifiable and attributable to their data creator's two types of identifiers are recommended:



#### A persistent identifier (PID) to your dataset

- The publication of data sets is becoming more and more important as a citable contribution to research. To become citable, you need to make sure that your datasets get a unique, persistent identifier. The Digital Object Identifier (DOI) is a well-known identifier in academia. Having a PID is an important aspect in making sure your data meet the F (Findability) and A (Accessibility) in FAIR data management.
- Persistent identifiers (PIDs) are important because they unambiguously identify your data and facilitate data citation. An example of a PID is a Digital Object Identifier (DOI). When depositing your data in a repository, make sure you select a repository that assigns a persistent identifier (for example [Zenodo](#)).

#### A persistent author identifier

- To make your research results even more connected you can create your personal persistent author identifier. The [ORCID](#) provides such a persistent digital identifier, distinguishing you from every other contributor and supporting automated linkages among all your professional activities. By creating and using an ORCID identification, you will be able to present all your growing work through one channel.

### 5.4.1.3 How to cite your data

- **Deposit your data in a data repository:** when you deposit your data in a (trusted) data repository, automatically, a persistent identifier is often assigned to your data sets.
- **Register for an ORCID iD**
- **Check how FAIR your data are**
- **Include persistent identifiers as a variable:** Include the persistent identifier to your dataset as a variable in your data file. For example, the database from the ISSP 2015 on Work Orientations (GESIS, n.d.) includes the following variable: name of the variable: DOI; variable label: "Digital Object Identifier". It has the same value for all the cases: [doi:10.4232/1.12848](https://doi.org/10.4232/1.12848). The link goes directly to the metadata at the GESIS data archive.

### 5.4.1.4 The importance of data citation

Citation is a special kind of descriptive metadata. As every scientist knows, in the course of publishing research it is essential to cite one's sources. The reference list at the end of a publication is crucial to establishing the credibility of the author's claims, as well as the provenance of a particular theory or



school of thought. This is no less true for datasets than for publications, but somehow the practice of proper data citation has been slow to take root in academic culture. All too often we see charts in articles such as, “Source: OECD” without any further information in the reference list to help a reader look up the original data.

Although this raises more issues than can be discussed here, a number of organizations and initiatives have attempted to rectify this by gathering evidence, publishing guidance and articles, applying pressure to bibliographic style authorities, and agreeing universal principles. In 2014, a group called [Force 11](#), which believes that research data should be treated as 'first class research objects' along with publications, issued a [Joint Declaration of Data Citation Principles](#), which has been endorsed by a number of scientific bodies and publishers as well as individuals.

#### 5.4.1.5 Why cite data?

“Data are a vital part of the scientific research process and proper citation should be a significant feature of research publications.”

Data citation:

- acknowledges the author's sources
- makes identifying data easier
- promotes the reproduction of research results
- makes it easier to find data
- allows the impact of data to be tracked
- provides a structure which recognizes and can reward the data creator”

By providing a data citation for your work, you make it easier for others using your data to cite it accurately in their work.

#### 5.4.2 Data paper

A data paper is a publication that is designed to make other researchers aware of data that is of potential use to them for scientific and educational purposes. [Open Health Data](#) publishes data papers, which provide a concise description of a dataset and where to find it.

Data papers can describe deposited data from studies that have not been published elsewhere (including replication research) but also from studies that have previously been published in another journal. As such, the data paper describes the methods used to create the dataset, its structure, its reuse potential, and a link to its location in a repository.

It is important to note that a data paper does not replace a research article, but rather complements it. When mentioning the data behind a study, a research paper should reference the data paper for further details. The data paper similarly should contain references to any research papers associated with the dataset.

For more information, please check the following link: <https://openhealthdata.metajnl.com/about/>

#### 5.4.3 Organizing data, naming convention and versioning

Naming records consistently, logically and in a predictable way, will distinguish similar records from one another at a glance, and by doing so, will facilitate the storage and retrieval of records which will enable users to browse file names more effectively and efficiently.







Naming records according to agreed conventions should also make file naming easier for colleagues because they will not have to re-think the process each time.

#### 5.4.3.1 *Best practice*

Research data files and folders need to be labelled and organized in a systematic way so that they are both identifiable and accessible for current and future users.

There are three main criteria to consider regarding the naming and labelling of research data files, namely:

- **Organization:** important for future access and retrieval, and needs to take into account the file naming constraints of the system where the file is located
- **Context:** this could include content specific or descriptive information, independent of where the data are stored.
- **Consistency:** choose a naming convention and ensure that the rules are followed systematically by always including the same information (such as date and time) in the same order (e.g. YYYYMMDD) Cf. [Naming conventions for research data files](#)

A number of common elements should be considered when developing a file naming strategy, including:

- Version number
- Date of creation
- Name of creator
- Description of content
- Name of research team/department associated with the data
- Project number or acronym

Moreover:

- Do not use generic file names that may conflict when moved from one location to another. Ensure filenames are independent of location and if you work on more than one computer ensure that your files are synchronized.
- Consider how scalable your file naming policy needs to be e.g. if you want to include the project number, do not limit your project number to two digits, or you can only have ninety-nine projects.
- Do not keep files in proprietary format. Export them in a common format when possible.
- Describe your data clearly. Do not use very short names (E.g. "S") for column header for instance.

#### 5.4.3.2 *Naming conventions for research data files*

In collecting material, you obviously have to organize it for the particular project, but most research, links from one project to another. Folder structuring and file naming is very important in case you want to be able to draw in the material, you have collected from one project to use it in another. The way to do this is partly through labelling the files in particular folders, giving them logical names and structures and ensuring that they have some architecture, some overarching principles. File and folder naming conventions are key to maintaining well-organized electronic directory and drive structures. In most cases, the policy for naming a file is left to individuals or to groups of individuals.

There are a number of easy-to-follow rules when naming data files:

- Keep file names short and relevant, generally about 25 characters is a sufficient length to capture enough descriptive information for naming a data file



- Do not use special characters in a filename such as **& \* % \$ £ ] { ! @** as these are often used for specific tasks in different operating systems
- Use underscores instead of full stops or spaces because, like special characters, these are parsed differently on different systems.
- The filename should include as much descriptive information that will assist identification independent of where it is stored.
- If including dates, format them consistently.
- Where possible, use file extensions to reflect accurately the software used to create the file. E.g. use **.por** for SPSS portable files, **.xls** or **.xlsx** for Excel files, **.ssd** or **.sas7bdat** as appropriate for SAS files, **.txt** for text files, etc.

#### 5.4.3.2.1 Naming files by chronology

If using a date use the format Year-Month-Day: YYYY-MM-DD or YYYY-MM or YYYY-YYYY. This will maintain chronological order of your files.

Ex: Files using this naming convention (**2006-03-24\_Attachment**) are easy to distinguish from one another, easier to browse and locate chronologically.

#### 5.4.3.2.2 Descriptive file naming

Keep file names short and relevant using sufficient characters to capture enough descriptive information.

- Example of good file naming: use **2010-08-11\_Bob\_30\_Birthday\_V1.png** where the filename represents the content more accurately. Using a version number convention also makes it easier to distinguish from other versions of the same file.
- Example of bad file naming: **Bob\_Latest\_Birthday\_110810\_old version.png** is a bad file naming since the date is ambiguous, and there could be a number of old versions. Provide an accurate description of the content of the file where possible.
- Another example of bad file naming: **Canon\_3776438656.raw**. This is an application or instrument generated filename lacking descriptive or context-specific information.

#### 5.4.3.2.3 Batch renaming of automatically generated files

Although all operating systems have in-built tools for managing files, there are software tools that can organize research data files and folders in a consistent and automated way through batch renaming (also known as mass file renaming or bulk renaming). Batch renaming software exists for most operating systems.

There are many situations in which batch renaming may be useful, such as:

- Where images from digital cameras are automatically assigned base filenames consisting of sequential numbers
- Where proprietary software or instrumentation generate crude, default or multiple filenames
- Where files are transferred from a system that supports spaces and/or non-English characters in filenames to one that does not (or vice versa). Batch renaming software can be used to substitute such characters with acceptable ones.

#### 5.4.3.3 Versioning

It is important to identify and distinguish versions of research data files consistently. This ensures that a clear audit trail exists for tracking the development of a data file and identifying earlier versions when needed. Thus, you will need to establish a method that makes sense to you and indicates the version of your data files.



- A common form for expressing data file versions is to use ordinal numbers (1,2,3 etc.) for major version changes and decimals for minor changes e.g. **v1, v1.1, v2.6**
- Beware of using confusing labels: **revision, final, final2, definitive\_copy** as you may find that these accumulate
- Record every change irrespective of how minor that change may be
- Discard or delete obsolete versions (whilst retaining the original 'raw' copy)
- Use an auto-backup facility (if available) rather than saving or archiving multiple versions
- Turn on versioning or tracking in collaborative documents or storage utilities such as **Wikis, GoogleDocs** etc.
- Consider using version control software e.g. **Subversion, TortoiseSVN**
- Some structured examples of maintaining version control [document name] [version number] [status: draft/final]:
  - Bob\_interview\_July2010\_V1\_DRAFT
  - Lipid-analysis-rate-V2\_definitive
  - 2001\_01\_28\_ILB\_CS3\_V6\_AB\_edited

#### 5.4.3.3.1 Why should you discard or delete obsolete versions of data files?

Too many similar or related files may be confusing, both to yourself and to anyone else wanting to access or use your data. You may think that you know which data file is which but that may not always be the case as time passes and the number of different file versions increase. It is easier to maintain a manageable number of versions with a clear naming structure. As long as the original 'raw' or definitive copy is retained and processing is well documented, the intermediate working files can and should be discarded.

#### 5.4.3.3.2 What advantages are there to a consistent approach to expressing data file versions?

A consistent approach to expressing data file versions means that you can easily identify the most recent copy of a file or a final version of a file without having to open each file (which would also require a good memory as to what was changed in the file and when!). File sharing can be made easier when a common file naming or versioning procedure is used with research colleagues or collaborators.

#### 5.4.3.4 Summary

- Research data files and folders need to be labelled and organized in a systematic way so that they are both identifiable and accessible for current and future users.
- Naming records according to agreed conventions should make file naming easier for colleagues because they will not have to re-think the process each time.
- One benefit of consistent research data file labelling is that files are not accidentally overwritten or deleted.
- It is important to consistently identify and distinguish versions of research data files. This ensures that a clear audit trail exists for tracking the development of a data file and identifying earlier versions when needed.

### 5.4.4 Documentation

#### 5.4.4.1 Why document your data?

While digital data by definition are machine-readable, understanding their meaning is a job for human beings. The importance of documenting your data during the collection and analysis phase of your research cannot be overestimated.

- **Help yourself:** You may be on intimate terms with your dataset while you are collecting and analyzing it, but remembering that the variable “**sglmemgp**” means single member of group,



or the exact procedure you used to transform or derive particular variables, could potentially become difficult months or years later;

- **Help others:** There are many reasons other people may want to examine or use your data to understand your findings, to verify your findings, to review your submitted publication, to replicate your results or to design a similar study.

#### 5.4.4.2 *Examples of data documentation*

Some examples of data documentation are:

- laboratory notebooks & experimental protocols
- questionnaires, codebooks, data dictionaries
- software syntax and output files
- information about equipment settings & instrument calibration
- database schema
- methodology reports
- provenance information about sources of derived or digitized data

Data documentation demonstrates adherence to standards of good practice, ethical integrity, and compliance with contractual provisions. They can play an important role in supporting claims relating to intellectual property developed by researchers, and even defending claims of scientific fraud.

Therefore, thorough and effective management of laboratory data and the routine documentation of all lab procedures is a highly important responsibility for all laboratory researchers.

##### 5.4.4.2.1 *Documentation example: Electronic Laboratory Notebooks*

Note on laboratory notebook: for laboratory researchers, laboratory notebooks are crucial components of data management.

#### **Organizing your data**

- Data and documents have a unique ID assigned to them at time of creation, which stays with them regardless of changes to the filename, preventing mix-ups with similarly named files.
- All files have appropriate metadata including creator, creation date, last modification, file type etc. assigned to them at time of creation.
- Files may be tagged according to their content, grouping them together and facilitating their rapid identification and retrieval during searches.
- Powerful searching capabilities allow files to be quickly identified via filters, tags, unique IDs, modification dates, creator etc.
- Files may be organized into notebooks and folders making them easy to locate and browse.
- Files and datasets may be linked within documents allowing results and methodologies for different experiments to be kept together.
- Automatic recording of document versions helps ensure that research is recorded in a way that supports protection of intellectual property.

#### **Sharing data with others**

- Documents, notebooks, folders and files may be shared in a controlled way between users enabling collaboration within and between groups.
- Data can be accessed 24/7.
- Principal investigators and lab administrators can control user access and privileges to documents and data (e.g. read only or editing capabilities).

#### **Integration with other tools**

- Users can access and make links to data on commercial file sharing apps like Dropbox and One Drive, and institutional and lab file storage facilities (Cf. [Storage](#))



- Users can export datasets directly to data repositories (Cf. [Sharing, and preservation](#))
- Through an API, other research tools can be integrated with the electronic lab notebook.

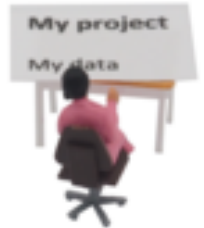
Example of a commercially available electronic laboratory notebook: [RSpace](#)

### 5.4.4.3 Levels of documentation

#### 5.4.4.3.1 Project-level documentation

Project-level documentation explains the aims of the study, what the research questions/hypotheses are, what methodologies were being used, what instruments and measures were being used, etc.

If a dataset is shared, a detailed technical report will need to be included for the user to understand how the data were collected and processed. You should also provide a sample bibliographic citation to indicate how you would like secondary users of your data to cite it in any publications, etc.



#### **For what purpose was data created?**

Describe the project history, its aims, objectives, concepts and hypotheses, including

- The title of the project
- Subtitle
- Author(s)/creator(s) of the dataset
- Other co-workers and their roles (person, research group or organization that participated in the study and their roles);
- The institution of the author(s)/creator(s)
- Funders
- Grant numbers
- References to related projects
  - Publications from the data

#### **What does the dataset contain**

Describe what is in a dataset

- Kind of data (interviews, images, questionnaires, etc.)
- File size (in bytes), file format of the data files and relationships between files
- Description of data file(s): version and edition, structure of the database, associations, links between files, external links, formats, compatibility

#### **How was data collected?**

Describe how the data was acquired

- The methodology and technique used in collecting and creating the data;
- Description of all the sources the data originate from (What is the subject of study? E.g. periodicals, datasets created by others?) together with an explanation of how and why it got to the present place (provenance)
- The methods/modes of data collection (for example):
  - The instruments, hardware and software used to collect the data
  - Digitization or transcription methods
  - Data collection protocols
- Sampling design and procedure
- Target population, units of observation.



### **Who collected the data and when?**

Describe the

- Data collector(s)
- Date of data collection
- Geographical coverage of the data (e.g. Nation).

### **How was the data processed?**

Describe your workflow and specific tools, instruments, procedures, hardware/software or protocols you might have used to process the data, like

- Data editing, data cleaning
- Coding and classification of data.

### **What possible manipulations were done to the data?**

Describe if and how the data was manipulated or modified

- Modifications made to data over time since their original creation and identification of different versions of datasets
- Other possible changes made to the data
- Anonymization
- For time series or longitudinal surveys: changes made to methodology, variable content, question text, variable labelling, measurements or sampling.

### **What were the quality assurance procedures?**

Describe how the quality of the data has been assured

- Checking for equipment and transcription errors
- Quality control of materials
- Data integrity checks
- Calibration procedures
- Data capture resolution and repetitions
- Other procedures related to data quality such as weighting, calibration, reasons for missing values, checks and corrections of transcripts, transformations.

### **How can the data be accessed?**

Describe the use and access conditions of the data

- Where the data can be found (which data repository)
- Permanent identifiers (PID)
- Access conditions such as embargo
- Parts of the data that are restricted or protected
- Licenses
- Data confidentiality
- Copyright and ownership issues
- Citation information.

#### **5.4.4.3.2 File or database level**

A **readme.txt** file is the classic way of accounting for all the files and folders in a project.

- How all the files (or tables in a database) that make up the dataset relate to each other?



- What format are they in?
- Whether they supersede or are superseded by previous files?

### 5.4.4.3.3 Data-level documentation

Data-level or object-level documentation, provides information at the level of individual objects such as: pictures, interview transcripts or variables in a database. You can embed data-level information in data files, for example: in interviews, it is best to write down the contextual and descriptive information about each interview at the beginning of each file. Then, Quantitative data variable and value names can be embedded within the data file itself.

#### 1.1.1.1.1.1 Data-level documentation for quantitative data

For quantitative data, document the following:

- **Information about the data file**  
Data type, file type and format, size, data processing scripts.
- **Information about the variables in the file**  
The names, labels and descriptions of variables, their values, a description of derived variables or, if applicable, frequencies, basic contingencies etc. The exact original wording of the question should also be available. Variable labels should:
  - Be brief with a maximum of 80 characters;
  - Indicate the unit of measurement, where applicable;
  - Reference the question number of a survey or questionnaire, where applicable.



Example of a variable and variable label

- Names, labels and descriptions for variables, records and their values
- Description of the missing values at each variable
- Description of the weighting variable
- Explanation or definition of codes and classification schemes used

Storing documentation: Whenever possible, embed data documentation within a file as in the following screenshot:

	Name	Type	Width	Decimals	Label	Values	Missing
175	quala10	Numeric	2	0	Which of the qualifications on this card do you have? 10	{-9, No ans...	-99 - -1
176	activb	Numeric	2	0	Activity status for last week	{-9, No ans...	-99 - -1
177	empstat	Numeric	2	0	Manager/Foreman	{-9, No ans...	-99 - -1
178	everjob	Numeric	2	0	Ever had paid employment or self-employed	{-9, No ans...	-99 - -1
179	ftptime	Numeric	2	0	Full-time or part-time	{-9, No ans...	-99 - -1
180	howlong	Numeric	2	0	How long have you been looking	{-9, No ans...	-99 - -1
181	wkstrt2	Numeric	2	0	Able to start work within 2 weeks (Government training scheme)	{-9, No ans...	-99 - -1
182	wklook4	Numeric	2	0	Looking paid work/govt scheme last 4 weeks	{-9, No ans...	-99 - -1
183	nemplee	Numeric	2	0	Number employed at place of work	{-9, No ans...	-99 - -1
184	nssec	Numeric	5	1	NS-SEC - long version (harmonised)	{-9.0, No a...	-99.0 - -1.0
185	othpaid	Numeric	2	0	Ever had other employment (waiting to start work)	{-9, No ans...	-99 - -1
186	payage	Numeric	3	0	Age when last had a paid job	{-9, No ans...	-99 - -1
187	paylast	Numeric	4	0	Year left last paid job	{-9, No ans...	-99 - -1
188	paymon	Numeric	2	0	Month last left paid job	{-9, No ans...	-99 - -1
189	sclass	Numeric	2	0	Social Class	{-9, No ans...	-99 - -1
190	seg	Numeric	2	0	Socio-Economic Group	{-9, No ans...	-99 - -1
191	sneemlee	Numeric	2	0	Self employed, how many employees	{-9, No ans...	-99 - -1
192	age	Numeric	3	0	Age last birthday	{-9, No ans...	-99 - -1



#### 1.1.1.1.1.2 Data-level documentation for qualitative data

Background and contextual information and participant details of interviews, observations or diaries can be described at the beginning of a file as a header or summary page.

For qualitative data, document the following

- Textual data file (for example, interview)
  - Key information of participants such as age, gender, occupation, location & relevant contextual information
  - For qualitative data collections (for example image or interview collections) you may wish to provide a data list that provides information that enables the identifying and locating of relevant items within a data collection
    - The list contains key biographical characteristics and thematic features of participants such as age, gender, occupation or location, and identifying details of the data items;
    - For image collections, the list holds key features for each item;
    - The list is created from an initial list of interviews, field notes or other materials provided by the data depositor.
- Audiovisual data files: For some types of data (image, audio or video files), the file format does not always allow recording background information in the beginning of the data file. In such cases, the best practice is to store background information in a manually created data list or a separate text file: a data list, which accompanies the data collection.
  - Provide the following information on each image: creator, date, location, subject, content, copyright, keywords, equipment used (Metadata)
  - Some image files have embedded technical metadata (You may use tools to extract technical metadata from images, such as [ExtractMetadata.com](http://ExtractMetadata.com))
- Periodicals, magazines, journal articles
- Among materials you use for qualitative data analysis, there may be online periodicals, magazines or journal articles. The information about all such resources must be kept in separate files:
  - Material collected from online periodicals: save references to web resources, like URLs, and do not forget they may change over time. To be sure information isn't lost, articles should be copied into a word processing program
  - Materials from periodicals: When articles, photographs and other material are collected from periodicals for research purposes, bibliographic information should be carefully detailed (author(s), title, date of publication etc.)
  - When you analyses articles, make a list of them, sort them alphabetically or chronologically in the order they were analyzed in the course of research

#### 5.4.4.3.4 Variable or item level

The key to understanding research results is, knowing exactly how an object of analysis came about. Not just, for example, a variable name at the top of a spreadsheet file, but the full label explaining the meaning of that variable in terms of how it was operationalized.

#### 5.4.4.4 Storing documentation:

- Write the documentation into a separate, well-structured file and associate that with the data file. You may use the same filename stem in order to strengthen the file-metadata association. For example **20130311\_interviews\_audio**, **20130311\_interviews\_trans**, **20130311\_interviews\_image** and **20130311\_interviews\_metadata**. The latter part of the name might be used to convey the specifics of the file. In this case “audio” means audio tape and “trans” a transcription of the audio tape.





- Data-level documentation can be embedded within a data file. For example, in interviews, it is best to write down the contextual and descriptive information about each interview at the beginning of each file;
- If you have a large amount of metadata or large amounts of data that will need metadata, you can use a standard specific database for the purpose (such as the [DDI Codebook](#) (DDI Alliance, 2017a)).

#### 5.4.4.5 Summary

There are many reasons why you need to document your data:

- Help you remember the details later
- Help others understand your research, verify your findings, review your submitted publication, replicate your results, or even archive your data for access and reuse.

Research data need to be documented at various levels:

- Project level
- File or database level
- Data-level
- Variable or item level.

Some examples of **data documentation** are:

- Laboratory notebooks (Cf. Documentation example Electronic Laboratory Notebooks)
- Methodology reports
- Questionnaires
- Software syntax.

Laboratory notebooks, for example, play an important role in supporting claims relating to intellectual property developed by University researchers, and even defending against claims of scientific fraud.

## 5.4.5 Metadata

### 5.4.5.1 What does metadata mean?

The term metadata is commonly defined as “data about data,” information that describes or contextualizes the data. The difference between documentation and metadata is that the first is meant to be read by humans and the second implies computer processing (though metadata may also be human-readable). In other terms, Metadata are descriptors that facilitate cataloguing data and data discovery. Metadata is intended for reading by machines.

The importance of metadata lies in the potential for machine-to-machine interoperability, providing the user with added functionality, or “actionable” information. When data is submitted to a trusted data repository, the archive generates machine-readable metadata, which helps to explain the purpose, origin, time, location, creator, terms of use, and access conditions of research data. The metadata describing your data supports findability, citation and reuse.

Rich metadata provides important context for the interpretation of your data and makes it easier for machines to conduct automated analysis. Follow standard metadata schemes, general ones such as Dublin Core (Cf. [Dublin Core Metadata Element Set](#)) or discipline specific.



Minimum information required in metadata for data generated during the OHEJP program:

- OHEJP
- Project name

### 5.4.5.2 Types of metadata

Three broad categories of metadata are:

- **Descriptive:** common fields such as title, author, abstract, keywords that help users to discover online sources through searching and browsing.
- **Administrative:** preservation, rights management, and technical metadata about formats.
- **Structural:** how different components of a set of associated data relate to one another, such as a schema describing relations between tables in a database.



Metadata may not be required if you are working alone on your own computer, but become crucial when data are shared online. Your data management plan should determine whether there is a need to apply metadata descriptors or tags at some point during your project. Metadata help to place your dataset in a broader context, allowing those outside your institution, discipline, or software environment to understand how to interpret your data.

<i>Descriptive</i>	<i>Subject, title, data creator, funder</i>
<i>Administrative</i>	<i>File formats, Rights statement, Version of data, Timestamp of transaction</i>
<i>Structural</i>	<i>data dictionary, taxonomy, database schema, variable list</i>

### 5.4.5.3 Metadata creation

Check out [The Dublin Core Metadata Generator](#) and see how metadata elements are converted into a machine-readable file in **.xml**. In addition, if you enjoy working with **.xml** schemas, get started in creating a codebook to accompany your dataset with the DDI codebook.

When you submit your dataset at a (trusted) data repository, machine-readable metadata will be added. (Cf. [Data repository](#))

### 5.4.5.4 Standard for Metadata examples

- “Metadata will be generated to describe the data generated in X format and will be stored alongside the data. X metadata standards will be applied during the creation of the metadata.”
- “The clinical data collected from this project will be documented using CDISC metadata standards.”

### 5.4.5.5 Summary

The term metadata is commonly defined as data about data. The importance of metadata lies in the potential for machine-to-machine interoperability, providing the user with an enhanced version of published data.

## 5.5 Making data openly “Accessible”

Make your data accessible by ensuring it:

- Is retrievable online using standardized protocols
- Has restrictions in place if necessary



Remember that not all data has to be made open. Data can be restricted and still be FAIR. However, if access is allowed, data should be retrievable without the need for specialized protocols. In addition, even if the full content is not made openly available, the data must be as findable as possible.

Data should be “**As Open as Possible, As Closed as Necessary**” (Cf. [H2020 AGA Annotated Model Grant Agreement](#))

- Which data produced and/or used in the project will be made openly available by default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
- Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for *opting out*.
- How will the data be made accessible (e.g. by deposition in a repository)?
- What methods or software tools are needed to access the data?
- Is documentation about the software needed to access the data included?
- Is it possible to include the relevant software (e.g. in open source code)?
- Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories that support open access where possible.
- Have you explored appropriate arrangements with the identified repository?
- If there are restrictions on use, how will access be provided?
- Is there a need for a data access committee?
- Are there well-described conditions for access (e.g. a machine-readable license)?
- How will the identity of the person accessing the data be ascertained?

### 5.5.1 How open is open?

Many in the open data movement believe that “non-commercial” licenses, those that allow copying except for purposes of profit, are not truly open, and may even damage the emerging “knowledge economy”. Of course, this is a matter of personal choice, unless a contract prevents you from making your data open. But, what if the for-profit company is adding value to the data, producing a valuable product, or providing a valuable service, in the course of earning a living? Would their motivation exist if there was no option to make a profit, however small?

Remember, you can put other restrictions on usage, such as the requirement to be attributed, or the requirement not to change the work in the course of re-using it (no derivatives).

#### 5.5.1.1 What is “linked open data”?

The approach of creating Linked Open Data (LOD) has become increasingly popular since the concept was described by Tim Berners-Lee, the inventor of the World Wide Web, in a speech in 2010. Ed Summers captured the essence of Berners-Lee's speech with his “5 stars of open linked data” which classifies the format of digital data in terms of its accessibility, sustainability and potential for secondary use by others. Summers notes that the effort and importance of getting the first three stars should not be trivialized.

This star system is a very non-technical description of LOD. Implementing it fully requires using a metadata standard called RDF ([Research Description Framework](#)) to mark up your data and create the links to other data in the LOD “cloud”. A search language called [SPARQL](#) is then used to query the RDF tags. Shared vocabularies are also required.

Many consider LOD the gold standard for open data. It is the essence of what *Berners-Lee* calls the ‘Semantic Web’ and is meant to be an evolutionary step from the far less structured Document Web



we now use. However, it does require that everyone ‘plays along’ and formats their data this way to reach critical mass.

- ★ make your stuff available on the web (whatever format)
- ★★ make it available as structured data (e.g. excel instead of image scan of a table)
- ★★★ non-proprietary format (e.g. csv instead of excel)
- ★★★★ use URLs to identify things, so that people can point at your stuff
- ★★★★★ link your data to other people’s data to provide context

### 5.5.2 Security

Data security means ensuring that research data are kept safe from corruption and that access is suitably controlled.

- Risks and how they will be managed;
- Access arrangements;
- Any arrangements, if needed, for safe and secure transfer of data collected in the field;
- Aspects of privacy of sensitive data should be covered;
- Steps to minimize leakage should be highlighted;
- Strategies to protect confidential data should be specified.

It is important to consider the security of your data to prevent:

- Accidental or malicious damage/modification to data.
- Theft of valuable data.
- Breach of confidentiality agreements and privacy laws.
- Premature release of data, which can void intellectual property claims.
- Release before data have been checked for accuracy and authenticity.

You need to consider the following questions for securing your research data:

- How will you manage access arrangements and data security?
- How will you enforce permissions, restrictions and embargoes?
- Other security issues such as sensitive data, off-network storage, storage on mobile devices (laptops, smartphones, flash drives, etc.), policy on making copies of data, etc. where relevant.

Securing research data is part of the issue of information technology security. You should always have up to date **anti-virus** software installed on your office and home computers.

If you have sensitive data that are covered by privacy laws or confidentiality agreements, it is best to store them on a computer that is not connected to any network. If this is not possible, then you can also consider **encrypting** your data or your PC.

However, you should also be aware of physical security. A computer that is not connected to a network is still vulnerable to **theft** and malicious damage/modification to data. Highly sensitive data, regardless of physical storage medium (laptop, USB flash drives, CDs/DVDs), should be stored in a locked room or safe when not in use. Such data should not be stored or transmitted without encryption.



### 5.5.2.1 *What is encryption*

Encryption is the process of converting data into an unreadable code. You must have access to a password or a secret encryption key to be able to read an encrypted file. Encrypting your data will help ensure your data remain safe from disclosure in the event that a laptop or other portable device such as a USB flash drive/memory stick ends up lost or stolen. Be advised, medium and high-risk personal data or business information must be encrypted.

BitLocker (Windows) and File Vault (Mac) are recommended software solution for encrypting data on computers and laptops.

Please note the following considerations before encrypting your data:

- Data encryption is not a substitute for other information protection controls.
- Data encryption is reliant on the creation of a strong password.
- Encrypted data **cannot** be recovered in the event of a failure.
- If the encryption key is lost, the disk image becomes corrupted, or the hard disk fails, any encrypted data will be lost.
- Establish a reliable and secure backup procedure for the data and any related passwords.

One of the things you need to consider when you are using encryption, is who actually has a backup of the password you are using for the encryption? Because actually, if you are the only person who knows that password, you want as few people as possible to know it. On the other hand, if you are the only person who knows it and for some reason you forget it, you have a huge problem yourself. Finally, if for some reason somebody has to, for an emergency purpose, access your data, it is very unlikely to happen without you being available.

### 5.5.2.2 *Encryption for portable devices*

Encryption of laptops is also an obvious step for laptop security. However, for laptop encryption, there is no “one size fits all” solution. Bit Locker, Pretty Good Privacy (PGP) and some other commercial software packages are available.

BitLocker is a full disk encryption feature included with the Ultimate and Enterprise editions of Windows Vista and Windows 7, the Pro and Enterprise editions of Windows 8 and Windows 8.1, the Pro, Enterprise, and Education editions of Windows 10, and Windows Server 2008 and later. It is designed to protect data by providing encryption for entire volumes.

Pretty Good Privacy (PGP) is a data encryption and decryption computer program that provides cryptographic privacy and authentication for data communication. It is often used to sign, encrypt, and decrypt (texts, e-mails, files, directories and whole disk partitions) and to increase the security of digital communications.

If you keep sensitive data on **USB flash drives** it is recommended that you use drives with encryption software. They give protection to the data if the drive is lost, especially important if the data contain personal or commercially sensitive data. Compared with “ordinary” USB flash drives they cost only a modest amount more so are a good buy. Your local IT support service will probably be happy to advice on the suitability of any encrypted flash drive that you are planning on purchasing.

### 5.5.2.3 *Data erasure*

Data should be erased and information with de-identified or anonymized data should be used in future research, once the research including sensitive information is done and the identifiable portions of the data is no more needed. This applies to paper records, as well as electronic records.



File deletion is not enough to ensure that sensitive electronic data are completely removed from a computer system. File deletion removes only the pointers to the disk sectors in which the data reside. Deleted files can be recovered using commonly available software tools.

To ensure the complete destruction of sensitive data, the main options are:

- Data erasure (aka data clearing or data wiping) e.g. removing all data while leaving the hard drive or other storage medium still operable. Current techniques may not, however, be completely successful on solid-state drives and USB flash drives
- Degaussing, which disturbs the magnetic alignment of magnetic storage media and in many cases makes newer media (such as hard drives and tapes) unusable
- Physical destruction, through disintegration, shredding, pulverizing or incineration.

More advice on secure deletion when disposing of a device is available at:

[https://www.youtube.com/watch?time\\_continue=151&v=Ylkg7-JOYX8](https://www.youtube.com/watch?time_continue=151&v=Ylkg7-JOYX8)

### 5.5.3 Storage

Your data are the lifeblood of your research. If you lose your data, recovery could be slow, costly or worse, it could be impossible. Therefore, through the course of your research, you must ensure that, regardless of format, all your research data are securely stored, backed up and maintained regularly. You should estimate the volume of data required for your project at an early stage, probably while drawing up your data management plan. It is also a good idea to consider including costs for data storage in funding proposals. These may all sound like common sense to you now, but the time will come when you will be glad that you have considered these issues right at the beginning of your research and taken the necessary precautions.

This section should specify how and where you would be storing the project data. It should specify the hardware platform and database software that is used for storage. Research data should be backed up from time to time to mitigate the risk of accidental corruption or deletion. The backup plan and version control mechanism should be specified under this section.

Storage & Backup :

- Storage location: Where (physically) data will be stored
- Backup provision
- Person or team responsible for backup
- Recovery procedures

#### 5.5.3.1 Storage location: Where to store your data?

Not necessarily opening it up, but keeping it somewhere safe for the long-term in a repository that does the following:

- Stores the data safely
- Make sure the data is findable
- Describes the data appropriately (metadata)
- Adds license information

You can deposit data to a general repository (e.g. [Zenodo](https://www.zenodo.org), [Harvard Dataverse](https://dataverse.harvard.edu/)) or a subject-specific repository (e.g. [Dryad](https://www.dryad.org/)). A sub-community One-Health EJP was set up on OpenAIRE platform: <https://www.zenodo.org/communities/ohejp/?page=1&size=20> . Search [www.re3data.org](https://www.re3data.org) if you are looking for a more suitable data repository for your discipline. See a demonstration of searching for research data repositories using the re3data directory (Cf. [Data repository](#)).



You can store your research data on networked drives, personal computers/laptops, and external storage devices.

#### 1.1.1.1.3 Networked drives

It is highly recommended that you store your research data on regularly backed up networked drives. This way you will ensure that your data will be:

- Stored in a single place and backed up regularly.
- Available to you as and when required.
- Stored securely minimizing the risk of loss, theft or unauthorized use.

#### 1.1.1.1.4 Personal computers and laptops

Personal computers (PCs) and laptops are convenient for storing your data while in use. However, they should not be used for storing master copies of your data. Local drives may fail or PCs and laptops may be lost or stolen leading to an inevitable loss of your data.

#### 1.1.1.1.5 External storage devices

External storage devices such as hard drives, USB flash drives (also known as memory sticks, USB keyrings or pen drives), Compact Discs (CDs) and Digital Video Discs (DVDs), can be an attractive option for storing your data due to their low cost and portability. However, they are not recommended for the long-term storage of your data, particularly your master copies:

- Their longevity is not guaranteed, especially if they are not stored correctly, for example CDs, DVDs and magnetic tapes degrade over the long term.
- They can be easily damaged, misplaced or lost.
- Errors writing to CDs and DVDs are common.
- They may not be big enough for all the research data, so multiple disks or drives may be needed.
- They pose a security risk due to their portability.

If you choose to use CDs, DVDs and USB flash drives (for example, for working data or extra backup copies), you should:

- Ensure your master copy is safe and is kept up to date on a networked drive.
- Choose high quality products from reputable manufacturers.
- Follow the instructions provided by the manufacturer for care and handling, including environmental conditions and labelling.
- Regularly check the media to make sure that they are not failing, and periodically 'refresh' the data (that is, copy to a new disk or new USB flash drive).
- Ensure that any private or confidential data is encrypted.

#### 1.1.1.1.6 Summary

Through the course of your research, you must ensure that you store your research data in a secure way and have backup copies in at least three locations that are maintained regularly.

You can store your research data, for example on:

- Your company networked drives (highly recommended)
- Personal computers and laptops (these should not be used for storing master copies of your data)
- In addition, external storage devices (not recommended for the long-term storage of your data, particularly your master copies).



### 5.5.3.2 DMP Data storage and preservation example

- “The research data from this project will be deposited with the institutional repository on the [organization identification].”
- “The research data from this project will be deposited with [repository] to ensure that the research community has long-term access to the data.”
- “By depositing data with [repository], our project will ensure that the research data are migrated to new formats, platforms, and storage media as required by good practice.”
- “In addition to distributing the data from a project Web site, future long-term use of the data will be ensured by placing a copy of the data into [repository], ensuring that best practices in digital preservation will safeguard the files.”
- “[Repository] will place a master copy of each digital file (e.g., research data files, documentation, and other related files) in Archival Storage, with several copies stored at designated locations and synchronized with the master through the Storage Resource Broker.”
- “The data will be processed and managed in a secure non-networked environment using virtual desktop technology.”
- “The data files from this study will be managed, processed, and stored in a secure environment (e.g., lockable computer systems with passwords, firewall system in place, power surge protection, virus/malicious intruder protection) and by controlling access to digital files with encryption and/or password protection. De-identified files will be deposited with [repository] whose security policy has been written according to best practices.”
- “Our research project will generate data from a large national sample. These data will be retained by [repository] as part of their permanent collection.”

### 5.5.4 Backup

Keeping backups is probably the most important data management task. There is a real risk of losing data through hard drive failure or accidental deletion. It is therefore recommended that you keep at least 3 copies of your data on at least 2 different media, keeping storage devices in separate locations with at least 1 off-site, and check that they work regularly. You should also have a policy for maintaining regular backups.

When considering your backup strategy, you need to know:

- How will you back up your data?
- Will all data, or only amended data, be backed up? (A backup of amended data is known as an “incremental backup”, while a backup of all data is known as a “full backup”).
- How often will full and incremental backups be made?
- How long will backups be stored?
- It is important to ascertain the back-up schedule and retention policies of any centralized backup.
- How much hard drive space or how many CDs/DVDs will be required to maintain this backup schedule?
- How will you keep track of different versions of data, especially when backing-up to multiple devices?
- If using versioning software, which software will you use (e.g., Tortoise, Subversion)?
- If the data are sensitive, how will they be stored securely and appropriately, and how will you manage the destruction of identifying data if required (e.g. at the end of your research)?
- What backup services are available that meet these needs and, if none, what alternatives are available?

**Important!** To ensure that your backup system is working properly, you should regularly test it by restoring your data files from your backups to check that you can read them.





### 5.5.4.1 Online backup

Remote, online or managed backup services provide users with an online system for storing and backing up computer files.

Typically, online backup services:

- Allow users to store and synchronize data files online and between computers.
- Employ cloud computing storage facilities (e.g. Amazon S3).
- Provide the first few gigabytes free and users pay for more facilities, including space.

Wikipedia has a thorough comparison of online backup services.

Some examples of online providers are:

#### PROVIDER

<a href="#"><u>Dropbox</u></a>	Dropbox is a Web based file hosting service operated by “Dropbox Inc.”. Dropbox uses cloud computing and synchronization to enable users to store and share files across the Internet. There are both free and paid services, each with varying options. Dropbox offers a relatively large number of user clients across a variety of desktop and mobile operating systems.
<a href="#"><u>Google Drive</u></a>	Google Drive is an online service, which provides 15GB of storage free, allowing files to be accessible from anywhere. A Google account is required to use the Drive service and enables the creation and editing of files directly within the service, providing the means to collaborate and share files with others (including documents, spreadsheets, presentations, forms, and drawings).
<a href="#"><u>Memopal</u></a>	Memopal is a cloud-based storage application and service that enables users to store and synchronize computer files and share files and folders with others using the Internet. Source Wikipedia: <a href="http://en.wikipedia.org/wiki/Memopal">http://en.wikipedia.org/wiki/Memopal</a>
<a href="#"><u>FilesAnywhere</u></a>	FilesAnywhere is one of the first cloud-based storage services to emerge and today continues to offer customers, both consumer and business, a means to back up, edit, sync, collaborate, and share data as well as catalog photos, videos, and music. Source Wikipedia: <a href="http://en.wikipedia.org/wiki/FilesAnywhere">http://en.wikipedia.org/wiki/FilesAnywhere</a>
<a href="#"><u>OneDrive</u></a>	OneDrive is a file hosting service that allows users to upload and synchronize files to a cloud storage and then access them from a Web browser or their local device. It is part of the suite of online services formerly known as Windows Live and allows users to keep the files private, share them with contacts, or make the files public. Publicly shared files do not require a Microsoft account to access. Source Wikipedia: <a href="http://en.wikipedia.org/wiki/OneDrive">http://en.wikipedia.org/wiki/OneDrive</a>
<a href="#"><u>Tresorit</u></a>	Tresorit is an online cloud storage service based in Switzerland and Hungary that emphasizes enhanced security and data encryption. The service offers users 5GB storage free. It has been likened to a high-security alternative to Dropbox. Source Wikipedia: <a href="http://en.wikipedia.org/wiki/Tresorit">http://en.wikipedia.org/wiki/Tresorit</a>
<a href="#"><u>iCloud</u></a>	iCloud is a cloud storage and cloud-computing service from Apple Inc. The service provides its users with means to store data such as documents, photos, and music on remote servers for download to iOS or Windows devices, to share and send data to other users, and to manage their Apple devices if lost or stolen.



### 5.5.4.2 Online backup: advantages & disadvantages

<i>Advantages</i>	<ul style="list-style-type: none"> <li>• No user intervention is required (changing tapes, labeling CDs, performing manual tasks)</li> <li>• Remote backup maintains data off-site</li> <li>• Most provide versioning and encryption</li> <li>• They are multi-platform</li> </ul>
<i>Disadvantages</i>	<ul style="list-style-type: none"> <li>• Many cloud storage services operate on servers physically located outside the European Economic Area (EEA). It is important to consider not using online storage platforms located outside the EEA to store sensitive personally identifiable data. That personal information should not be transferred outside the EEA without adequate protection.</li> <li>• Restoration of data may be slow (dependent upon network bandwidth)</li> <li>• Stored data may not be entirely private (if unencrypted)</li> <li>• Service provider may go out of business</li> <li>• Protracted intellectual property rights/copyright/data protection licenses</li> <li>• Vendor lock-in, e.g. vendor’s proprietary formats may make migration to another vendor complex and expensive</li> </ul>

### 5.5.4.3 Long-term data preservation: the risk of loss

Preserving research data for future reuse can be particularly problematic due to the sheer amount of data being generated as well as the wide variety and complexity of formats and types. Unlike traditional “hardcopy”, objects such as books or publications, where the user has unmediated access to the content (even the simplest piece of digital research data) requires software to render it. Such environments evolve and change at a rapid pace, threatening the continuity of access to the data. Physical storage media, data formats, hardware, and software, all become obsolete over time, posing significant threats to the survival of the content. If you do not archive or manage your data in a systematic way, there is a real danger, and one that increases with the passage of time, that they will effectively be lost.

While paper records can last for centuries, millennia even, digital data, specifically the bits and bytes of which it is comprised, can deteriorate quickly. While the rate of deterioration will vary depending on the storage medium used - magnetic, optical, etc. all media decay over time (“bit rot”), introducing errors that reduce their ability to be read accurately. This is known as “Data Fragility” and here are some suggestions to minimize “bit rot”

- **Refreshment:** move data files onto new storage media well within the projected lifespan of the media.
- **Replication:** by keeping more than one copy of a data file, in different locations, the risk of losing a readable copy over time is reduced.

The pace of technological advance has resulted in ever more obsolescence as new developments overtake and make redundant existing technologies. This problem can occur with computer software when a new version of a software product is incompatible with the superseded versions. In addition, if a software manufacturer ceases production, there may be no alternative newer version capable of running on later operating systems. This is known as “Software Obsolescence” and here are some strategies to overcome it:



- **Migration:** when a new software version has become established, the data file is converted or “migrated” to the new software version or package.
- **Emulation:** recreate the functionality of the obsolete software package on a new operating system.
- **Format conversion:** the most pro-active method is to select a neutral or non-proprietary format that is most easily imported into a number of suitable software programs, or that is based on a universal standard.

Given the issues raised, planning for preservation is essential. The [Open Archival Information Standard Reference Model](#) (OAIS), originally developed by space scientists, is the pre-eminent model for preservation planning and is an International Standards Organization (ISO).

The Digital Curation Lifecycle Model is useful for mapping the planning, preservation, and curation activities associated with a digital object into a lifecycle view.

#### 5.5.4.4 Summary

A preservation plans should include the following:

- Regular back up schedules (including multiple copies in multiple locations)
- Strategies to prevent data loss
- File format migration plans
- Check sums (bit integrity checking)
- Version control
- Data security
- Data storage media
- Copyright and licensing

#### 5.5.5 Data repository

Depositing your research data in a data archive or data repository will facilitate its discovery and preservation. It will also ensure the long-term preservation of your data for future access by you and other researchers. While you may share your data informally by emailing it to requestors or posting it to a website, informal methods of sharing such as these will make it difficult for people to find and access your data both now and in the long-term.

There are many compelling reasons why you should consider depositing your research data into an institutional or national data repository or subject / domain-specific data archive.

The advantages of data deposit within a data repository or data archive:

- Your data are kept safe in a secure environment
- Your data are regularly backed up and preserved for future use
- Your data can be discovered by search engines and included in online catalogues
- The intellectual property rights and licensing of your data are managed
- Access to your data can be administered and usage monitored
- The visibility of your data can be enhanced, thus enabling more use and citation

##### 5.5.5.1 How to select a data repository?

In deciding where to store your data, you may have a number of choices about who will look after it. The choice may be straightforward if you have an established data management facility in your domain or institution, or even within your research group or department. When data preservation standards



or norms exist in your discipline, they should be followed. Your research funder may recommend a data center or self-deposit archive.

In order of preference:

- Use an external data archive or repository already established for your research domain to preserve the data according to recognized standards in your discipline.

Examples of discipline-specific repositories:

- Biological sciences: nucleic acid sequence (eg:European Nuclotide Archive- ENA, GenBank), functional genomics, which bridge disparate research disciplines (European Genome-Phenome Archive – EGA), metabolomics (MetaboLights), proteomics (PRIDE),
  - Modelling: mathematical and modelling resources (BioModels Database, Kinetic Models of Biological Systems – KiMoSys), Network Data Exchange – NDEx),
  - Health Sciences: immunology (ImmPort), pathogen-focused resources (Eukaryotic Pathogen Database Resources – EuPathDB, VectorBase), repositories suitable for restricted data access (Research Domain Criteria Database-RDoCdb).
- If available, use an institutional research data repository, or your research group’s established data management facilities.
  - Use a community-recognized data repository such as [Zenodo](https://www.zenodo.org). Zenodo is a service that enables researchers, scientists, EU projects and institutions to share and showcase multidisciplinary research results (data and publications) that are not part of existing institutional or subject-based repositories. A sub-community One-Health EJP on OpenAIRE platform was set up: <https://www.zenodo.org/communities/ohejp/?page=1&size=20> . On this repository, you can submit your deliverables and other data related to your project. A link will be made between the repository and the OHEJP website. (Cf. Annex: [How to use the OH-EJP community repository on Zenodo?](#))
  - Search for other data repositories at [re3data](https://re3data.org). On top of specific research disciplines, you can filter on access categories, data usage licenses, trustworthy data repositories (with a certificate or explicitly adhering to archival standards) and on the availability of persistent identifier.



**Remember**, you do not need to keep everything! Work with your library to help you determine which data you need to retain for validation and/or reuse.



**When choosing a repository, it is important to consider factors such as whether the repository:**

- Gives your submitted dataset a persistent and unique identifier. This is essential for sustainable citations, both for data and publications, and to make sure that research outputs in disparate repositories can be linked back to particular researchers and grants.
- Provides a landing page for each dataset, with metadata that helps others find it, tell what it is, relate it to publications, and cite it. This makes your research more visible and stimulates reuse of the data.
- Helps you to track how the data has been used by providing access and download statistics.
- Responds to community needs and is “preferably” certified as a “trustworthy data repository”, with an explicit ambition to keep the data available in the long term.
- Matches your particular data needs (e.g. formats accepted, access, backup, recovery and sustainability of the service). Most of this information should be contained within the data repository’s policy pages.
- Offers clear terms and conditions that meet legal requirements (e.g. for data protection) and allow reuse without unnecessary licensing conditions.
- Provides guidance on how to cite the data that has been deposited.
- Charges for its services.

**Your institution may offer Research Data Management support to help you deal with these issues and get the most out of the investment put into your research. This could involve:**

- Registering datasets with the institution’s Data Catalogue to help make the research more visible. National registry services are also being established to harvest institutional data catalogue records to make the data visible at a national level.
- Depositing the dataset with an institutional repository to maintain a long-term record of its safekeeping and, if it is publicly available, the access and download statistics.
- Providing advice on rich metadata and other documentation, to help make the data both discoverable and understandable.
- Selecting an appropriate license for using your data, also when you make them available Open Access.

### 5.5.6 Data protection

If your research involves human subjects, you will need to consider both legal and ethical obligations regarding sharing your data.

Data protection refers to the rights of the individuals whose data are being collected, held, and processed. Individuals have the right to have inaccuracies corrected and to know what data are being held and how they are being used.

Data protection legislation has changed since 25 May 2018 when the General Data Protection Regulation (GDPR) came into force replacing the existing regulations. Information relating to the new legislation is available [EU GDPR](#) and in this [Wikipedia](#) article

#### 5.5.6.1 Data protection, access rights and procedure

It is accepted practice to share the final research data. A well-drafted license agreement will help with this case. If data is not shared, it should be backed up by a strong argument.





### 5.5.6.2 Intellectual Property Rights (IPR)

Intellectual Property Rights (IPR) can be defined as rights acquired over any work created or invented with the intellectual effort of an individual. Common types of IPR include copyrights, patents, trademarks, geographical indications, industrial design rights, integrated circuits, design layouts, and confidential information (trade secrets).

Under intellectual property law, owners are granted certain exclusive rights, such as:

- The ability to:
  - Publish to various markets or assign that right to another
  - License the manufacture and distribution of inventions
  - Sue in case of unlawful or deceptive copying
- Moral rights (which are waivable but not assignable).

As a researcher, you should clarify who has primary ownership of the data, and whose rights should be considered when making decisions about the management and dissemination of the data (such as funders, your institution, research subjects, collaborators, publishers and the public). Ownership and rights will determine how the data can be managed into the future, so these should be documented early in a project through data management planning.

#### 5.5.6.2.1 Who owns the IPR in my work?

Usually, the employer is the first owner of any copyright in the work achieved by an employee in the course of his employment (subject to any agreement to the contrary).” However, some facilities have incentive schemes, which allow staff a share of revenue from intellectual property they have created. You should ask your institution for guidance on the ownership and management of any intellectual property arising from your work, particularly any that has potential for commercialization. In some cases, external academic or commercial collaborators may have intellectual property rights in research outputs. Normally there are consortium agreements or legal contracts associated with such collaborations. While commercially sensitive data are often not shared widely, there are options for doing so, such as after filing for a patent to protect the commercial application of the idea or invention.

#### 5.5.6.2.2 IPR: What is different about data?

It depends. “Research data are collected, observed, or created, for the purposes of analysis to produce and validate original research results.” Different domains have vastly different forms of research data, from straightforward numeric lists to highly annotated audio/video materials.

Some data, such as photographs or video, could be treated as an original work and enjoy normal copyright protections.

Databases may be protected by the “database right” arising from the European Database Directive (1996), in recognition of the work involved in creation and arrangement. The *sui generis* (Latin for “of its own kind”) protection for databases means they do not need to meet the criterion of originality used for copyright protection.

Other data may simply be facts and not subject to any IPR. However, a collection of “just facts” could be protected, either under the database right because of its structure and arrangement, or under copyright by the particular presentation of those facts: such as the layout of a website, or the typographical arrangement of a printed directory.

Different legal jurisdictions treat data differently too. Creativity or originality may be a factor as to whether data attracts copyright, such as in the US for the former, or Australia for the latter. Just as



typography is protected in regular works, the arrangement and structure of a database may (or may not) be protected.

The European Union Database Directive offers protection in some European countries for the contents of a database where a substantial investment was made to obtain, verify or present them, although exemptions may exist for teaching or research purposes.

Biological sequence data could also contain intellectual property, particularly if there is scope for their incorporation into biotech products.

Establishing who has legal rights to your data and whether you have rights to use others' data might not be an easy task.

#### 5.5.6.2.3 Copyright & IPR reminder

- Name(s) of the owner(s) of the data
- License(s) for reuse which will be applied (e.g. one of the licenses available from Creative Commons or Open Data Commons) (Cf. [Open data licensing](#))
- Restriction on third party use
- Any expected delay to data sharing e.g. pending a patent application or embargo related to publication in a journal

#### 5.5.6.3 Freedom of Information

Freedom of Information (Fol) refers to a body of legislation that establishes the right of the public or individual members of the public, to be given access to information from public bodies. Although exemptions exist, including an exemption for data intended for future publication, these may be overruled on grounds of public interest upon appeal.

#### 5.5.6.4 Open Government Data

The open data movement, promoted by [OKF](#) and others, builds on earlier and continuing movements to promote open source software and open access literature, and has succeeded in pushing US and UK government agencies to “open their data” to a great extent. Although in the US, government information has always been free from copyright restrictions, in the UK there has been a more recent move from use of “Crown copyright” status to use of the Open Government License, allowing users to derive new datasets from government sources and publish them freely.

Always check the terms and conditions of any internet source of data before doing so.

### 5.5.7 Ethical review of research projects

A number of ethical requirements apply to the management of research data, particularly where the research involves human subjects. Ethical considerations include the purpose and nature of the research itself, the nature of consent obtained (e.g. opt-in versus opt-out participation) and what data need to be safeguarded during analysis, and destroyed after their use.

#### 5.5.7.1 Which research is subject to ethical review?

Any research involving the use of human participants or live animals will usually be subject to ethical review. Similarly, research which involves referencing individual subjects (people) or storing identifiers for individuals will usually be subject to ethical review, unless the data is obtained through pure observational studies of public behavior.

“Pure” observational studies:



- Are of human action that occurs in a forum open to the general public
- Are non-invasive
- Require no interaction with participants
- Do not identify participants

Professional societies representing particular academic fields tend to publish ethical guidelines. These require regular updating as norms, professional consensus and technologies change over time. In terms of data sharing, there is currently a tension between the risks of personal disclosure and the pressure to make data openly available as part of the record of research.

#### 5.5.7.2 *Privacy and consent*

The “right to privacy” generally refers to the state of being free from intrusion or disturbance in one's private life or affairs. Your research proposal should outline strategies to protect subjects' privacy including how the investigator will access information from or about participants. Privacy and confidentiality are related but are not the same thing. Privacy relates to the individual or subject, whereas confidentiality relates to the actions of the researcher.

There are many different ways of obtaining consent from your research subjects. The form of consent affects not only how you conduct your research but also who can have access to any personal data you hold. In most cases, you must describe to the research subject what you intend to do with their data, who will have access to it and how will it be published before obtaining their consent: this is called *informed consent*.

It is worth considering who needs to access the personal data and what needs to be done to the data in order for it to be shared publicly or with other researchers when deciding on the form of consent you will use. Anonymized data does not require consent for sharing or publication, but it is considered ethical to inform your subjects about what will become of the data.

#### 5.5.7.3 *Confidentiality*

Confidentiality refers to the researcher's agreement with the participant about how the participant's identifiable private information will be handled, managed, and disseminated. The research proposal should outline strategies to maintain confidentiality of identifiable data, including controls on storage, handling, and sharing of personal data.

You can minimize the risk of disclosing confidential information when designing your research by considering the following factors and approaches:

- If possible, collect the necessary data without using personally identifying information.
- If personally identifying information is required, de-identify your data upon collection or as soon as possible thereafter.
- Avoid transmitting unencrypted personal data electronically.

Other considerations include retention of original collection instruments, such as paper questionnaires or interview recordings. Once these are transferred into an analysis package, and transcript and quality assured, or validated, there may no longer be a reason to retain them. Questions of which data to keep and for how long need to be considered in the context of your ability to maintain the confidentiality of your subjects' information, and should be planned.

#### 5.5.7.4 *Reminders for Ethical & Legal compliance*

##### Ethics & Legal Compliance

- Details of consent needed for data preservation and sharing





- Steps to be taken if needed to protect the identity of any participants
- Steps to be taken if needed to ensure sensitive data is stored and transferred securely
- Anonymization procedures
- Expected embargo period
- Data owner

## 5.6 Making data “Interoperable”

In broad terms, interoperability is the ability of different information and communications technology systems and software applications to communicate, to exchange data accurately, effectively, and consistently, and to use the information that has been exchanged. Data interoperability is the ability to interpret correctly data across systems or organizational boundaries.

**Interoperable data** means it can be integrated with other data, applications and workflows. Think about not creating data with proprietary software and making it available in open formats. Remember to use community agreed schemas, controlled vocabularies, keywords, thesauri or ontologies where possible.

- Are the data produced in the project interoperable, that is allowing data exchange and reuse between researchers, institutions, organizations, countries, etc. e.g. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins?
- What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable? (Cf. [Dublin Core Metadata Element Set](#))
- Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
- Will you provide mappings to more commonly used ontologies in case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies?

### 5.6.1 The different levels of interoperability

In general, there are seven basic levels of different levels of interoperability. These levels include:

- **Level Zero or No Interoperability:** This is usually characterized by stand-alone systems, which have no interoperability.
- **Level One or Technical Interoperability:** This level of interoperability involves the use of a communication protocol for the exchange of data between systems. Technical interoperability establishes harmonization at the plug and play, signal and protocol level.
- **Level Two or Syntactic interoperability:** This is the ability of two or more systems to exchange data and services using a common interoperability protocol such as the High Level Architecture (HLA).
- **Level 3 or Semantic Interoperability:** Semantic interoperability refers to the ability of two or more systems to automatically and accurately, interpret the information exchanged, in order to produce useful results as defined by the end users of the systems. Semantic interoperability is also used in a more general sense to refer to the ability of two or more systems to exchange information with an unambiguous and shared meaning. Semantic interoperability implies that the precise meaning of the exchanged information is understood by the communicating systems. Hence, the systems are able to recognize and process semantically equivalent information homogeneously, even if their instances are heterogeneously represented, that is, if they are differently structured, and/or using different terminology or different natural language. Semantic interoperability can thus be said to be distinct from the other levels of interoperability because it ensures that the receiving system understands the meaning of the



exchange information, even when the algorithms used by the receiving system are unknown to the sending system.

- **Pragmatic Interoperability:** This level of interoperability is achieved when the interoperating systems are aware of the methods and procedures that each other are employing. This implies that the use of the data or the context of its application is understood by the participating systems.
- **Dynamic Interoperability:** Two or more systems are said to have attained dynamic interoperability when they are able to comprehend the state changes that occur in the assumptions and constraints that they are making over time, and they are able to take advantage of those changes.
- **Conceptual Interoperability:** Conceptual interoperability is reached if the assumptions and constraints of the meaningful abstraction of reality are aligned.

For more information regarding interoperability in health care, please refer to the publication: [Interoperability in Healthcare: Benefits, Challenges and Resolutions](#)

Make your data interoperable by using:

- Common formats and standards (Cf. [Data formats](#))
- Controlled vocabularies

## 5.7 Increase data “Reuse”, through clarifying licenses

Selection & Preservation

- Details of which data should be retained, shared and/or preserved, with particular reference to contractual, legal or regulatory requirements
- Foreseeable research uses for the data
- Length of time for which data will (or should) be kept beyond the life of the project
- The repository or archive where the data will be held, and any association charges
- Time and effort needed to prepare data for preservation & sharing
- Which data to retain, share & preserve
- Expected future research reuse
- Which repository will be used if any
- Length of time data to be preserved

Make your data reusable by ensuring it:

- Is well-documented
- Has clear license and provenance information

**Create documentation**, e.g. a **README** file to help ensure that your data can be correctly interpreted and reanalyzed by others. A README plain text file should contain the following information:

- For each filename, a short description of what data it includes, optionally describing the relationship to the tables, figures, or sections within the accompanying publication
- For tabular data: definitions of column headings and row labels, data codes (including missing data) & measurement units
- Any data processing steps, especially if not described in the publication, that may affect interpretation of results
- A description of what associated datasets are stored elsewhere, if applicable
- Whom to contact with questions



P.S: If text formatting is important for your README, PDF format could also be used.

Data should have a clear license to govern the terms of its reuse

- How will the data be licensed to permit the widest reuse possible?
- When will the data be made available for reuse? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
- Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the reuse of some data is restricted, explain why.
- How long is it intended that the data remains reusable?
- Are data quality assurance processes described?

Good practice in data documentation, metadata and data citation can all contribute to reaching a better standard of reproducibility of research. Reproducibility is a fundamental principle of the scientific method.

### 5.7.1 Open data licensing

When sharing data, it is important to consider how you want your data to be reused. You can then apply a relevant license that most closely reflects those intended uses. Applying an explicit license removes any ambiguity over what users can and cannot do with your data.

Lawyers can design licenses to meet specific criteria. However, there are number of open licenses developed for widespread use on the internet that anyone can apply.

The *Open Knowledge Foundation's* (OKF) definition of “open knowledge” says that knowledge is open if “one is free to use, reuse, and redistribute it without legal, social or technological restriction.”

Similarly, the [Panton Principles for Open Data in Science](#) state, “for science to effectively function, and for society to reap the full benefits from scientific endeavors, it is crucial that science data be made open.”

Open data therefore means, not only do users have the right to download and use the data, but they have the right to make copies for their own purposes (including data mining and other machine processing). An open license clarifies the IPR in a given work and gives others permission to use it as they wish but with certain conditions attached, including providing a citation to the original work as is normally done within the scholarly publishing system.

It is very important that, where appropriate, you take into consideration both privacy (which concerns the rights of the individual) and confidentiality (which is concerned with the obligations of the researcher). Publication of anonymized data does not require formal consent of the subjects, but it is considered ethical to inform your subjects about what will become of the data.

As a researcher, you should clarify ownership of your research data, e.g. Intellectual Property Rights, and your obligations pertaining to Freedom of Information legislation in the country where you are working, before a project starts. Remember, facts are not copyrightable, whereas the structure of a database could be, under database right. Biological sequence information could be patentable if it is part of an invention.

**Not all data can or should be shared.** Issues such as ownership, commercial value, and confidentiality need to be considered.





Preserving research data for future reuse can be problematic due to the huge amount of data being generated as well as the wide variety and complexity of formats. If you do not archive or manage your data in a systematic way, there is a danger that it will effectively be lost.






It is also important to consider how you want your data to be reused, and to apply a relevant license that most closely reflects intended use. Applying an explicit license removes any ambiguity over what may and may not be done with the data.




- A **CC0 license** makes it clear to users that “no rights are reserved and a work may be released entirely into the public domain”.
- A **CC-BY license** makes it clear to users that they must attribute the work to the creator or author e.g. they must give credit by citing the creator.

Data that has been marked up using the [RDF](#) metadata standard, so that it may be queried using the [SPARQL](#) search language can be described as Linked Open Data.

### 5.7.2 Creative Commons 4.0 licenses explained

There are five main CC licence categories, each of which has its own distinctive logo, reproduced below, although a licence may also be indicated by the common abbreviation of its description.

-  (BY) **Attribution** You must credit the creator of the work.
-  (NC) **Non-commercial** You may not use the work for commercial purposes.
-  (ND) **No derivative works** You may not alter, transform or build upon this work.
-  (SA) **Share alike** If you alter, transform or build upon this work, you may distribute the resulting work only under a licence identical to this one.
-  **CC0** - whereby no rights are reserved and a work may be released entirely into the public domain.

You may also encounter one of six additional basic CC licences, which are a combination of two or more of the above elements, and which will be expressed, for example, thus:    or BY-NC-ND.

[Creative Commons](#) (CC) licenses is a US-based non-profit organization, which has been a leader in developing legal tools for sharing creative works over the internet since the 2000s.

Creative Commons licenses are popular on the internet because they provide robust legal code combined with a human-readable summary that is understandable at a glance, and a machine-to-machine layer of code that will help make information resources interoperable across systems.

In the suite of 4.0 licenses released in 2013, CC has worked with legal experts in a wide range of countries to produce a set of licenses that work well for both research data and expressive works because they:

- Conform to both copyright law and database rights where applicable
- Do not need to be ported for use in particular countries, as they are applicable in all jurisdictions



You can select an appropriate license and generate text for your web page or document by using this online [CC](#) tool.

Several open data advocacy groups, including [DataCite](#), recommend **CCO** (CC Zero) for datasets to allow maximum freedom to end-users who may be combining and “mashing up” several sources. However, for academic data creators, whose careers depend on their publication record, **CC-BY** may be more palatable, in which the user is expected to provide a citation to the original source.

### 5.7.3 Licensing Example

- “The main output from this project is field data. We recognize that these data are the property of X and hence we will be asking their permission to license these data to Y for use in their exploration program.”
- “There is an agreement regarding the right of the original data collector, creator, or PI for first use of the data. The specified embargo period associated with the data being submitting extends from [date] until [date]. The embargo will be lifted by [date].”

### 5.7.4 Policies and provision for reuse & re-distribution examples

- “The data gathered will use a copyrighted instrument for some questions. A reproduction of the instrument will be provided to [repository] as documentation for the data deposited with the intention that the instrument be distributed under “fair use” to permit data sharing, but it may not be re-disseminated by users.”
- “The project team will create a dedicated Web site to manage and distribute the data because the audience for the data is small and has a tradition of interacting as a community. The site will be established using a content management system like Drupal or Joomla so that data users can participate in adding site content over time, making the site self-sustaining. The site will be available at an **.org** location. For preservation, we will supply periodic copies of the data to [repository]. That repository will be the ultimate home for the data”.
- “Users of field data should acknowledge and/or offer co-authorship to the investigators who collected the data.”
- “The data to be produced will be of interest to demographers studying family formation practices in early adulthood across different racial and ethnic groups.”
- “In addition to the research community, we expect these data will be used by practitioners and policymakers.”



## 6 Annexes

### 6.1 How to use the DMPonline.be tool?

DMPonline tool is web-based tool to help researchers to write Data Management Plans and to share them. It includes requirements and guidance from funders (such as EC), universities and other group.

DMPonline tool was developed by the Digital Curation Centre: <https://dmponline.dcc.ac.uk> . Various local instances of DMPonline were now developed in different countries. We propose to use the DMPonline which is available in Belgium; <https://dmponline.be> .

Prior to register to DMPonline.be, you need to perform a few steps. The first step is to ask the DMP team of OHEJP to create a DMP template for your project. You can do this demand by email to [Valerie.DeWaele@sciensano.be](mailto:Valerie.DeWaele@sciensano.be) .

To access the tool, you need to sign in either with your institutional account or with your [ORCID](#) ID.

The screenshot shows the homepage of DMPonline.be. At the top left is the logo 'DMP ONLINE .BE'. On the right, there are navigation links for 'Home', 'About', and 'Help'. The main content is divided into two columns. The left column has a 'Welcome' heading and text explaining the tool's purpose and listing participating institutions: 'Instituut voor Natuur- en Bosonderzoek', 'Université Libre de Bruxelles', 'Universiteit Antwerpen', 'Universiteit Gent', 'Universiteit Hasselt', 'Vrije Universiteit Brussel', and 'Wetenschappelijk Instituut Volksgezondheid – Institut Scientifique de Santé Publique (Sciensano)'. It also lists institutions that joined in 2018: 'Université Catholique de Louvain', 'Université de Liège', 'Université de Mons', 'Université de Namur', and 'Vlaamse Instelling voor Technologisch Onderzoek'. The right column is titled 'Sign in' and offers two main options: 'with your institutional account' and 'or with your ORCID ID'. Under the institutional account option, there is a list of 'Sign in with' links for various universities: Sciensano, UAntwerp, UCLouvain, UGent, UHasselt, ULB, ULiège, UMon, UNamur, VITO, and VUB. At the bottom of the page, there are links for 'Contact us', 'Disclaimer and terms of use', and 'Privacy Statement', along with a note 'Based on work by Digital Curation Centre (DCC)'.

The second step is therefore to create your [ORCID](#) ID. ORCID provides a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized.



Sign into ORCID or [Register now](#)

Personal account	Institutional account
------------------	-----------------------

Sign in with your ORCID account

Email or ORCID iD

ORCID password

When filling your account details, please select that everyone should be able to see information added to your ORCID record by default. You will always be able to modify this visibility setting later in your account settings if needed.

### Visibility settings

Your ORCID iD connects with your ORCID record that can contain links to your research activities, affiliations, awards, other versions of your name, and more. You control this content and who can see it.

**By default, who should be able to see information added to your ORCID Record?**

- Everyone** (87% of users choose this) \* ?
- Trusted parties** (5% of users choose this)
- Only me** (8% of users choose this)

Your ORCID account has been created.

Go to [dmponline.be](https://dmponline.be) and login with you ORCID ID. DMPonline will request access to your ORCID record; you authorize this request. Note, please check your spam filter to obtain the email with access information.

**DMPonline.be**

has asked for the following access to your ORCID Record



Read your information with visibility set to Trusted Parties

This application will not be able to see your ORCID password or any other information in your ORCID record with visibility set to Only me. You can manage permission granted to this and other Trusted Organizations in your [account settings](#).



Then, DMPonline.be will be automatically launched, and you can edit your profile.

Note: If you check your ORCID account, you will see in your account settings that DMPonline.be is listed as a trusted organization.

## Trusted organizations <sup>?</sup>

Trusted organization	Approval date	Access type
DMPonline.be <a href="https://dmponline.be">https://dmponline.be</a>	2019-02-12	Read your information with visibility set to Trusted Parties

After having edit your account, you can view the plan that was created for your project by the DMP team. You can then select the action that you want perform, such as edit and share.

View plans Create plan About Help

### My plans

The table below lists the plans that you have created, and any that have been shared with you by others. These can be edited, shared, exported or deleted at anytime.

Filter plans <input type="text" value="Filter"/>				
Name	Owner	Shared?	Last edited	Select an action
Promoting One Health in Europe through joint actions on foodborne zoonoses, antimicrobial resistance and emerging microbiological hazards	Me	Yes (with 2 people)	18-12-2018	View Edit Share Export Delete

Create plan

The plan created follows the guidance of Horizon 2020.

Plan details Horizon 2020 FAIR DMP

- Version information (4 questions, 3 answered) +
- 1. Data summary (6 questions, 5 answered) +
  - 2.1 FAIR data: Making data findable, including provisions for metadata (6 questions, 5 answered) +
  - 2.2. FAIR data: Making data openly accessible (5 questions, 5 answered) +
  - 2.3. FAIR data: Making data interoperable (2 questions, 2 answered) +
  - 2.4. FAIR data: Increase data re-use (through clarifying licenses) (5 questions, 5 answered) +
- 3. Allocation of resources (3 questions, 3 answered) +
- 4. Data security (1 question, 1 answered) +
- 5. Ethical aspects (1 question, 1 answered) +
- 6. Other issues (1 question, 1 answered) +

Export





Then, you will open and answer each question. To answer the questions, you can use guidance from the present document or from the DMPonline tool directly.

**Version information** (4 questions, 3 answered)

Version number

1.0

**Guidance** Add comment

**Guidance**

Indicate the current version of the DMP. E.g. v1.0, 2.0, ...

**Save**

Answered about 2 hours ago by valerie.dewaele@sciensano.be

**Description**

Initial version of the DMP (submitted in month 6 of the program)

**Guidance** Add comment

**Guidance**

Describe the current version of the DMP. E.g. update of the initial DMP (submitted in month 5) for the first periodic evaluation of the project.

**DCC guidance on Metadata**

**Questions to consider:**

- How will you capture / create the metadata?
- Can any of this information be created automatically?
- What metadata standards will you use and why?

**Guidance:**

Metadata should be created to describe the data and aid discovery. Consider how you will capture this information and where it will be recorded e.g. in a database with links to each item, in a 'readme' text file, in file headers etc.

Researchers are strongly encouraged to use community standards to describe and structure data, where these are in place. The DCC offers a [catalogue of disciplinary metadata standards](#).

You can also decide to share your DMP with other partners, who can also participate at the creation of the DMP. Sections are locked for editing when they are being edited by another colleague.

**Plan details** | **Horizon 2020 FAIR DMP** | **Share** | **Export**

You can give other users access to your plan and clarify each collaborator's role here. There are five roles/permission levels:

- Users with 'read only' access can only read the plan.
- Users with 'edit' access can contribute to the plan.
- 'Data Contacts' are any contact persons for the plan other than the principal investigator. They can contribute to the plan.
- 'Co-owners' can also contribute to the plan, but additionally can edit the plan details and control access to the plan.
- 'Principal Investigators' are the main researchers of the project associated with a plan. They can also contribute to the plan, but additionally can edit the plan details and control access to the plan. By default, the creator ('owner') of a plan is listed as principal investigator, but this can be changed and another principal investigator can be added.

Add each collaborator in turn by entering their email address below, choosing a role/permission level and clicking 'Add collaborator'. Any Principal Investigators and Data Contacts you add will also appear in your plan details.

Those you invite will receive an email notification that they have access to this plan. A notification is also issued when a user's permission level is changed.

**Collaborators**

Email address	Permissions	
Valerie De Waele	Co-owner	<a href="#">Remove user access</a>
valerie.dewaele@sciensano.be	Owner	
valerie.dewaele@sciensano.be	Principal Investigator	<a href="#">Remove user access</a>

When you have finished to answer the questions, you can export the DMP as plain text, PDF, html,...

When your DMP is developed, you can publish it on the OHEJP community of the platform Zenodo (see Annex "[How to use the OHEJP community repository on Zenodo?](#)"). A link between the published DMP and the OHEJP website is then made.



## 6.2 How to use the OHEJP community repository on Zenodo?

You can share your data and software on zenodo: <https://www.openaire.eu/zenodo/>. Zenodo provides a rich interface which enables linking research outputs to datasets and funding information. Zenodo allows you to link uploads to grants from more than 11 funders such as European Commission.

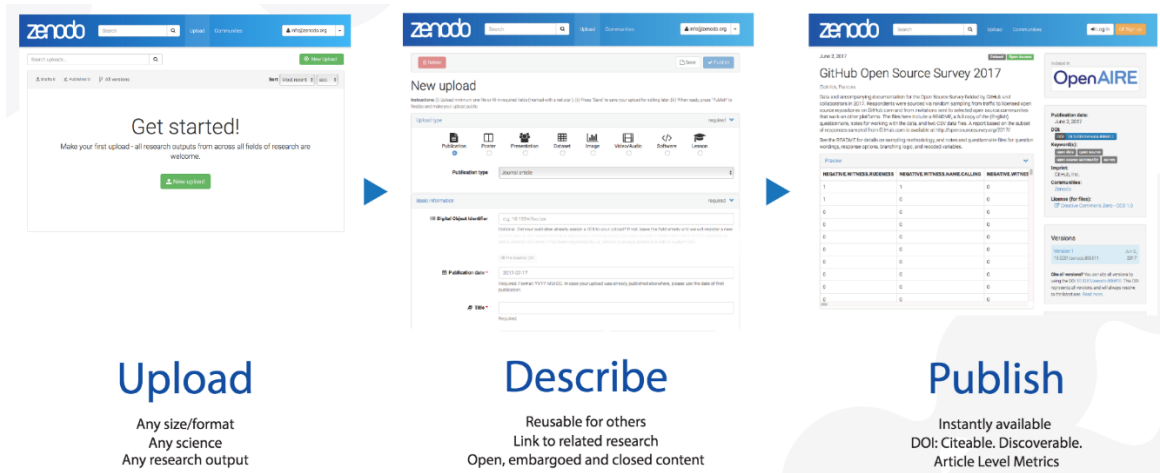
Your data uploaded on Zenodo is stored in the CERN Data Center. The limit of total files size per record is 50GB (max 100 files). One-time 100GB quota can be requested and granted on a case-by-case basis. Zenodo is also integrated with GitHub.

Zenodo makes data **Findable, Accessible, Interoperable and Reusable**. All uploads get a Digital Object Identifier (DOI) to make them easily and uniquely citeable. Zenodo allows for uploading under a variety of different licenses and access levels. Research materials can set to share with reviewers only, and also embargoed. All open content is harvestable via OAI-PMH by third parties.

To use it, a sub-community One Health EJP repository was set up on OpenAIRE platform: <https://www.zenodo.org/communities/ohejp/?page=1&size=20>.



### Share your research output in 3 steps!



Sign-in locally or by using your GitHub or ORCID credentials. Easily upload files of up to 50 GB.

Next, describe your content so others can find it. Add funding and license information and pre-reserve a DOI for your upload. Optionally, you can select a community.

Submit to finalize your upload and publish. After a quick spam-check, your content will be available immediately. Open access uploads will also be visible on Zenodo's front-page.



In detailed:

### Upload your data.

## New upload

One-Health EJP

**Instructions:** (i) Upload minimum one file or fill-in required fields (marked with a red star). (ii) Press "Save" to save your upload for editing later. (iii) When ready, press "Publish" to finalize and make your upload public.

Filename (1 files)	Size	Progress	Delete
D_DMP_WP4_OHEJP_20180629.pdf	774 Kb		

Note: File addition, removal or modification are not allowed after you have published your upload. This is because a Digital Object Identifier (DOI) is registered with [DataCite](#) for each upload.

(minimum 1 file required, max 50 GB per dataset - [contact us](#) for larger datasets)

### Select the community One Health EJP.

Communities recommended

Start typing a community name...

### Select the type of data.

Upload type required

Publication

Poster

Presentation

Dataset

Image

Video/Audio

Software

Lesson

Other

Publication type:

### Reserve the DOI of your data.

Basic information required

**Digital Object Identifier**

Optional. Did your publisher already assign a DOI to your upload? If not, leave the field empty and we will register a new DOI for you. A DOI allows others to easily and unambiguously cite your upload. Please note that it is NOT possible to edit a Zenodo DOI once it has been registered by us, while it is always possible to edit a custom DOI.

Reserve DOI

**Publication date \***

Required. Format: YYYY-MM-DD. In case your upload was already published elsewhere, please use the date of first publication.

**Title \***

Required.

**Authors \***

Optional.



It is recommended to add at least the following keywords:

- Programme acronym: OHEJP
- Project acronym: for example, NOVA
- Data description: for example, data management plan

Select the appropriate access right. It is recommended to select the open access. However, not all data should have open access.

Select the funding. For OHEJP, you can enter the grants 773830. Therefore EC is directly informed of the data publication under OHEJP grants agreement.

In addition, it is also possible to archive and assign a DOI to GitHub repositories using Zenodo. A guide is available on the following link: <https://guides.github.com/activities/citable-code/>.



### 6.3 Dublin Core Metadata Element Set

<b>Contributor</b>	
<b>URI</b>	<a href="http://purl.org/dc/elements/1.1/contributor">http://purl.org/dc/elements/1.1/contributor</a>
<b>Label</b>	Contributor
<b>Definition</b>	An entity responsible for contributing to the resource.
<b>Comment</b>	Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
<b>Coverage</b>	
<b>URI</b>	<a href="http://purl.org/dc/elements/1.1/coverage">http://purl.org/dc/elements/1.1/coverage</a>
<b>Label</b>	Coverage
<b>Definition</b>	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
<b>Comment</b>	Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names [TGN]. Where appropriate, named places or periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges.
<b>References</b>	[TGN] <a href="http://www.getty.edu/research/tools/vocabulary/tgn/index.html">http://www.getty.edu/research/tools/vocabulary/tgn/index.html</a>
<b>Creator</b>	
<b>URI</b>	<a href="http://purl.org/dc/elements/1.1/creator">http://purl.org/dc/elements/1.1/creator</a>
<b>Label</b>	Creator
<b>Definition</b>	An entity primarily responsible for making the resource.
<b>Comment</b>	Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
<b>Date</b>	
<b>URI</b>	<a href="http://purl.org/dc/elements/1.1/date">http://purl.org/dc/elements/1.1/date</a>
<b>Label</b>	Date
<b>Definition</b>	A point or period associated with an event in the lifecycle of the resource.
<b>Comment</b>	Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 [W3CDTF].
<b>References</b>	[W3CDTF] <a href="http://www.w3.org/TR/NOTE-datetime">http://www.w3.org/TR/NOTE-datetime</a>
<b>Description</b>	
<b>URI</b>	<a href="http://purl.org/dc/elements/1.1/description">http://purl.org/dc/elements/1.1/description</a>
<b>Label</b>	Description
<b>Definition</b>	An account of the resource.
<b>Comment</b>	Description may include but is not limited to an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
<b>Format</b>	
<b>URI</b>	<a href="http://purl.org/dc/elements/1.1/format">http://purl.org/dc/elements/1.1/format</a>
<b>Label</b>	Format
<b>Definition</b>	The file format, physical medium, or dimensions of the resource.
<b>Comment</b>	Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME].



**References** [MIME] <http://www.iana.org/assignments/media-types/>

**Identifier**

**URI** <http://purl.org/dc/elements/1.1/identifier>

**Label** Identifier

**Definition** An unambiguous reference to the resource within a given context.

**Comment** Recommended best practice is to identify the resource by means of a string conforming to a formal identification system.

**Language**

**URI** <http://purl.org/dc/elements/1.1/language>

**Label** Language

**Definition** A language of the resource.

**Comment** Recommended best practice is to use a controlled vocabulary such as RFC 4646 [RFC4646].

**References** [RFC4646] <http://www.ietf.org/rfc/rfc4646.txt>

**Publisher**

**URI** <http://purl.org/dc/elements/1.1/publisher>

**Label** Publisher

**Definition** An entity responsible for making the resource available.

**Comment** Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.

**Relation**

**URI** <http://purl.org/dc/elements/1.1/relation>

**Label** Relation

**Definition** A related resource.

**Comment** Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.

**Rights**

**URI** <http://purl.org/dc/elements/1.1/rights>

**Label** Rights

**Definition** Information about rights held in and over the resource.

**Comment** Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights.

**Source**

**URI** <http://purl.org/dc/elements/1.1/source>

**Label** Source

**Definition** A related resource from which the described resource is derived.

**Comment** The described resource may be derived from the related resource (in whole or in part). Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.

**Subject**

**URI** <http://purl.org/dc/elements/1.1/subject>

**Label** Subject

**Definition** The topic of the resource.

**Comment** Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary.

**Title**

**URI** <http://purl.org/dc/elements/1.1/title>

**Label** Title



<b>Definition</b>	A name given to the resource.
<b>Comment</b>	Typically, a Title will be a name by which the resource is formally known.
<b>Type</b>	
<b>URI</b>	<a href="http://purl.org/dc/elements/1.1/type">http://purl.org/dc/elements/1.1/type</a>
<b>Label</b>	Type
<b>Definition</b>	The nature or genre of the resource.
<b>Comment</b>	Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.
<b>References</b>	[DCMITYPE] <a href="http://dublincore.org/documents/dcmi-type-vocabulary/">http://dublincore.org/documents/dcmi-type-vocabulary/</a>



## 6.4 Data Management Plan Checklist Sample

DCC Checklist	DCC Guidance and questions to consider
<b>Administrative Data</b>	
ID	A pertinent ID as determined by the funder and/or institution.
Funder	State research funder if relevant
Grant Reference Number	Enter grant reference number if applicable [POST-AWARD DMPs ONLY]
Project Name	If applying for funding, state the name exactly as in the grant proposal.
Project Description	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- What is the nature of your research project?</li> <li>- What research questions are you addressing?</li> <li>- For what purpose are the data being collected or created?</li> </ul> <p><b>Guidance:</b></p> <p>Briefly summarise the type of study (or studies) to help others understand the purposes for which the data are being collected or created.</p>
PI / Researcher	Name of Principal Investigator(s) or main researcher(s) on the project.
PI / Researcher ID	E.g ORCID <a href="http://orcid.org/">http://orcid.org/</a>
Project Data Contact	Name (if different to above), telephone and email contact details
Date of First Version	Date the first version of the DMP was completed
Date of Last Update	Date the DMP was last changed
Related Policies	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- Are there any existing procedures that you will base your approach on?</li> <li>- Does your department/group have data management guidelines?</li> <li>- Does your institution have a data protection or security policy that you will follow?</li> <li>- Does your institution have a Research Data Management (RDM) policy?</li> <li>- Does your funder have a Research Data Management policy?</li> <li>- Are there any formal standards that you will adopt?</li> </ul> <p><b>Guidance:</b></p> <p>List any other relevant funder, institutional, departmental or group policies on data management, data sharing and data security. Some of the information you give in the remainder of the DMP will be determined by the content of other policies. If so, point/link to them here.</p>
<b>Data Collection</b>	
What data will you collect or create?	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- What type, format and volume of data?</li> <li>- Do your chosen formats and software enable sharing and long-term access to the data?</li> <li>- Are there any existing data that you can reuse?</li> </ul> <p><b>Guidance:</b></p> <p>Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.</p>
How will the data be collected or created?	<p><b>Questions to Consider:</b></p> <ul style="list-style-type: none"> <li>- What standards or methodologies will you use?</li> <li>- How will you structure and name your folders and files?</li> <li>- How will you handle versioning?</li> <li>- What quality assurance processes will you adopt?</li> </ul> <p><b>Guidance:</b></p> <p>Outline how the data will be collected/created and which community data standards (if any) will be used. Consider how the data will be organised during the project, mentioning</p>





	for example naming conventions, version control and folder structures. Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeat samples or measurements, standardised data capture or recording, data entry validation, peer review of data or representation with controlled vocabularies.
<b>Documentation and Metadata</b>	
What documentation and metadata will accompany the data?	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- What information is needed for the data to be to be read and interpreted in the future?</li> <li>- How will you capture / create this documentation and metadata?</li> <li>- What metadata standards will you use and why?</li> </ul> <p><b>Guidance:</b></p> <p>Describe the types of documentation that will accompany the data to help secondary users to understand and reuse it. This should at least include basic details that will help people to find the data, including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed.</p> <p>Documentation may also include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, any assumptions made, and the format and file type of the data. Consider how you will capture this information and where it will be recorded. Wherever possible you should identify and use existing community standards.</p>
<b>Ethics and Legal Compliance</b>	
How will you manage any ethical issues?	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- Have you gained consent for data preservation and sharing?</li> <li>- How will you protect the identity of participants if required? e.g. via anonymisation</li> <li>- How will sensitive data be handled to ensure it is stored and transferred securely?</li> </ul> <p><b>Guidance:</b></p> <p>Ethical issues affect how you store data, who can see/use it and how long it is kept. Managing ethical concerns may include: anonymisation of data; referral to departmental or institutional ethics committees; and formal consent agreements. You should show that you are aware of any issues and have planned accordingly. If you are carrying out research involving human participants, you must also ensure that consent is requested to allow data to be shared and reused.</p>
How will you manage copyright and Intellectual Property Rights (IPR) issues?	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- Who owns the data?</li> <li>- How will the data be licensed for reuse?</li> <li>- Are there any restrictions on the reuse of third-party data?</li> <li>- Will data sharing be postponed / restricted e.g. to publish or seek patents?</li> </ul> <p><b>Guidance:</b></p> <p>State who will own the copyright and IPR of any data that you will collect or create, along with the licence(s) for its use and reuse. For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. Consider any relevant funder, institutional, departmental or group policies on copyright or IPR. Also consider permissions to reuse third-party data and any restrictions needed on data sharing.</p>
<b>Storage and Backup</b>	
How will the data be stored and backed up during the research?	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- Do you have sufficient storage or will you need to include charges for additional services?</li> <li>- How will the data be backed up?</li> <li>- Who will be responsible for backup and recovery?</li> <li>- How will the data be recovered in the event of an incident?</li> </ul> <p><b>Guidance:</b></p> <p>State how often the data will be backed up and to which locations. How many copies are being made? Storing data on laptops, computer hard drives or external storage devices alone is very risky. The use of robust, managed storage provided by university IT teams is preferable. Similarly, it is normally better to use automatic backup services provided by IT Services than rely on manual processes. If you choose to use a third-party service, you</p>



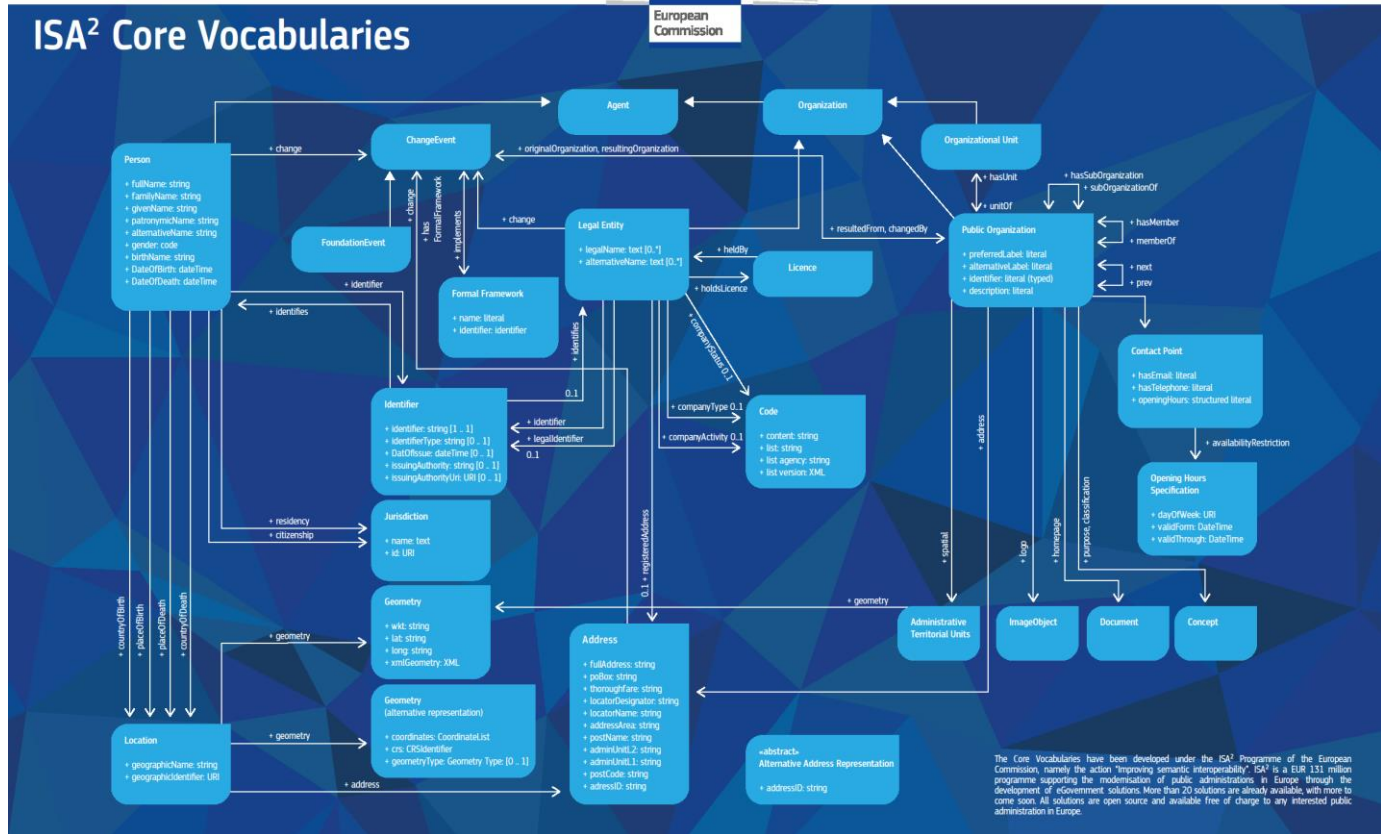
	should ensure that this does not conflict with any funder, institutional, departmental or group policies, for example in terms of the legal jurisdiction in which data are held or the protection of sensitive data.
How will you manage access and security?	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- What are the risks to data security and how will these be managed?</li> <li>- How will you control access to keep the data secure?</li> <li>- How will you ensure that collaborators can access your data securely?</li> <li>- If creating or collecting data in the field how will you ensure its safe transfer into your main secured systems?</li> </ul> <p><b>Guidance:</b></p> <p>If your data is confidential (e.g. personal data not already in the public domain, confidential information or trade secrets), you should outline any appropriate security measures and note any formal standards that you will comply with e.g. ISO 27001.</p>
<b>Selection and Preservation</b>	
Which data should be retained, shared, and/or preserved?	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- What data must be retained/destroyed for contractual, legal, or regulatory purposes?</li> <li>- How will you decide what other data to keep?</li> <li>- What are the foreseeable research uses for the data?</li> <li>- How long will the data be retained and preserved?</li> </ul> <p><b>Guidance:</b></p> <p>Consider how the data may be reused e.g. to validate your research findings, conduct new studies, or for teaching. Decide which data to keep and for how long. This could be based on any obligations to retain certain data, the potential reuse value, what is economically viable to keep, and any additional effort required to prepare the data for data sharing and preservation. Remember to consider any additional effort required to prepare the data for sharing and preservation, such as changing file formats.</p>
What is the long-term preservation plan for the dataset?	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- Where e.g. in which repository or archive will the data be held?</li> <li>- What costs if any will your selected data repository or archive charge?</li> <li>- Have you costed in time and effort to prepare the data for sharing / preservation?</li> </ul> <p><b>Guidance:</b></p> <p>Consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.</p>
<b>Data Sharing</b>	
How will you share the data?	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- How will potential users find out about your data?</li> <li>- With whom will you share the data, and under what conditions?</li> <li>- Will you share data via a repository, handle requests directly or use another mechanism?</li> <li>- When will you make the data available?</li> <li>- Will you pursue getting a persistent identifier for your data?</li> </ul> <p><b>Guidance:</b></p> <p>Consider where, how, and to whom data with acknowledged long-term value should be made available. The methods used to share data will be dependent on a number of factors such as the type, size, complexity and sensitivity of data. If possible, mention earlier examples to show a track record of effective data sharing. Consider how people might acknowledge the reuse of your data.</p>
Are any restrictions on data sharing required?	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- What action will you take to overcome or minimise restrictions?</li> <li>- For how long do you need exclusive use of the data and why?</li> <li>- Will a data sharing agreement (or equivalent) be required?</li> </ul> <p><b>Guidance:</b></p> <p>Outline any expected difficulties in sharing data with acknowledged long-term value,</p>



	<p>along with causes and possible measures to overcome these. Restrictions may be due to confidentiality, lack of consent agreements or IPR, for example. Consider whether a non-disclosure agreement would give sufficient protection for confidential data.</p>
<p><b>Responsibilities and Resources</b></p>	
<p>Who will be responsible for data management?</p>	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?</li> <li>- Who will be responsible for each data management activity?</li> <li>- How will responsibilities be split across partner sites in collaborative research projects?</li> <li>- Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between partners?</li> </ul> <p><b>Guidance:</b></p> <p>Outline the roles and responsibilities for all activities e.g. data capture, metadata production, data quality, storage and backup, data archiving &amp; data sharing. Consider who will be responsible for ensuring relevant policies will be respected. Individuals should be named where possible.</p>
<p>What resources will you require to deliver your plan?</p>	<p><b>Questions to consider:</b></p> <ul style="list-style-type: none"> <li>- Is additional specialist expertise (or training for existing staff) required?</li> <li>- Do you require hardware or software which is additional or exceptional to existing institutional provision?</li> <li>- Will charges be applied by data repositories?</li> </ul> <p><b>Guidance:</b></p> <p>Carefully consider any resources needed to deliver the plan, e.g. software, hardware, technical expertise, etc. Where dedicated resources are needed, these should be outlined and justified.</p>



# 6.5 Core Vocabularies sample





## 6.6 A Data Management Plan Example from DMPTool Public Plan

---

### Root Trait Genetic Characterization

*A Data Management Plan created using DMPTool*

Creator: Alfredo Delgado

Affiliation: Texas A&M University

Template: National Science Foundation (NSF)

Last modified: 11-07-2017

**Copyright information:**

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

---



# Root Trait Genetic Characterization

---

## Data and Materials Produced

Because of the interdisciplinary nature of this proposal, data produced will range from observational and experimental data collected in the lab and field to large bioinformatics datasets and datasets generated through computer-simulated experiments. These data will be mainly in digital format. Several projects will generate 'omic level data; genomics and phenomics. Ground penetrating radar (GPR) and spectral imagery image files.

## Standards, Formats and Metadata

Data gathered will typically be in the following formats: MS Excel (.xlsx), MS Word (.docx), Comma Separated Values (.csv), Portable Document Format (.pdf), Joint Photographic Experts Group (.jpg), Tagged Image File Format (.tiff), and GPR files requiring proprietary software K2FastWave (.dt).

## Roles and Responsibilities

This large interdisciplinary project will employ a standardized data management program. Each data set will be linked to a project description that describes the purpose of the research, the methods used to generate the data and the experimental design, the period of time data were collected and if the data has been updated. The Fellow will be responsible for ensuring the implementation of the data management plan with a specific check at each quarterly review. Maintained and updated laboratory notebook, either digital or hard copy will be required. Here we will implement best practices followed by industry to assure documentation of the generation of intellectual property. Each quarter, Fellow will have a data and materials review with the current hosting sponsor to provide accountability.

## Dissemination Methods

To facilitate file access and sharing, we will develop a detailed plan for sharing data between collaborators on each project, including the use of secure cloud-based access such as Dropbox, and Google Drive. Generally, participants will be expected to archive and make final datasets publicly available within two years of collection, or as soon as they are published, whichever comes first. Work will be undertaken to begin an implementation process on data to CassavaBase to consolidate all data in a specific database.

## Policies for Data Sharing and Public Access

As part of facilitating increasingly complex webs of collaboration, as well as holding members of a collaboration responsible for the data they produce, expectations for project deliverables and plans for disseminating deliverables, when applicable, will be developed at the start of a project and revised as required during the collaboration. Examples of steps collaborators will take to facilitate productive policies for data re-use and re-sharing include:

- Creating a list of participants, by section of a project, for all projects being proposed so that credit can be correctly attributed,
- Including each contributor's expectations for acknowledgement,
- Specifying if data are under license such as common data licenses from Creative Commons or Open Data Commons.

## Archiving, Storage and Preservation

Until the full integration of data to the CassavaBase system, cloud archiving via Dropbox and a redundant Google Drive, as well as hard copy, will be the form of storage.



## 7 Further support in developing your DMP

- The Research Data Alliance provides a [Metadata Standards Directory](#), which can be used to search for discipline-specific standards and associated tools.
- The [EUDAT B2SHARE](#) tool includes a built-in license wizard that facilitates the selection of an adequate license for research data.
- Useful listings of repositories include: [Registry of Research Data Repositories](#)
- Some repositories like [Zenodo](#) allow researchers to deposit both publications and data, while providing tools to link them.
- Other useful tools include [DMP online](#) and platforms for making individual scientific observations available such as [ScienceMatters](#).

<i>Managing and sharing data: a best practice guide</i>	<ul style="list-style-type: none"> <li>• 40 Pages describing the best practice for managing and sharing data for researchers <a href="http://data-archive.ac.uk/media/2894/managingsharing.pdf">http://data-archive.ac.uk/media/2894/managingsharing.pdf</a></li> </ul>
<i>How to select a data repository?</i>	<ul style="list-style-type: none"> <li>• <a href="https://www.openaire.eu/opendatapilot-repository-guide">https://www.openaire.eu/opendatapilot-repository-guide</a></li> </ul>
<i>ERC Template for the Data Management Plan</i>	<ul style="list-style-type: none"> <li>• <a href="https://www.google.com/url?sa=t&amp;rct=j&amp;q=&amp;esrc=s&amp;source=web&amp;cd=5&amp;cad=rja&amp;uact=8&amp;ved=2ahUKewi89JCTIKLfAhVFJFAKHeKzAvEQFjAEegQICBAC&amp;url=http%3A%2F%2Fec.europa.eu%2Fresearch%2Fparticipants%2Fdata%2Fref%2Fh2020%2Fgm%2Freporting%2Fh2020-erc-tpl-oa-data-mgt-plan_en.docx&amp;usg=AOvVaw2RjDvssk_dARf6qDnTX3ij">https://www.google.com/url?sa=t&amp;rct=j&amp;q=&amp;esrc=s&amp;source=web&amp;cd=5&amp;cad=rja&amp;uact=8&amp;ved=2ahUKewi89JCTIKLfAhVFJFAKHeKzAvEQFjAEegQICBAC&amp;url=http%3A%2F%2Fec.europa.eu%2Fresearch%2Fparticipants%2Fdata%2Fref%2Fh2020%2Fgm%2Freporting%2Fh2020-erc-tpl-oa-data-mgt-plan_en.docx&amp;usg=AOvVaw2RjDvssk_dARf6qDnTX3ij</a></li> </ul>
<i>Template Horizon 2020 Data Management Plan (DMP)</i>	<ul style="list-style-type: none"> <li>• <a href="https://www.google.com/url?sa=t&amp;rct=j&amp;q=&amp;esrc=s&amp;source=web&amp;cd=6&amp;cad=rja&amp;uact=8&amp;ved=2ahUKewi89JCTIKLfAhVFJFAKHeKzAvEQFjAFegQIBhAC&amp;url=http%3A%2F%2Fec.europa.eu%2Fresearch%2Fparticipants%2Fdata%2Fref%2Fh2020%2Fgm%2Freporting%2Fh2020-tpl-oa-data-mgt-plan_en.docx&amp;usg=AOvVaw2l1Qci8gIHw-qx6BRTEqCH">https://www.google.com/url?sa=t&amp;rct=j&amp;q=&amp;esrc=s&amp;source=web&amp;cd=6&amp;cad=rja&amp;uact=8&amp;ved=2ahUKewi89JCTIKLfAhVFJFAKHeKzAvEQFjAFegQIBhAC&amp;url=http%3A%2F%2Fec.europa.eu%2Fresearch%2Fparticipants%2Fdata%2Fref%2Fh2020%2Fgm%2Freporting%2Fh2020-tpl-oa-data-mgt-plan_en.docx&amp;usg=AOvVaw2l1Qci8gIHw-qx6BRTEqCH</a></li> </ul>
<i>Data security further reading links</i>	<ul style="list-style-type: none"> <li>• A short video describing data security: <a href="https://www.youtube.com/watch?time_continue=151&amp;v=Ylkg7-JOYX8">https://www.youtube.com/watch?time_continue=151&amp;v=Ylkg7-JOYX8</a></li> <li>• Tips and techniques for managing your passwords: <a href="https://www.ed.ac.uk/infosec/how-to-protect/lock-your-devices/passwords">https://www.ed.ac.uk/infosec/how-to-protect/lock-your-devices/passwords</a></li> </ul>
<i>FAIR Data</i>	<ul style="list-style-type: none"> <li>• <a href="https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/1.-Plan/FAIR-data">https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/1.-Plan/FAIR-data</a></li> </ul>
<i>FAIR data principles</i>	<ul style="list-style-type: none"> <li>• <a href="http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf">http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf</a></li> </ul>
<i>Research data management</i>	<ul style="list-style-type: none"> <li>• An overview of recent development in the Netherland <a href="https://dans.knaw.nl/en/about/organisation-and-policy/information-">https://dans.knaw.nl/en/about/organisation-and-policy/information-</a></li> </ul>



	<a href="#">material/Whitepaper_ResearchdatamanagementAnoverview_DEF.pdf</a>
<i>Data Cleansing Cost</i>	<ul style="list-style-type: none"><li>• <a href="http://datascope.net/cost-of-data-cleansing/">http://datascope.net/cost-of-data-cleansing/</a></li></ul>
<i>Amnesia, Data anonymization tool</i>	<ul style="list-style-type: none"><li>• <a href="https://amnesia.openaire.eu/index.html">https://amnesia.openaire.eu/index.html</a></li></ul>
<i>Data Archives, description and Metadata</i>	<ul style="list-style-type: none"><li>• <a href="https://data-archive.ac.uk/help/user-faq">https://data-archive.ac.uk/help/user-faq</a></li></ul>
<i>DMPOnline Belgium</i>	<ul style="list-style-type: none"><li>• Tool for creating Data Management Plans <a href="https://dmponline.be/">https://dmponline.be/</a></li></ul>
<i>Metadata Generator</i>	<ul style="list-style-type: none"><li>• <a href="http://nsteffel.github.io/dublin_core_generator/">http://nsteffel.github.io/dublin_core_generator/</a></li></ul>
<i>Metadata Extractor tool</i>	<ul style="list-style-type: none"><li>• <a href="https://www.extractmetadata.com/">https://www.extractmetadata.com/</a></li></ul>
<i>Create a codebook</i>	<ul style="list-style-type: none"><li>• DDI Codebook: <a href="http://www.ddialliance.org/training/getting-started-new-content/create-a-codebook">http://www.ddialliance.org/training/getting-started-new-content/create-a-codebook</a></li></ul>
<i>Persistent identifiers</i>	<ul style="list-style-type: none"><li>• <a href="https://www.openaire.eu/what-is-a-persistent-identifier">https://www.openaire.eu/what-is-a-persistent-identifier</a></li></ul>
<i>Zenodo</i>	<ul style="list-style-type: none"><li>• Your research output is stored safely for the future in the same cloud infrastructure as CERN's <a href="https://zenodo.org/">https://zenodo.org/</a></li></ul>
<i>Adding publication to Zenodo</i>	<ul style="list-style-type: none"><li>• A step by step guide to adding publication to Zenodo <a href="https://www.structuralbiology.eu/help/other/zenodo-upload-guidelines">https://www.structuralbiology.eu/help/other/zenodo-upload-guidelines</a></li></ul>
<i>What is Metadata</i>	<ul style="list-style-type: none"><li>• <a href="https://www.openaire.eu/what-is-metadata">https://www.openaire.eu/what-is-metadata</a></li></ul>
<i>Dublin Core Metadata Initiative</i>	<ul style="list-style-type: none"><li>• <a href="http://dublincore.org/documents/dces/">http://dublincore.org/documents/dces/</a></li></ul>
<i>DCC metadata directory</i>	<ul style="list-style-type: none"><li>• <a href="http://www.dcc.ac.uk/resources/metadata-standards">http://www.dcc.ac.uk/resources/metadata-standards</a></li></ul>
<i>FAIR sharing</i>	<ul style="list-style-type: none"><li>• A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies. <a href="https://fairsharing.org/">https://fairsharing.org/</a></li></ul>
<i>Registry of Research Data Repositories</i>	<ul style="list-style-type: none"><li>• <a href="http://www.re3data.org">www.re3data.org</a></li></ul>





<i>H2020 AGA</i>	<ul style="list-style-type: none"> <li>• H2020 Program Annotated Model Grant Agreement <a href="http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf">http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf</a></li> </ul>
<i>H2020: Ethics</i>	<ul style="list-style-type: none"> <li>• <a href="http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm">http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm</a></li> </ul>
<i>H2020 Open Access &amp; Data Management</i>	<ul style="list-style-type: none"> <li>• <a href="http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm">http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm</a></li> </ul>
<i>H2020: Funding</i>	<ul style="list-style-type: none"> <li>• H2020 Participant Portal <a href="http://ec.europa.eu/research/participants/docs/h2020-funding-guide/index_en.htm">http://ec.europa.eu/research/participants/docs/h2020-funding-guide/index_en.htm</a></li> </ul>
<i>Metadata Standard Directory Working Group</i>	<ul style="list-style-type: none"> <li>• <a href="http://rd-alliance.github.io/metadata-directory/">http://rd-alliance.github.io/metadata-directory/</a></li> </ul>
<i>H2020 FAIR Guideline &amp; DMP Template</i>	<ul style="list-style-type: none"> <li>• Guidelines on FAIR Data Management in Horizon 2020 <a href="http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf">http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf</a></li> </ul>
<i>H2020 DMP Template</i>	<ul style="list-style-type: none"> <li>• <a href="http://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf">http://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf</a></li> </ul>
<i>H2020 FAIR Guideline</i>	<ul style="list-style-type: none"> <li>• Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 <a href="http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf">http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf</a></li> </ul>
<i>DMP &amp; DMP Checklist</i>	<ul style="list-style-type: none"> <li>• <a href="http://www.dcc.ac.uk/resources/data-management-plans">www.dcc.ac.uk/resources/data-management-plans</a></li> <li>• <a href="http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf">http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf</a></li> </ul>
<i>A Data Management Plan Checklist</i>	<ul style="list-style-type: none"> <li>• <a href="https://www.cessda.eu/content/download/3844/35033/file/20171117DMPQuestionsCESSDAExpertTourGuide.pdf">https://www.cessda.eu/content/download/3844/35033/file/20171117DMPQuestionsCESSDAExpertTourGuide.pdf</a></li> </ul>
<i>Data Management plans examples</i>	<ul style="list-style-type: none"> <li>• <a href="https://www.lib.ncsu.edu/data-management/dmp_examples">https://www.lib.ncsu.edu/data-management/dmp_examples</a></li> <li>• <a href="http://libguides.gwumc.edu/c.php?g=27812&amp;p=170533">http://libguides.gwumc.edu/c.php?g=27812&amp;p=170533</a></li> <li>• <a href="https://www.lib.umn.edu/datamanagement/DMP/example">https://www.lib.umn.edu/datamanagement/DMP/example</a></li> </ul>