Hands-On DMP Creation Exercise

In this exercise, we will guide you through the process of creating a Data Management Plan (DMP) based on the European Commission Horizon 2020 DMP template, using DMPonline—an online tool that is jointly developed by the Digital Curation Centre (DCC) and the University of California Curation Center (UC3).

Participants are encouraged to use their own research projects or project plans. Alternatively an example project will be provided.

DMPonline Set-Up

Please **access the DMPonline home page**, by following this link: **https://dmponline.dcc.ac.uk/**

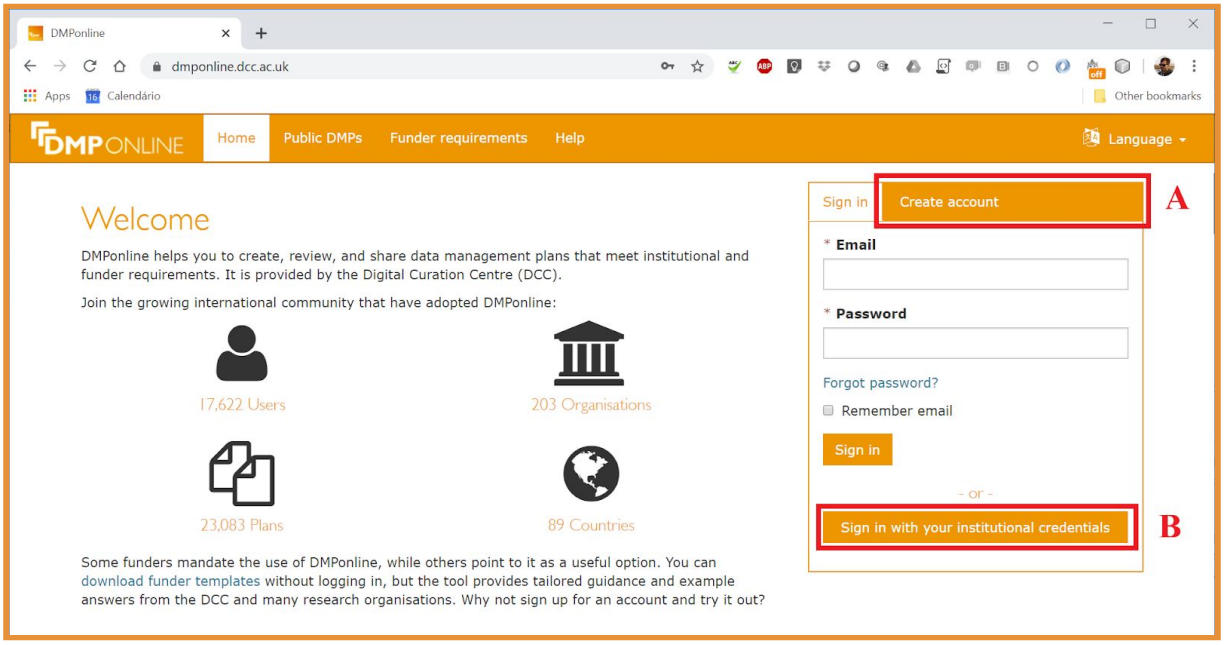

Figure 1. DMPonline home page

As can be seen in figure 1, you are presented with a form that allows you to either sign in, or create an account.

DMPonline allows users to sign in in two distinct ways: (1) through a standard user account and (2) using institutional credentials.

If you already have a DMPonline account, or valid institutional credentials, jump ahead to "Sign In", **otherwise, click on button A,** as seen in figure 1.

Register

You will be shown a "create account" form, as can be seen in figure 2. Please **fill in all the required fields** and **submit** your registration request, **by clicking on the "Create account" button**.

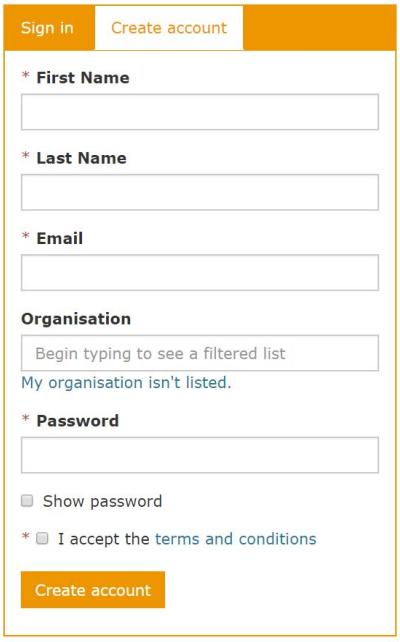

Figure 2. "Create account" form

If your organisation is not listed in the create account form, **click on the link "My organisation isn't listed"**, and you will be able to provide the name of your organisation.

Sign In

By clicking on button B (as seen in figure 1), you'll be redirected to the page displayed in figure 3. In this page you'll be asked to type in or select the name of your institution.

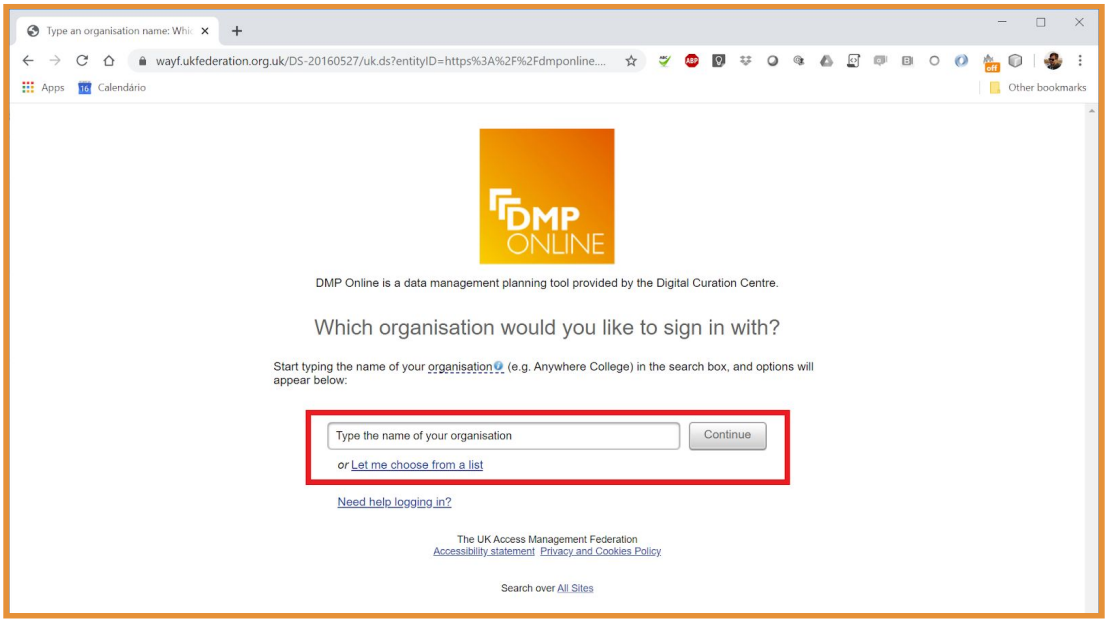

Figure 3. Institutional Credentials Sign In

If your institution is listed, then you will be redirected to your institutional sign in form. You will then be able to use DMPonline.

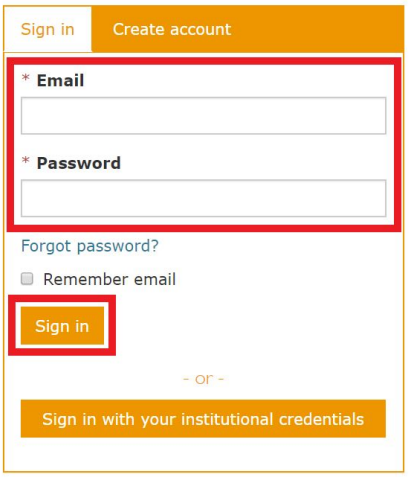

Figure 4. Sign in form.

If you have otherwise registered (as instructed in section 2.1), you can proceed to sign in. You can do this by **filling in both your email address and password**, and then **clicking on the sign in button, as seen in figure 4.**

Once you've successfully signed in to DMPonline, you'll be directed to the DMPonline dashboard (see figure 5), where you'll be able to create DMP documents.

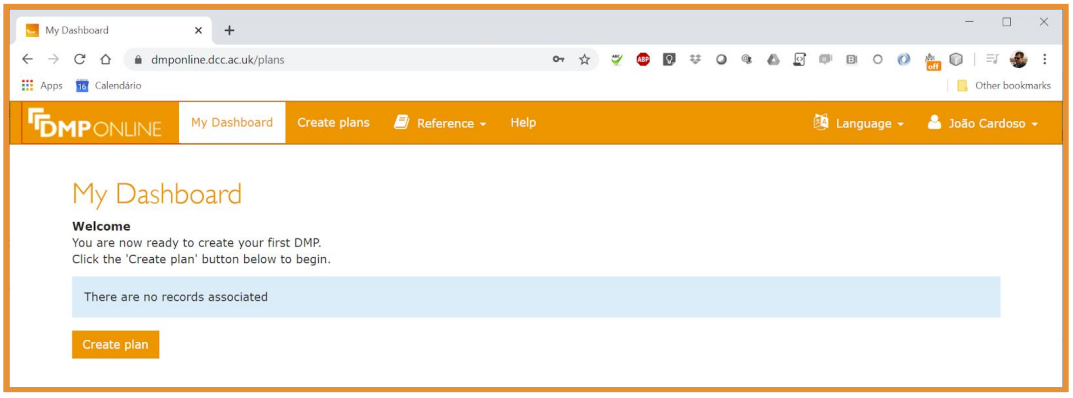

Figure 5. DMPonline dashboard.

DMP Set-Up

In DMPonline, the DMP creation process begins in the dashboard. Once you **click on the "Create plan" button** (seen in figure 5), you will be shown a "create a new plan" form, as is seen in figure 6.

You will be asked to answer three questions:

A. What research project are you planning?

In this question you should state the project title, as is in the proposal. Alternatively, if you're working with fictional data, tick the box stating that this is a mock project.

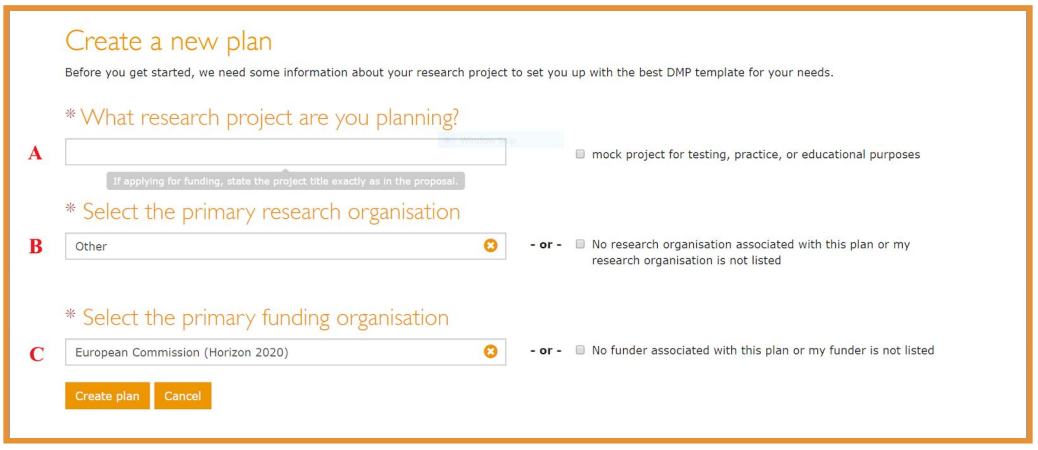

Figure 6. "Create a new plan" form.

B. Select the primary research organisation

In this question you should select your organisation if it is listed. If it is not listed, tick the appropriate box.

C. Select the primary funding organisation

In this question you should **select "European Commission (Horizon 2020)"**, as we'll be using the H2020 template for DMP creation.

Project Details

Once you've filled in the three questions, proceed to **click on the "Create plan" button**. You will be directed to the **Project Details** section of the DMP, as can be seen in figure 7.

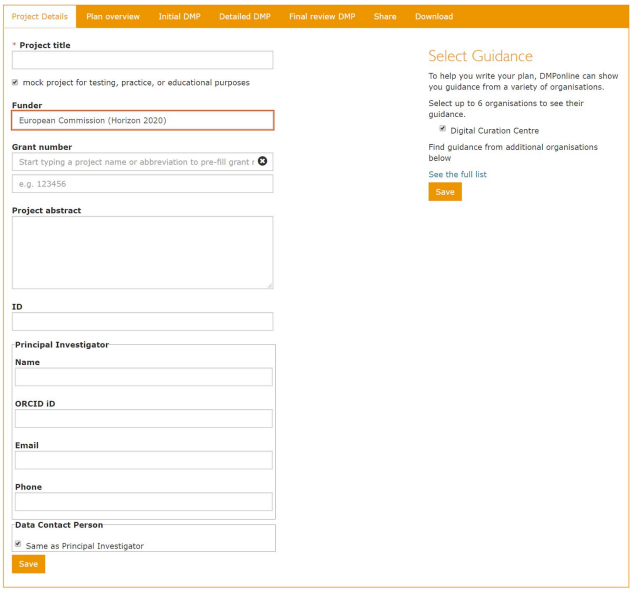

Figure 7. Project Details

In this section you'll be asked to provide information on the following topics:

- Project title

- Funder
- Grant
- Project abstract
- DMP identifier
- Principal investigator
- Data contact person

You may notice that some of the information may already have been automatically filled in. In particular the **funder**, should already list **"European Commission (Horizon 2020)"**

Fill in the necessary information and **click on the "Save" button**. You will then be directed to the **"Plan overview"** section of the DMP, as seen in figure 8.

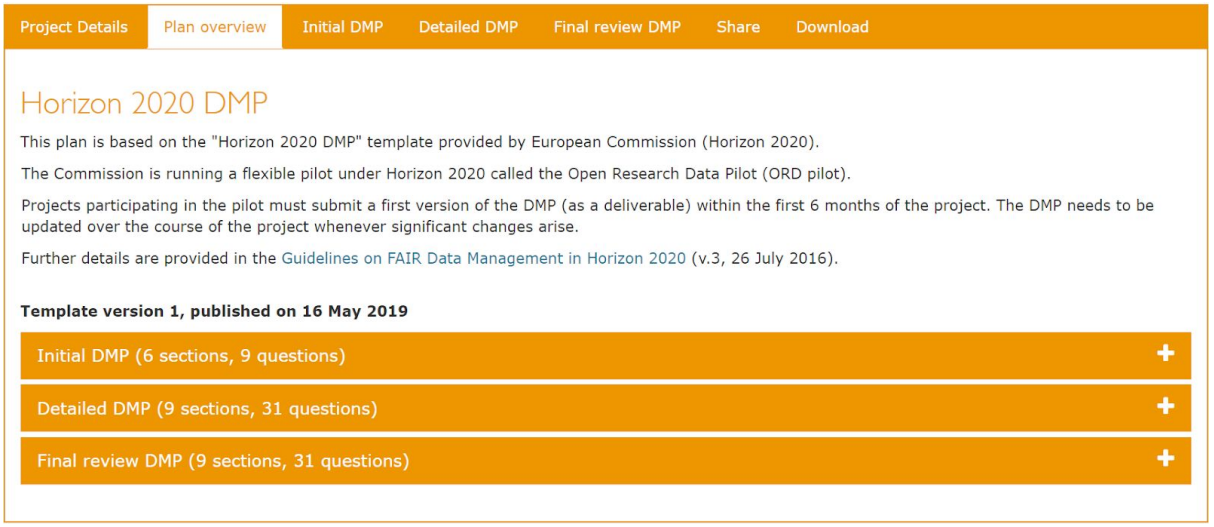

Figure 8. Plan overview.

As the name implies, the "Plan overview" section gives you an overall overview on the three types of DMP that can be created based on the Horizon 2020 DMP template. They are:

- **Initial DMP**, an initial version of the DMP that needs to be submitted by month 6 of the project. It does not require detailed answers for all questions.

- **Detailed DMP**, a standard detailed version of the DMP. These versions should be submitted periodically, as the DMP is intended to be a living document. Over the progress of the project the DMP versions should progress to a finer level of granularity.

- **Final review DMP,** if there are no other periodic reviews foreseen within the grant agreement, then a final version of the DMP needs to be made in time for the final review at the latest.

For the purpose of this hands-on exercise **we'll assume that no other previous DMP versions were created** and will be creating a **Final review DMP**.

H2020 DMP Creation

1. Data Summary

In this section of the H2020 DMP Template you're asked to write free-text answers to the following seven questions on the data that is to be generated throughout the course of the project. Although the template gives you leeway with respect to the granularity of your description (i.e. you are allowed to discuss all datasets collectively or each dataset individually), we recommend that you be as granular as required to capture the singularities of your datasets. That is to say, you should only group datasets for which the answer to the question is the same.

1.1. State the purpose of the data collection/generation

Identify the broad purpose for which data is to be collected/created in this project. If different datasets are to be collected/generated for different purposes, you should list them separately. However, only a connection to the general purpose(s) of the project is expected at this stage.

1.2. Explain the relation to the objectives of the project

Detail further the overview provided in the previous question, by clarifying the connections between datasets (individually or grouped) and the objectives of the project to which they contribute. We recommend that your answer be organized by project objective.

1.3. Specify the types and formats of data generated/collected

List the types of data and corresponding file formats of all datasets to be generated/collected, including an explanation of why the file formats were selected (e.g. open file format, the most popular file format, staff expertise). Open file formats should be preferred over proprietary ones, in order to ensure long-term usability of data.

1.4. Specify if existing data is being re-used (if any)

If the project will rely in part or whole on existing datasets (e.g. from a previous project, from a public domain database), these datasets should be listed here.

1.5. State the origin of the data

Detail the provenance of all data used in the project, both pre-existing and newly-created. For pre-existing data, you should list the current location of the data and, if applicable, the project under which it was generated. For newly-created data, you should describe who will generate the data and how it will be generated.

1.6. State the expected size of the data (if known)

Estimate the expected size or quantify the exact size (for existing data) of the data generated/collected in the course of the project. We recommend that you provide both the total size and the size per dataset.

1.7. Outline the data utility: to whom will it be useful

List the third parties that might have interest in the data generated/collected in the project as well as the potential uses of the data.

2. FAIR Data

The FAIR data principles (i.e. findable, accessible, interoperable and reusable) are being strongly advocated by the European Commission, thanks in part to the efforts of ELIXIR. Thus, the next four sections of the H2020 DMP Template are devoted exclusively to describing the 'FAIRness' of your data.

2.1. Making data findable, including provisions for metadata

FAIR Data, part I: Findability
Data is findable if it is richly and unambiguously described with metadata and keywords that enable its adequate indexation by search engines.

2.1.1. Outline the discoverability of data (metadata provision)

Describe the practices you will adopt to increase data discoverability, such as using persistent identifiers, providing rich metadata, using controlled vocabularies, and/or providing keywords. Only an overview is expected in this question, as you'll be able to detail each aspect further below.

2.1.2. Outline the identifiability of data and refer to standard identification mechanism.

Specify the methods that are used to attribute data identifiers, be they persistent identifiers such as Digital Object Identifiers (DOI) or Persistent Uniform Resource Locators (PURL) or local identifiers such as database accession numbers. You should describe how the identifiers are attributed.

2.1.3. Outline naming conventions used.

Detail any naming conventions, such as controlled vocabularies and ontologies that are being used to in the (meta)data in this project. We recommend that you list naming conventions by type of data and/or domain they cover.
If you are unaware of naming conventions in the research domain of your project, you can browse ontology lookup services such as BioPortal or repositories of standards such as FAIRsharing.org to identify applicable controlled vocabularies and ontologies.

2.1.4. Outline the approach towards search keyword.

If keywords are being added to increase data discoverability, describe what are the criteria for selection. We recommend that you adopt keywords from an ontology detailing scientific domains, such as EDAM.

2.1.5. Outline the approach for clear versioning.

Detail how data versioning is handled in the context of the project. We recommend either the software development practice of attributing a 3-digit version identifier following the logic major.minor[.revision] (e.g. 0.9, 1.1, 2.0.1) or the simpler approach of using the timestamp of creation/edition as a version identifier.

2.1.6. Specify standards for metadata creation (if any).

List any metadata standards that are being used in the context of this project. If none are available in your community, detail how metadata is handled in your project.
If you are unaware of whether there are metadata standards in the domain of your project, you can check FAIRsharing.org.

2.2. Making data openly accessible [FAIR data]

FAIR Data, part II: Accessibility
Data is accessible if it is made available through an established and well-documented protocol, if necessary, with authorization and authentication functionalities. If particular software is needed to access or interpret the data, that software should itself be accessible.

2.2.1. Specify which data will be made openly available?

List what data is going to be made available, and if some data is to remain under closed or controlled access, provide a justification as to why. We recommend that you organize this information by dataset.

2.2.2. Specify how the data will be made available.

Describe the means by which the data will be made available, such as deposition in public repositories or hosting in a database developed for the purpose.
If you are unaware of public repositories that are suitable for the type of data in your research project, you can consult the Registry of Research Data Repositories.

2.2.3. Specify 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)?

Detail the methods or protocols by which the day can be accessed, and if applicable, the software that is necessary to access it and whether the documentation and source code for that software are accessible.
In the straightforward case of textual or tabular data published in a public repository, you need only state that the data will be accessible through HTTP.

2.2.4. Specify where the data and associated metadata, documentation and code are deposited.

Detail the specific repositories, databases or other hosting services where the data, metadata, documentation, code, and other artifacts associated with the project will be deposited.
Again, if you are unaware of public repositories that are suitable for the type of data in your research project, you can consult the Registry of Research Data Repositories.

2.2.5. Specify how access will be provided in case there are any restrictions.

If your data will be under controlled access, describe the method through which this control will be enacted (e.g. a data access committee) and clarify the criteria for data access.

2.3. Making data interoperable [FAIR data]

FAIR Data, part III: Interoperability
Data is interoperable if it can be easily combined with other related data and/or if they can be interpreted by available (open) software applications.

2.3.1. Assess the interoperability of your data.

Describe what interoperability concerns pertain to your data and what you will do to address them, such as following metadata standards, using controlled vocabularies, using established data formats, etc. Note that some data is fairly interoperable *a priori* (e.g. sequencing data in FASTQ format) whereas other data requires more care to become interoperable.

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

For each of the distinct data types in your project, provide information regarding any controlled vocabularies or ontologies that are being used with the intent of promoting interoperability.
Again, you can browse ontology lookup services such as BioPortal or repositories of standards such as FAIRsharing.org to identify applicable controlled vocabularies and ontologies.

2.4 Increase data reuse (through clarifying licences) [FAIR data]

FAIR Data, part IV: Reusability
Data is reusable if it is Findable, Accessible and Interoperable, and if it is released under a clear usage license, that should be as broad as possible.

2.4.1. Specify how the data will be licenced to permit the widest reuse possible.

Indicate under which licence each of your datasets will be published, and why that license is the broadest it can be.

2.4.2. Specify when the data will be made available for reuse.

Provide a time-frame for when the data associated with the project will be made available for reuse, regardless of controlled access. If embargo periods are applicable, specify to which data do they apply and explain why they are required.

2.4.3. Specify whether the data produced and/or used in the project is usable by third parties.

Indicate whether your datasets will be reusable (i.e. made available, in a form amenable to reuse, under a license that permits it) by third parties, and at what stage of the project (e.g. during the project, after the project, after an embargo period). If applicable, provide a justification for any data that will not be made reusable.

2.4.4. Describe data quality assurance processes.

Detail the data quality assurance processes used to ensure the quality of the data. These can include processes such as data profiling to discover inconsistencies and data anomalies, or data cleaning to remove outliers or handle missing data, but also simple practices such as using checksums.

2.4.5. Specify the length of time for which the data will remain re-usable.

Describe for how long the data that has been marked as reusable will remain available for reuse by third parties. This time period may be impacted by factors such as funding, allocated resources, software maintenance, etc. If data is to be deposited in a public repository with stable funding (e.g. GenBank) you can state that it will remain re-usable for the foreseeable future.

3. Allocation of resources

This section addresses the distribution of resources in your project. You will be asked to consider the costs associated with making your data FAIR and to establish associations between data management tasks and project staff members.

3.1. Estimate the costs for making your data FAIR.

Provide an appropriate estimate for the costs of making your data FAIR, detailed by category (e.g. data cleaning, data annotation) and specify your provisions for covering these costs.

3.2. Clearly identify responsibilities for data management in your project.

Associate all data management tasks with the various participants in the project.

3.3. Describe costs and potential value of long term preservation.

Discuss the costs associated with long term preservation of your data (which can be null if you publish data in a stable public repository, or substantial if you store it in local hardware) and the value of doing so. Some datasets retain their value to science for a long time, whereas others rapidly lose value as time progresses.

4. Data security

This section focuses on the secure storage of the data that is generated/collected in the context of the project.

4.1. Address data recovery as well as secure storage and transfer of sensitive data.

Describe what policies are in place for the secure storage, recovery and transfer of data (e.g. backup types, frequency, repositories, etc.). Additionally list the staff members responsible for data storage tasks.

5. Ethical aspects

This section addresses any legal or ethical issues that might impact data sharing. Typically these issues are analysed by an ethics committee, resulting in an ethics report that can be submitted along with the project proposal.

5.1. To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former.

Describe any potential legal or ethical issues that require approval by an ethics committee (e.g. personal data, sensitive data, conflicts of interest, etc.)

6. Other

This section addresses other data management practices and policies that were not directly covered in the previous sections.

6.1. Refer to other national/funder/sectional/departmental procedures for data management that you are using (if any).

In this final question you should address any other data management policies or practices that are being followed in this project, and are specific to your context.

Sharing and Exporting the DMP

Once **all of the questions have been addressed**, you can **proceed to export your DMP**. However, **before you do this**, you need to **address the sharing definitions**. You can do this, by opening the DMP share section, as seen in figure 9.

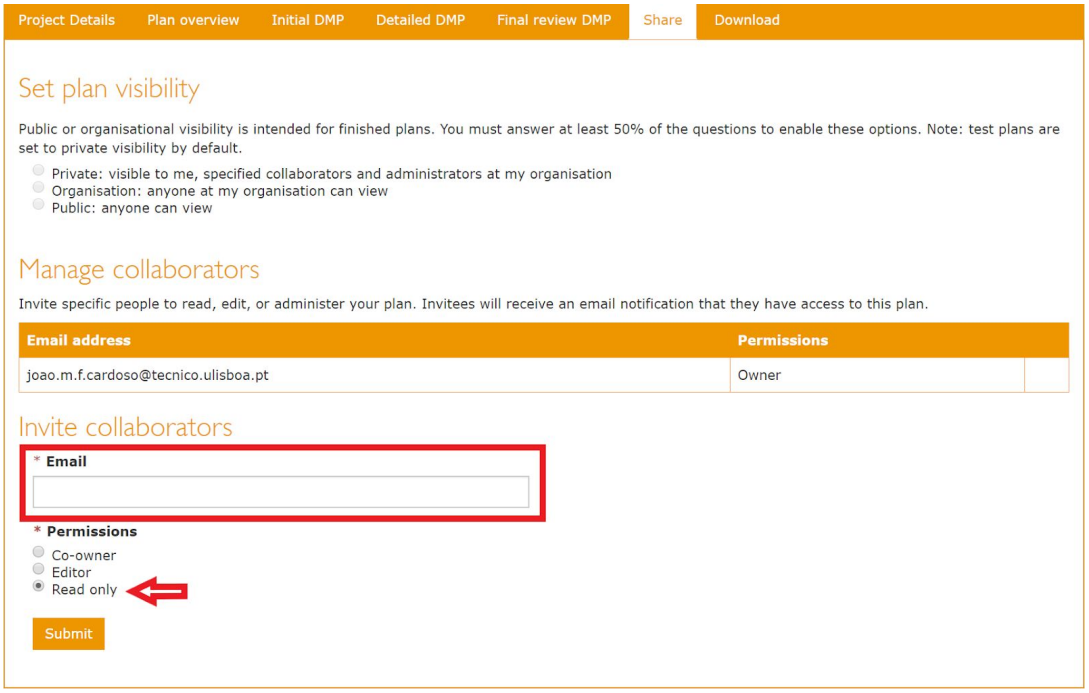

Figure 9. DMP share section

You now need to decide, based on the context of your project, on how to set the plan visibility setting. You can choose between three options:

- Private, visible only to you.
- Organisation, only visible to members of your organisation.
- Public, visible to every user.

In taking your decision, you should consider the types of data, ethical issues that are present in your project.

In accordance to the terms and conditions of this workshop, we now request that you now invite one of the monitors to become a collaborator of your DMP. In order to do this you should input a provided email address, in the appropriate field (see figure 9). Subsequently you should set the permissions for the monitor as “read only” (as indicated in figure 9).

Once you’ve set the sharing definitions, and established a collaboration with a monitor, **click on the “Submit” button**.

You can now proceed to the “Download” section of your DMP Project, as seen in figure 10.

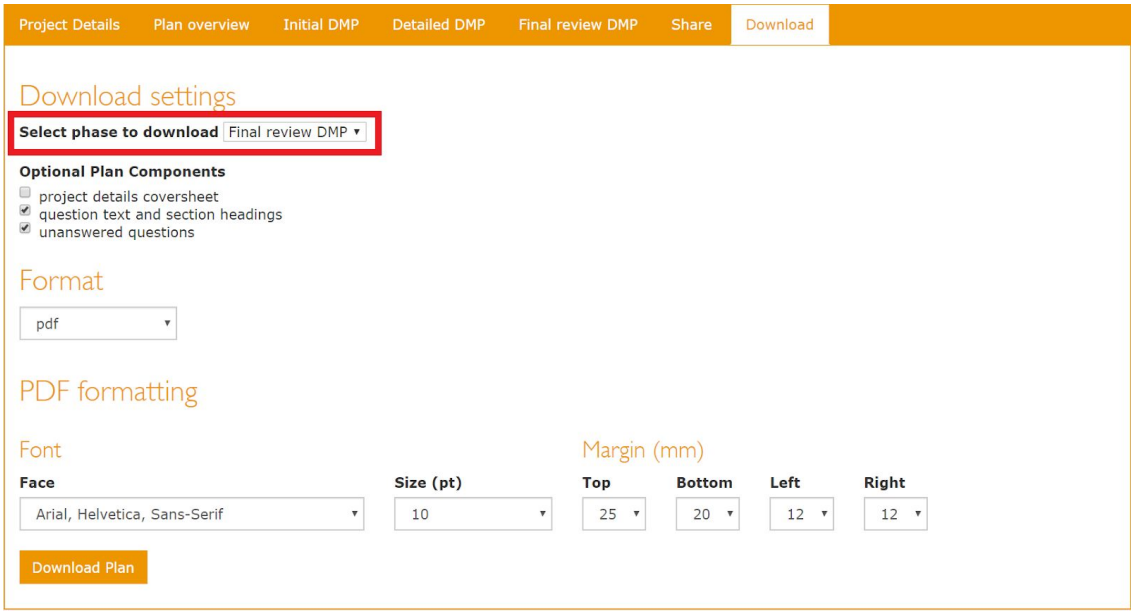

Figure 10. DMP download section.