Study Unit 5

FAIR Data Management

FAIR Data Management Outline

- Research Data Management
- Data Life Cycle and its components
- Different actors in research and their affect for research
- FAIR Data Principles
- Why do we use FAIR Data Principles in Data Management?
- Core Principles for FAIR Data Management

- FAIR Data Management Plan
- Core requirements for FDMP
- Tools for FAIR Data Management Plan
- Creating FAIR Data Management Plan

Study Session Duration

This Study Session requires a 4 hours of formal study time.

You may spend an additional 2-3 hours for revision

Introduction

This unit focus on FAIR Data Management and its core principles. The requirements for a good data management as well as the platform for creating FAIR data is covered.

You will learn what kind of questions you need to refer to make a good Data Management Plan and which tools you might use for creating a FAIR Data Management Plan? Along with that, you will have the practice of creating a FAIR Data Management plan yourself.

Learning Outcomes of Study Unit 5

Upon completion of this study unit, you should be able to:

5.1 Enumerate the importance of managing research data

5.2 Discuss Data Management and its importance

5.3 Demonstrate Data Life Cycle components and what Data Management entails by looking at it.

5.4 Describe FAIR Data Principles, purposes of their usage and identification of elements that help make data FAIR

5.5 Construct data management plan sequentially or step by step.

5.6 Explain core Principles for FAIR Data Management plan, its requirements and compatibility with the FAIR Data Principles.

5.7 Demonstrate ability write and use online platforms for creating FAIR Data Management Plan

5.8 Understand the added value of making data management plans in research projects.

5.1 Introduction to research data management

In this module, firstly, you will understand why data management is necessary to improve the efficiency of the data-driven research process, and why it is necessary to maintain data integrity and obtain reproducible results. Second, you will be provided an overview of all stages involved in the data lifecycle and highlight how they can benefit from effective FAIR Data Management. Third, you will be introduced tools to implement FAIR Data Management plans.

5.1.1 What is Data

In order to understand data management, let's firstly understand data itself. Data can be defined as "facts and statistics collected together for reference and analysis" [1]. This means that data is collected or created for analytical purposes that can lead to the solution of many problems that arise in various areas and aspects of life. Data is becoming an important aspect of life these days as the volume of data produced increases and the better it is curated, the better the knowledge is applied, and therefore the better the results.

5.2 Data Management

Data management is the process of receiving, storing, organizing and maintaining data created and collected during a project and beyond. It concerns the efficient and proper handling of data throughout the life of a project and afterwards. In data management, the concept of the data cycle is often used to help realize the scope and meaning of data management. Figure 1 shows such data lifecycle. This includes planning, collecting, analyzing, publishing / sharing, preserving and reusing. All these stages are required to ensure that data is collected in correct format and well curated and can be reused afterwards.

A good Data Management facilitates clear communication between researchers participating within the research and passing the information to new researchers.

When thinking about and performing data management, it is very important to understand the data in terms of its life cycle and the life cycle of the research project. From project planning to archiving, proper data management occurs throughout the entire research lifecycle.

5.3 Data life cycle

Each stage of the lifecycle produces specific data products and requires a variety of considerations, responsibilities, and activities. Once data are created, which is represented in Figure 1, the data undergo subsequent stages, processing, analyzing, preserving, providing access to, and re-using the data. All these steps enable all stakeholders to make the most of the data produced.

Discovery and Planning - researchers will need to determine what type and format of data they are going to collect. This may involve collecting new data, combining existing data sets or analyzing existing data.

- Define type and format of data
- Consider privacy and confidentiality issues, including data regulations
- Consider documentation and metadata standards that will be used for data description
- Consider potential re-users of the project

Initial Data Collection -

Data Preparation and Analysis

Publication and Sharing

Long-Term Management

Good data management starts early in the data management lifecycle. In order to create a good data management, you need to properly plan everything at the very beginning. Thus, planning where to store your data, and what kind of format of data to use, to whom they are accessible makes a good data management.

Figure 5.1: Data life cycle

Source: https://ddialliance.org/welcome-to-the-data-documentation-initiative

5.4 Fair data principles

Data management is essential driver of research and clinical practice. Collection, storage, access, sharing and analytics depend on the correct and consistent use of data management principles by investigators. Since 2016, the FAIR (findable, accessible, interoperable, and reusable) guiding

principles for research data management have been resonating in scientific communities and since 2020 in clinical practice. The primary goal of FAIR is to reuse scientific data, and thus, making data accessible both for humans and machines.

FAIR Data principles are the way of facilitating knowledge discovery from any data. FAIR is introducing itself not as a standard, but as the set of principles in order to facilitate the process of re-use of data (Mons, 2018). These principles provide guidance for scientific data management and stewardship and are relevant to all stakeholders in the current digital ecosystem. They directly address data producers and data publishers to promote the maximum use of data.

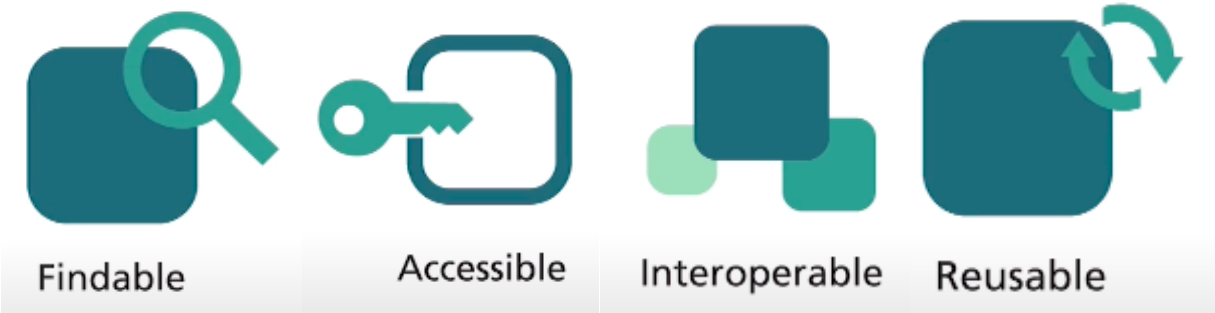

Figure 5.2: FAIR Data Principles

- **Findable** – others can discover your data; Data and metadata should be easily findable by both humans and computers.
- **Accessible** – your data can be made available for others; Users need to know how data can be accessed, possibly including authentication and authorization.
- **Interoperable** – your data can be integrated with other data or can be easily used by machines.
- **Reusable** – your data can be used for new research; metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Enabling data to be findable, accessible, interoperable, and reusable might strengthen data sharing, reduce duplicated efforts, and move towards harmonizing data from heterogeneous disconnected data stores.

In the below given picture we thoroughly learn 15 Principles of FAIR and how they are related to FAIR Data Management.

Figure 5.3 Australian Research Data Commons

5.4.1 Findable

- Meta (data) needs to be assigned with globally unique and persistent identifiers. Globally unique and persistent identifiers allow meta(data) to be discovered globally
- metadata is described in a rich manner so that human and machine can understand what stores the exact dataset
- Metadata consists of the identifier of the data set. Both metadata and data are located in

separately, mentioning a data set's globally unique and persistent identifier in the metadata.

- Data can be found through the search engine (Wilkinson et al., 2016). In this regard, indexing helps to easily find data across the internet and it works for almost all ordinary data, although scholarly research data requires better approach for indexing (Wilkinson et al., 2016).

5.4.2 Accessible

The second aspect is "A" is not necessarily mean open, it means that data is open under specific conditions.

- Meta (data) are retrievable by their identifier using a standardized communication protocol (Wilkinson et al., 2016).
- The protocol is open, free and universally implementable. For data to be maximally reused, the protocol should be with no-cost and open, therefore, globally implementable to promote data retrieval.
- The protocol allows for an authentication and authorization where necessary (Wilkinson et al., 2016). The evaluation tested metadata for the ability to implement authentication and authorization in its resolution protocol.
- Metadata should be accessible even when the data is no longer available (Wilkinson et al., 2016). The evaluation was done to test if the metadata contains a persistence policy, explicitly identified by a persistence Policy key (in hashed data) or a predicate in Linked Data.

5.4.3 Interoperable

- Meta(data) use a formal, accessible, shared, and broadly applicable language for knowledge representation
- Meta (data) use vocabularies that follow the FAIR Principles (Wilkinson et al., 2016). The controllable vocabulary that used to describe datasets needs to be documented and well-regulated using globally unique and persistent identifiers. The documentation should be

easily findable and accessible to anyone interested in using the data set.

- Meta (data) include qualified references to other meta(data) (Wilkinson et al., 2016). To be more precise, specifying if one data set builds on another one, if supplementary data sets are needed to complete the data, or if the corresponding information is stored in another data set. The scientific links connecting the data sets need to be described. Besides, all data sets need to be properly cited, including their globally unique and persistent identifiers.

5.4.4 Reusable

- Meta (data) are released with a clear and accessible data usage license (Wilkinson et al., 2016). Although 'I' aspect covers the components of technical interoperability, R1.1 deals with legal interoperability, covering the aspect of what usage rights attached to data.
- Data (Meta) are associated with detailed provenance.
- Data (Meta) meet domain-relevant community standards.

These principles ensure that (Meta) data is made available from different sources, which is machine-actionable, and ready for use whenever needed in accordance with legal, ethical, disciplinary and regulatory frameworks and regulations.

The FAIR Data Principles should be applied throughout the entire data life cycle, and the planning is made at the very beginning in order to be able to use your data for analysis, so that the use of the data is maximized.

5.5 Data Management Plan (DMP)

Before we get started with the module, let's take a look at what the Data Management Plan is in the data lifecycle. How important is it to draw up a plan and what is the role of FAIR in it?

The term DMP refers to a Data Management Plan, is a formal document developed at the beginning of a project which describes how you will manage your data and document it for the duration of your project and gives instructions on naming conventions, metadata structure, storage of the data, and how to make data available.

The DMP defines strategy that covers produced data, volumes, metadata requirements, data retention periods, data disposal, requirements and tools for processing and analysis. The DMP should clarify all aspects of data management between the facility and users before starting the project.

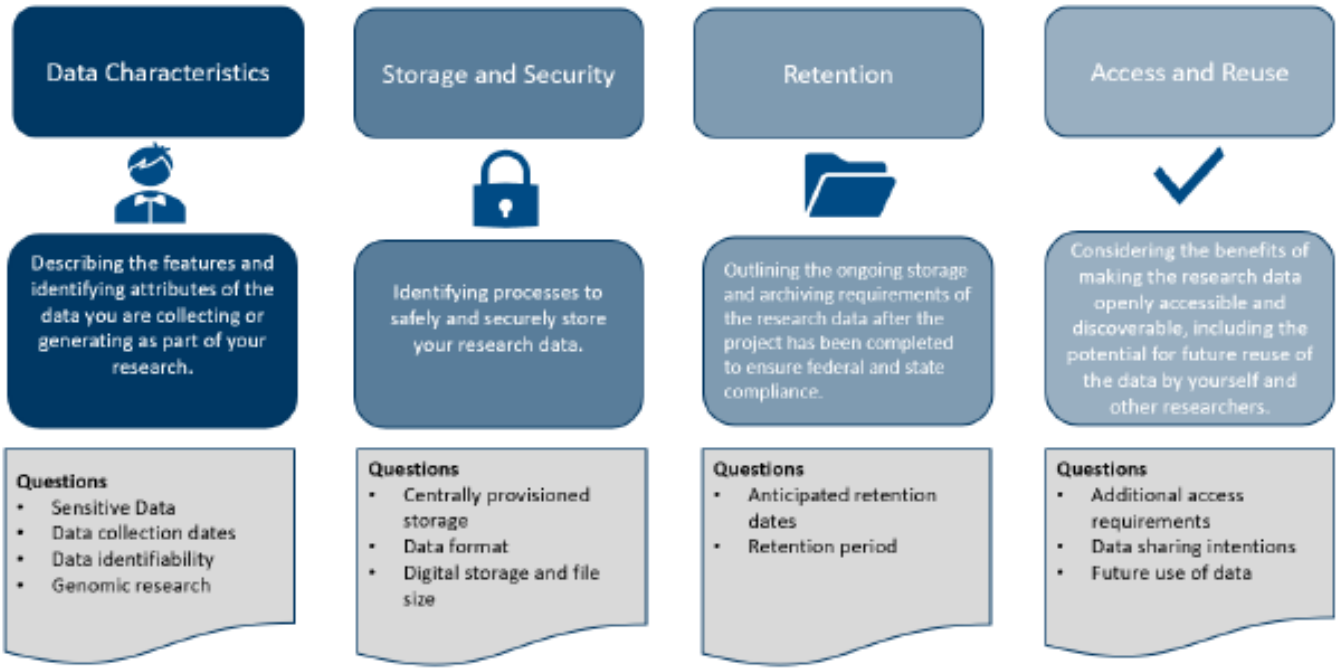

Figure 5.4:

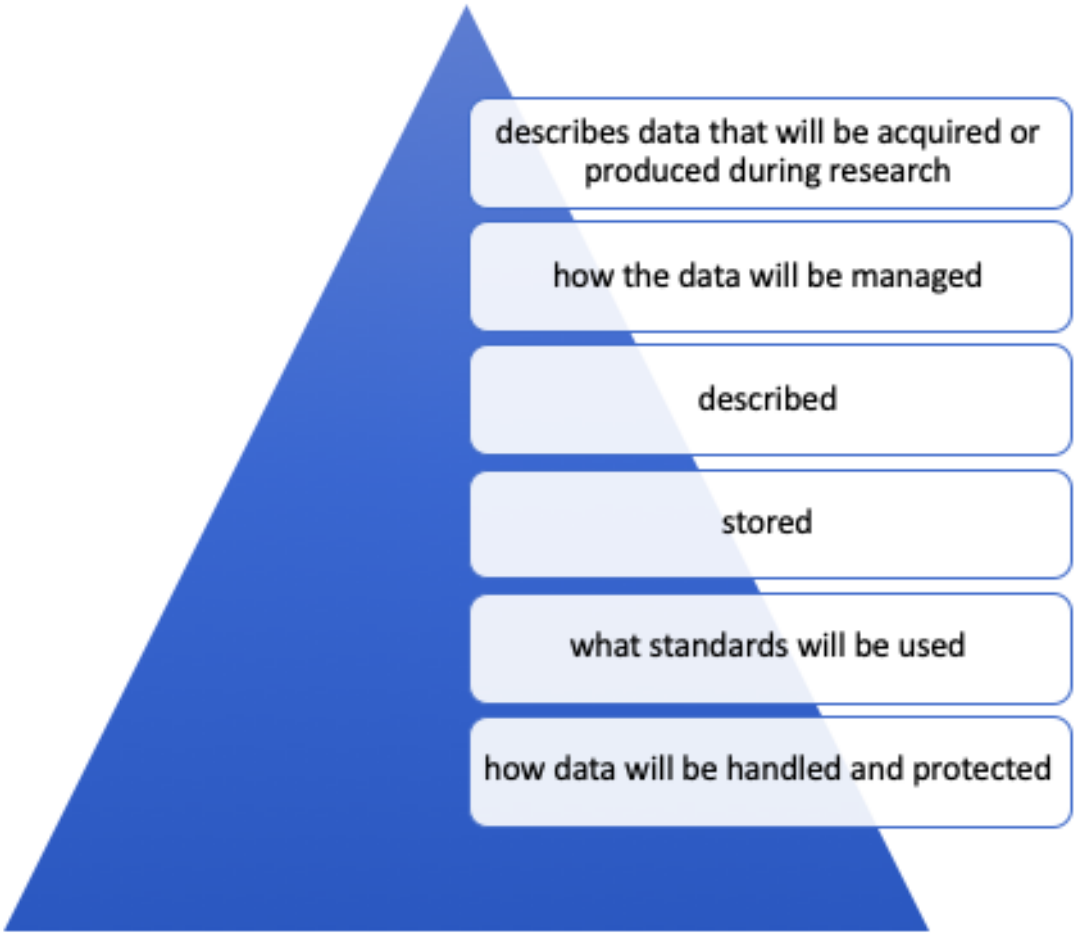

Figure 5.5: Data Management Plan

Thinking about DMP from the beginning of your project will ensure that you are well prepared. The following questions provide a good Data Management Plan.

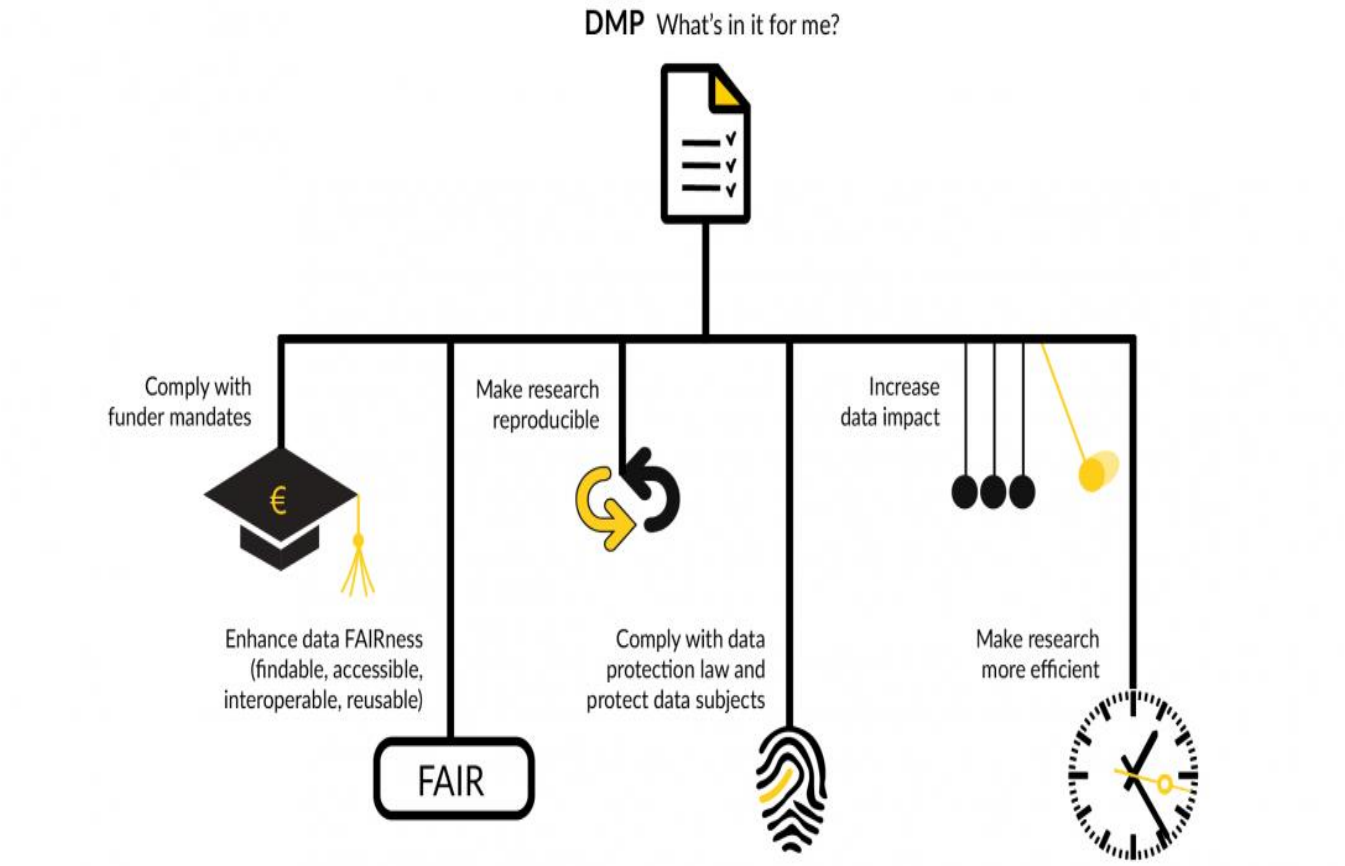

Figure 5.6

https://www.uu.nl/sites/default/files/styles/image_1085x580/public/rdm_benefitsdmp.png?mt=1590481711&itok=BHidbtpy

Step 1. Where will the data come from?

Possible sources of research data will come from interviews, questionnaires, images, lab experiments and so on. Can you list any special tools or software that are required to create, manipulate, or visualize data?

Will your data comply with the GDPR (General Data Protection Regulation) in case if you are carrying out research inside Europe or working with any European data?

Step 2: What format is the data in?

What format will the data be in… RDF, Triple stores, csv, txt, excel, word files, open source, proprietary software. It is important to think about reuse, interoperability and long-term preservation of the data at the beginning stage and make sure the file format facilitates that.

Step 3: How will the data be organized, documented and described?

Data documentation ensures that your data can be understood and interpreted by any user. It will explain how your data was created, what the context for the data is, the structure of the data and its content, and any manipulations that were performed on the data.

Step 4: How will you store the data and ensure it is secure?

Where will the data be stored and what media will be used for storage? Is it just for a short time or are you thinking about saving it for the future?

Step 5: Research Ethics and Intellectual property

Are there ethical or legal reasons why you cannot share the data? For example, you told respondents that data would be anonymized and transferred, and in this regard, remember that there are additional responsibilities under the GDPR (General Data Protection Regulation).

Who owns the data? This is especially important if your project is multi-agency in nature, as ownership needs to be clearly defined from the beginning.

Step 6: Data Sharing

Who will be the audience for the data?

Are there people other than the authors who are allowed to view or use the data?

Step 7: Implementation of the Plan

How will you ensure that the plan is implemented and who will be responsible for keeping it?

How often will you review and update your plan?

Difference between Data Management Plan and FAIR Data Management Plan.

5.6 FAIR Data Management Plan

5.6.1 Guidelines on FAIR Data Management

The full document is here:
https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

In this module, you will work with the Data Stewardship Wizard tool which helps to select domain-specific standards, repositories and data policies as part of an intelligent questionnaire that guides users through the extensive requirements that need to be met to achieve good and FAIR Data management.

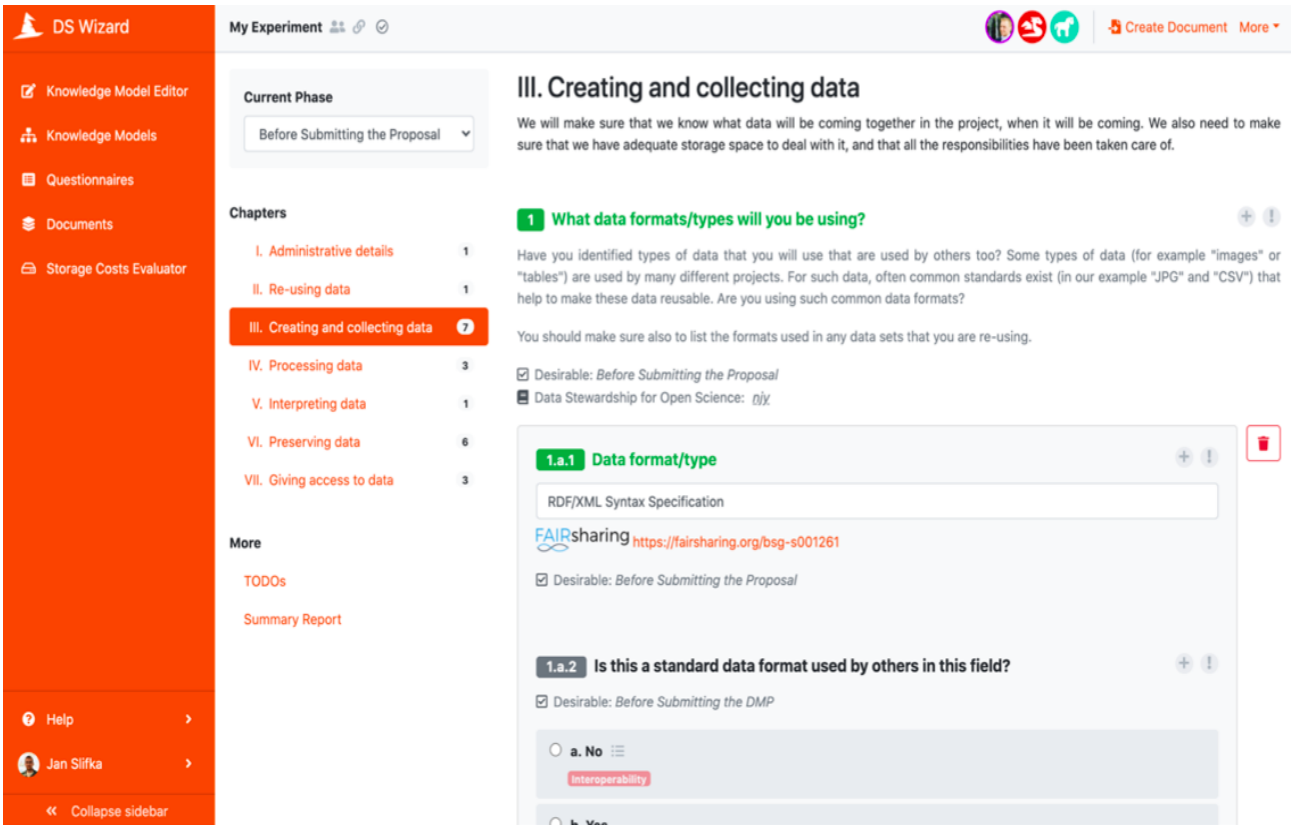

Figure 5.7: DSW Wizard Platform

FAIR Data Management is data administration for producing findable, accessible, interoperable, and re-usable data. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation-solution. Data Management Plans (DMPs) are an essential component of good data management. A DMP describes the data management life cycle for the data to be collected, processed, and/or generated. In general terms, your research data should be 'FAIR', that is findable, accessible, interoperable and re-usable.
Developing data management plans requires dealing with the following topics and the answer the following questions:

5.6.2 Description and collection of data or reuse of existing data

a) How will data be collected or produced?
   i) Define the goal of the data collection or creation
   ii) Provide information about origin of data
   iii) Is existing data re-used?
b) What data will be collected or produced?
   i) Define formats and types of data
   ii) Specify the size of data expected

c) Findable:
   i) What kind of metadata and documentation will be following data?
      (1) Define methodology of data collection and way of organizing data
      (2) Define methodology for creation of search keywords and clear versioning
      (3) Provide identification for your data and refer to the standard mechanism such as Digital Object Identifiers.
   ii) What kind of quality control measures of data will be used?
d) Accessible:
   i) How and when will data be shared? Are there possible restrictions to data sharing and embargo reasons?

(1) Specify is data openly available? In case some data is kept private, provide rationale reason
(2) Provide information how data can be made available?
(3) Provide information about software tools or methods needed to access data?
(4) Provide information about data location where it deposited

e) Interoperable:
   i) Define data and metadata vocabularies, standards, or methodologies to facilitate interoperability

f) Re-usable:
   i) How the data can be reusable?
      (1) Is data will be licensed to permit re-use possible?
      (2) Define the embargo for data
      (3) Is there any restriction to re-use of data? Explain why
      (4) Define period for the data will remain re-usable?

5.6.3 Storage and resource allocation

g) How will data and metadata be stored and backed up during the research process?
h) How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?
i) How costs of making your data FAIR is covered?
   - Describe data long-term preservation costs and coverage
   - Define potential value of data

5.6.4 Data security and ethical requirements, codes of conduct

j) How will data security and protection of sensitive data be taken care of during the research?
k) If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?

l) How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

m) How will possible ethical issues can be taken into account, and codes of conduct followed?

5.7 Platforms for creation a FAIR Data Management Plan

5.7.1 Data Stewardship Wizard

In order to write a data management plan, many researchers use an online tool such as Data Stewardship Wizard. It provides many features such as guidance with smart questionnaires, FAIR metrics, online collaboration and many other. In order to create DMP with this tool:

1. Go to the official website of Data Stewardship Wizard.
2. Sign in or create account
3. Create project and fill in the fields as follows (see figure below):
    1. Write a name of the project
    2. Select the knowledge model from the list
    3. Select tags:
        - If you select some of them, you will be given a questions based on those tags
        - if you select none of them, you will be given all questions from all tags

    Tags are assigned according to the projects: Horizon 2020, Science Europe and the last one maDMP that is machine-actionable. The ladder one is opposite for those static documents mostly in free text, which means they are not machine-actionable and cannot be easily exchanged across research tools and systems.

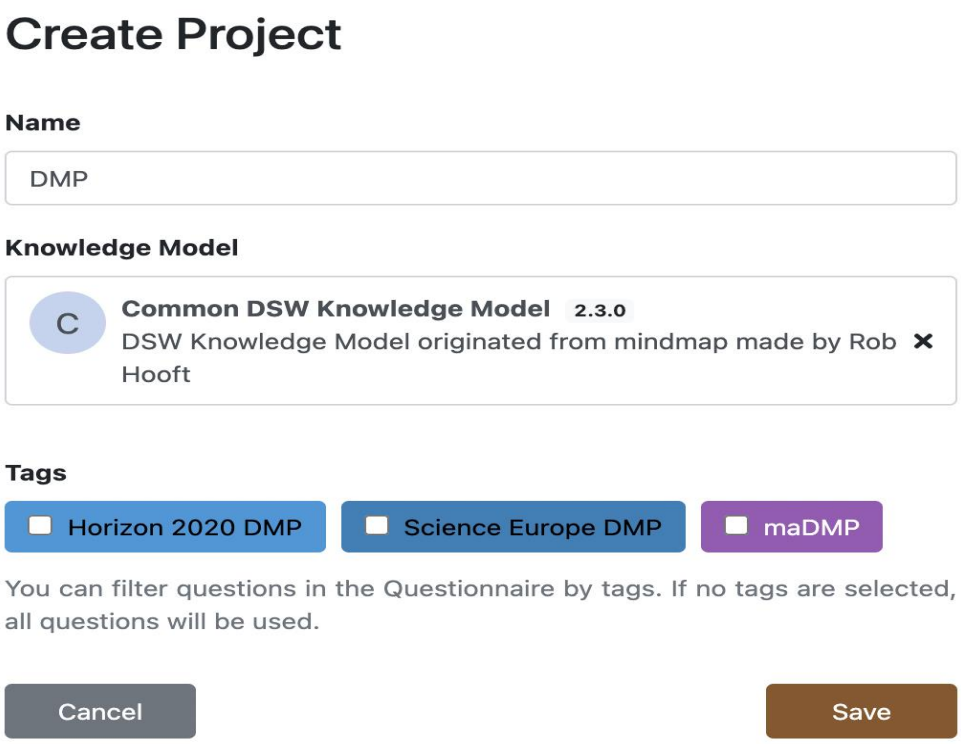

Figure 5.8: DSW project creation.

4. After successful completion of previous steps, you will see the main project page as shown below. It has five essential parts:
    1. Left-side menu where you can switch between projects and their knowledge model
    2. Tabs on the top of the page where you can use features of website such as:
        - Questionnaire – all questions regarding how to make data FAIR is inside of this section
        - TODOs – list of questions you selected as TODO
        - Metrics – metric information about how FAIR are your data
        - Preview – it is preview of machine-actionable formats of DMP like in JSON
        - Documents – it is section where DMP can be downloaded
        - Settings – section where default template, formats and name can be changed

- Share (on the right top corner) – share project with others

3. Content part – where you can see questions and fields to fill answers.
4. Chapters (inside the Questionnaire section) – FAIR principles distributed to separate chapters and each chapter provide question regarding those chapters.
5. Current phase (inside the Questionnaire section) – it highlights the desirable questions to answer for current stage. It provides three stages:
    - Before submitting the Proposal
    - Before submitting the DMP
    - Before finishing the Project

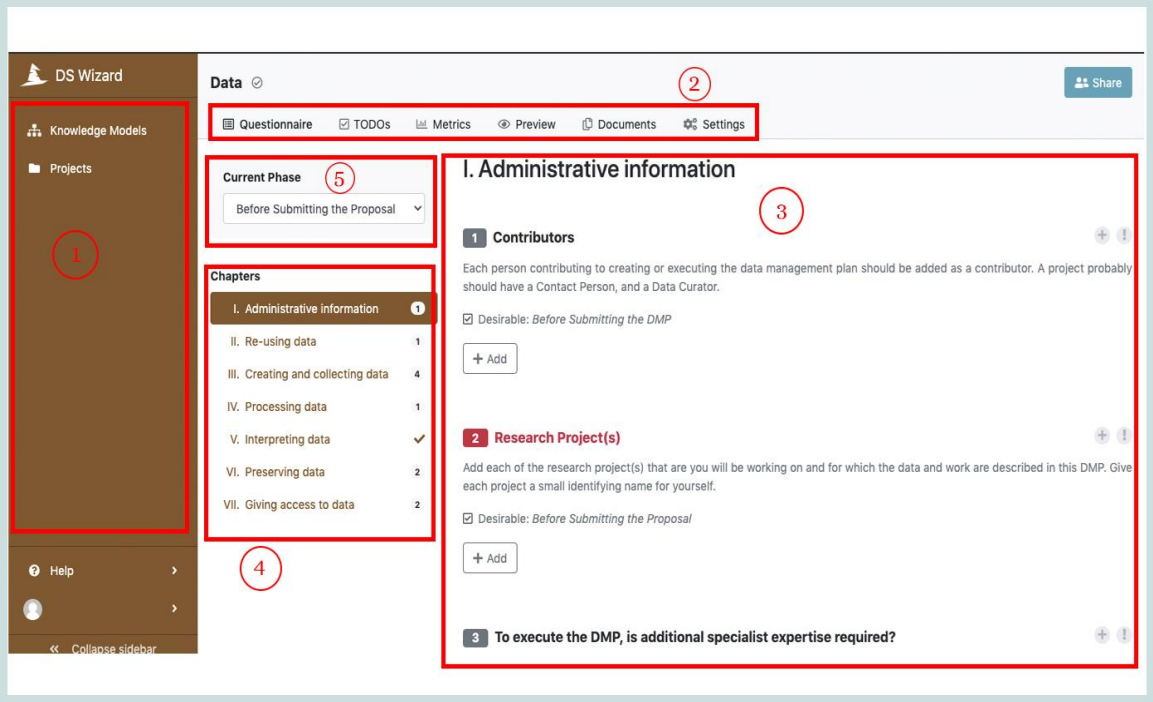

Figure 5.9: DSW Essential parts

The chapters as an essential part of the Questionnaire section. It has seven chapters:

1. Administrative information - contains questions regarding the research project details, information about contributors, and expertise required. Questions make sure that data is findable, and it has a persistent identifier and rich metadata.

2. Re-using data - information about pre-existing data
3. Creating and collecting data - collecting information regarding data and storage space required for data
4. Processing data - questions regarding automatically processing data
5. Interpreting data - the last step of processing data, and it requires visualization and data integration as questions regarding interoperability will come
6. Preserving data - question regarding data publication and long-term archiving
7. Giving access to data - questions regarding access to data.

The chapters are static and only questions inside the chapters can be different based on the tags you select.

5.8 Data Management Plan Online

DMP Online is also one of the popular tools to create a DMP. It provides many templates which are ready to use. Many researcher funders use this tool and if you choose one of the funders form list then you will be given a template of this funder. In order to create you DMP:

1. Go to website by link https://dmponline.dcc.ac.uk/
2. Create account or sign with your organizational or institutional account
3. Then you need to fill your title of research project. Then, you can fill research and funding organizations in case it is needed. Otherwise, you can select “No” check list as shown below.

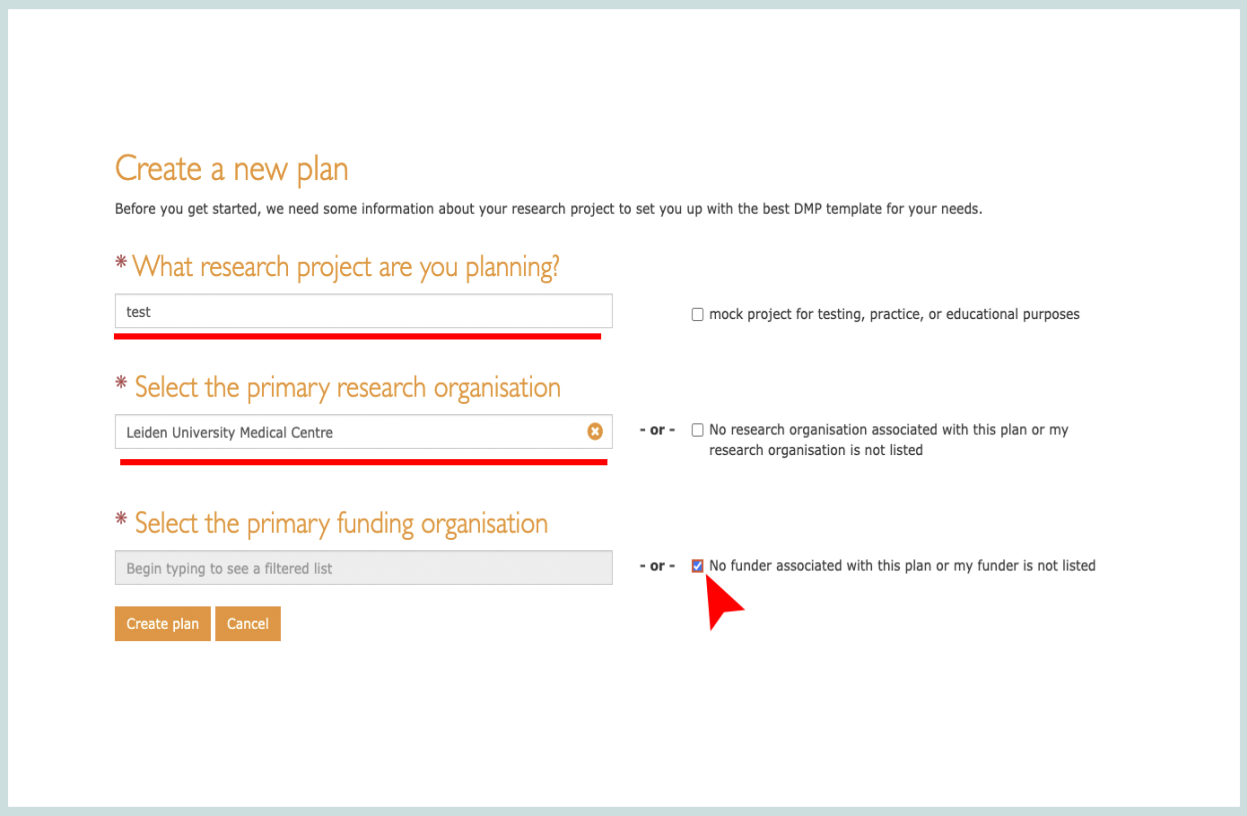

Figure 5.10: DMP online creation

4. Then you will be given a DMP overview plan as shown next image. Each section has questions to answer. Note that you should start filling as much as you can. It has to be completed during your research project, and try to update plans after some period of time

Figure 5.11: DMP overview plan

For instance, 6, 12 months. Then, you also can see additional questions to consider on the right side of the page. Besides, you can also read the guidance regarding questions as shown below.

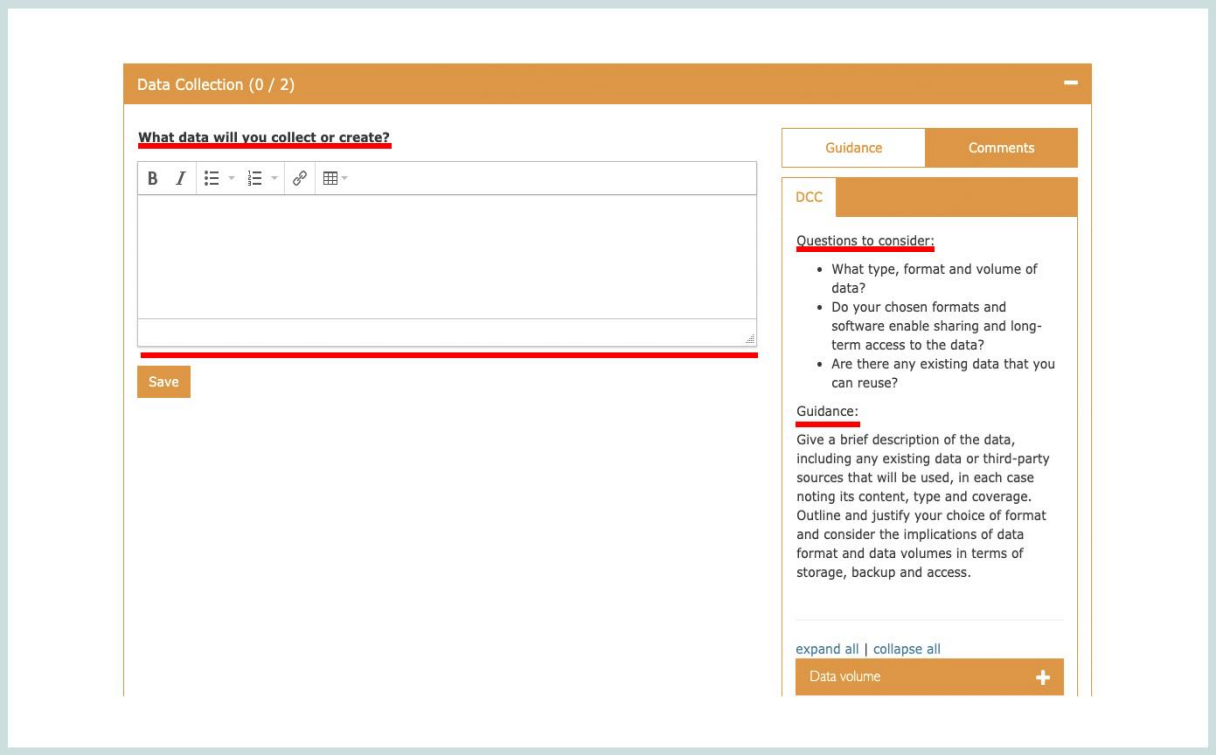

Figure 5.12: Additional questions to consider and guidance section

5. You can update your DMP or you can create a new plan. Then, you will see the list of your templates under your dashboard.
6. You can also share your DMP with others in order to get feedback. To share your DMP with fellow researchers or a data supporter. To do that go to the tab Share, fill in the e-mail address of the specific person and give edit rights as shown below.

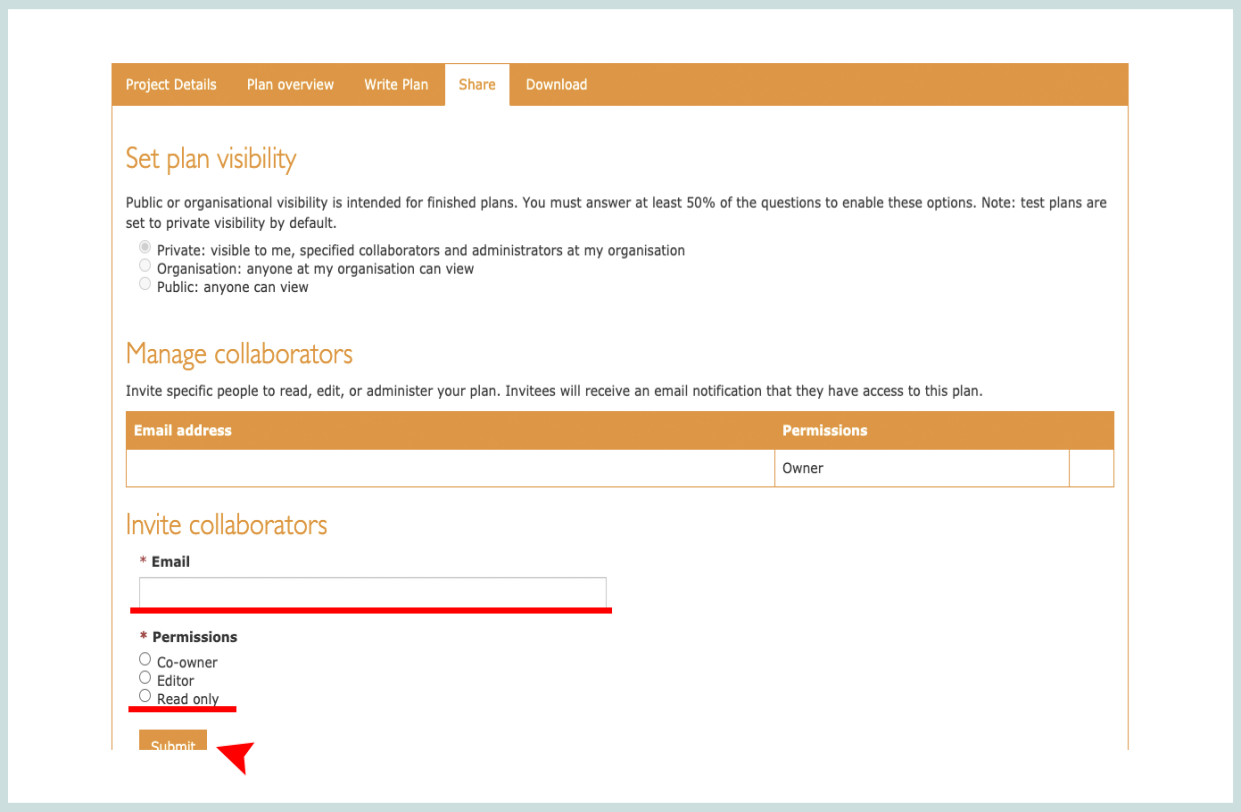

Figure 5.13: Sharing DMP with others

7. You can also download DMP in a Word or PDF format. To download go to the tab Download. Then, you are able to choose components that you want to see in file, choose format, and change style based on your or organization preferences as shown below.

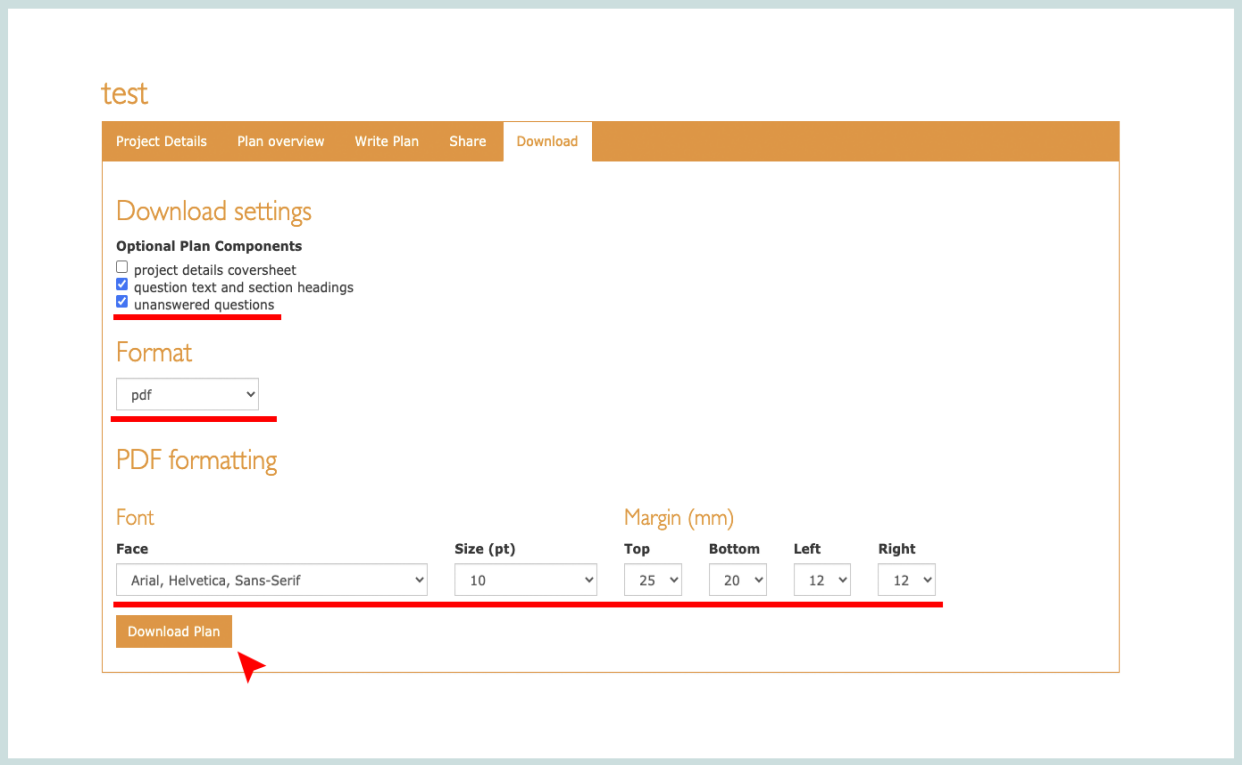

Figure 5.14: DMP download in a Word or PDF format.

References

[1] Data [Internet] Oxford, UK: Oxford University Press; 2014. [cited 25 Feb 2021]. < http://www.oxforddictionaries.com/us/definition/american_english/data>. [Google Scholar]

[1] *https://intranet.ecu.edu.au/research/for-research-staff/research-integrity/responsible-research/research-data-management/data-management-plan*