

Core DS Knowledge Model

Organization	ELIXIR Norway
Created by	Rukaya Sarah Johaadien (rukayasj@uio.no)
Based on	RI gap analysis, 0.0.1 (elixir.no:ri-elixir-norway:0.0.1)
Project Phase	Before Submitting the DMP
Created at	11 Mar 2021

I. Administrative details

Report

Indications

Answered (current phase)	13 / 13
Answered	22 / 22

Metrics

No metrics for this chapter.

Questions

1 Contributors

Each person contributing to creating or executing the data management plan should be added as a contributor. A project probably should have a Contact Person, and a Data Curator.

Tags: *maDMP, Science Europe DMP*

Answers

1.b.1 Name

Tags: *maDMP, Science Europe DMP*

✓ *Rukaya Johaadien*

1.b.2 E-mail address

Tags: *maDMP, Science Europe DMP*

✓ *rukayasj@uio.no*

1.b.3 ORCID Identifier

Tags: *maDMP, Science Europe DMP*

✓ *0000-0002-2857-2276*

1.b.4 Affiliation

Tags: *Science Europe DMP*

✓ *University of Oslo*

1.b.5 Role

Roles in a project should be given as they are defined by [datacite](#).

You should specify at least one "Contact Person". If your project has a work package for data management, identify the leader of that work package as "Data Curator".

Tags: *maDMP, Science Europe DMP*

✓ *a. Contact Person*

1.c.1 Name

Tags: *maDMP, Science Europe DMP*

✓ *Rukaya Johaadien*

1.c.2 E-mail address

Tags: *maDMP, Science Europe DMP*

✓ *rukayasj@uio.no*

1.c.3 ORCID Identifier

Tags: *maDMP, Science Europe DMP*

✓ *0000-0002-2857-2276*

1.c.4 Affiliation

Tags: *Science Europe DMP*

✓ *University of Oslo*

1.c.5 Role

Roles in a project should be given as they are defined by [datacite](#).

You should specify at least one "Contact Person". If your project has a work package for data management, identify the leader of that work package as "Data Curator".

Tags: *maDMP, Science Europe DMP*

✓ *a. Contact Person*

2 RI

Add each of the project(s) that you will be working on and for which the data and work are described in this DMP. Give each project a small identifying name for yourself.

Tags: *maDMP, Science Europe DMP*

Answers

2.b.1 RI name

Tags: *maDMP, Science Europe DMP*

✓ GBIF Norway - Radiolaria dataset

2.b.2 Project short discription

Tags: *maDMP, Science Europe DMP*

✓ This dataset is the result of a radiolarian sampling program from Korsfjorden (60.19110 N,5.23136 E), on the west coast of Norway near Bergen, that took place from November 1969 and ended in May 1971. At the sampling location the fjord depth is 670 m. The water column was divided into 100 m depth zones (100-0 m, 200-100 m,600-500 m, and 650-600 m) except for the deepest zone where the sampling started ca. 20 m above the mud line to avoid contamination of the net. At each sampling date all depth zones were sampled. A plankton net with 63 µm mesh opening was used. This dataset also includes the phaeodarians found in this dataset. A total of 72 species was observed. A discussion of this dataset can be found in: Bjørklund, K.R., 1974. The seasonal occurrence and depth zonation of radiolarians in Korsfjorden, Western Norway. *Sarsia*, 56: 13-42.

2.b.3 Date the RI will started

Tags: *Science Europe DMP, maDMP*

✓ 2010

2.b.4 Date the RI funding will end

Tags: *Science Europe DMP, maDMP*

✓ 2020

2.b.5 Funding

Add all the funding that are part of this project.

Tags: *maDMP, Science Europe DMP*

Answers

2.b.5.b.1 Funder

Specify the name of the funder that you ask for funding for your project. If the funder is not present in the suggested list, please specify a complete URL to the funder web site.

Tags: *Science Europe DMP, maDMP*

✓ *Norges Forskningsråd*

 <http://dx.doi.org/10.13039/501100005416>

2.b.5.b.2 Funding status

Tags: *Science Europe DMP, maDMP*

✓ *c. Granted*

2.b.5.b.3 Grant number

Tags: *maDMP, Science Europe DMP*

✓ *Unsure*

3 To execute the DMP, is additional specialist expertise required?

Tags: *Science Europe DMP*

✓ *a. No*

4 Do you require hardware or software in addition to what is currently available in the participating institutions?

✓ *a. No*

II. Re-using data

Before you decide to embark on any new study, it is nowadays good practice to check all options to re-use existing available data, either collected or generated by yourself in an earlier project, or data from others (Barend Mons calls this "Other People's Data And Services" or OPEDAS). This can include reusable data that have been created for an earlier study, and also so-called "reference data" which is used by many projects.

It is not because we can generate massive amounts of data that we always need to do so. Creating data with public money is bringing with it the responsibility to treat those data well and (if potentially useful) make them available for re-use by others. And the circle is only complete if such data is actually re-used.

Report

Indications

Answered (current phase)	5 / 5
Answered	5 / 5

Metrics

No metrics for this chapter.

Questions

1 Describe the utility of data produced at the RI; to whom might it be useful?

✓ *Any scientific researcher using radiolaria biodiversity data*

2 Is there pre-existing data?

Are there any data sets available in the world that are relevant to your planned research?

Tags: *maDMP, Science Europe DMP*

Data Stewardship for Open Science: [*atq*](#)

✓ *b. Yes*

2.b.1 Will you be using any pre-existing data (including other people's data)?

Will you be referring to any earlier measured data, reference data, or data that should be mined from existing literature? Your own data as well as data from others?

Tags: *maDMP, Science Europe DMP*

Data Stewardship for Open Science: [*ezi*](#)

✓ *a. No*

Did you research all the data that exists? You may not be aware of all existing data that could be available. Although using and/or integrating existing data sets may pose a challenge, it will normally be cheaper than collecting everything yourself. Even if you decide not to use an existing data set, it is better to do this as a conscious decision.

2.b.2 Do you need to harmonize different sources of existing data?

If you are combining data from different sources, harmonization may be required. You may need to re-analyse some original data.

☰ Data Stewardship for Open Science: [wht](#)

✓ a. No

2.b.3 Will you be using data that needs to be (re-)made computer readable first?

Some old data may need to be recovered, e.g. from tables in scientific papers or may be punch cards.

☰ Data Stewardship for Open Science: [pth](#)

✓ a. No

III. Creating and collecting data

We will make sure that we know what data will be generated at the RI and when it will be generated. We also need to make sure that there will be adequate storage space to deal with it, and that all the responsibilities have been taken care of.

Report

Indications

Answered (current phase)	31 / 31
Answered	36 / 37

Metrics

Metric	Score
Findability	0.67
Accessibility	0.5
Interoperability	1
Reusability	1

Questions

1 What data formats/types will you/your users be using?

Have you identified types of data that you will use that are used by others too? Some types of data (for example "images" or "tables") are used by many different projects. For such data, often common standards exist (in our example "PNG" and "CSV") that help to make these data reusable. Are you using such common data formats?

You should make sure also to list the formats used in any data sets that you are re-using.

🔖 Tags: *Science Europe DMP*

☰ Data Stewardship for Open Science: [njy](#)

Answers

1.b.1 Data format/type

🔖 Tags: *Science Europe DMP*

✓ *Darwin Core*



<https://fairsharing.org/bsq-s000195>

1.b.2 Is this a standard data format used by others in this field?

Tags: *Science Europe DMP*

✓ *b. Yes*

1.b.3 Does this data format enable sharing and long term archiving?

Complicated (binary) file formats tend to change over time, and software may not stay compatible with older versions. Also, some formats hamper long term usability by making use of patents or being hampered by restrictive licensing.

Tags: *Science Europe DMP*

✓ *b. Yes*

1.b.4 What volume of data of this type will you be working with?

Tags: *Science Europe DMP*

✓ *a. So small that it is not a problem*

2 Will you/your users be using new types of data?

Sometimes the type of data you collect can not be stored in a commonly used data format. In such cases you may need to make your own, keeping interoperability as high as possible.

Data Stewardship for Open Science: [ikk](#)

✓ *a. No, all of my data will fit in common formats*

3 How will you/your users be storing metadata?

For the re-usability of your data by yourself or others at a later stage, a lot of information about the data, how it was collected and how it can be used should be stored with the data. Such data about the data is called metadata, and this set of questions are about this metadata.

[SEEK](#) is a webtool to store (meta)data and provenance. The public global instance [FAIRDOMHub](#) is free to users in Norway. SEEK can be integrated with the data storage and analysis platform for users in Norway [NeLS](#).

Data Stewardship for Open Science: [rhm](#)

External Links: [SEEK](#)

✓ *a. Explore*

3.a.1 Do suitable 'Minimal Metadata About ...' (MIA...) standards exist for your experiments?

External Links: [FAIRsharing repository of standards](#)

✓ *b. Yes*

3.a.2 Do you know how and when you will be collecting the necessary metadata?

Often it is easiest to make sure you collect the metadata as early as possible.

External Links: [FAIRsharing repository of standards](#)

✓ *b. Yes*

3.a.3 Will you consider re-usability of your data beyond your original purpose?

Adding more than the strict minimum metadata about your experiment will possibly allow more wide re-use of your data, with associated higher data citation rates. Please note that it is not easy for yourself to see all other ways in which others could be reusing your data.

✓ *b. Yes, I will document more metadata than needed for reproducibility*

3.a.3.b.1 How will you balance the extra efforts with the potential for added reusability?

✓ *c. I will collect all metadata I can gather and document the data set beyond minimal standards*

3.a.3.b.2 Do you need to exchange your data with others?

✓ *b. Yes*

3.a.4 Did you consider how to monitor data integrity?

Working with large amounts of heterogenous data in a larger research group has implications for the data integrity. How do you make sure every step of the workflow is done with the right version of the data? How do you handle the situation when a mistake is uncovered? Will you be able to redo the strict minimum data handling?

☰ Data Stewardship for Open Science: [spg](#)

✓ *a. Explore*

3.a.4.a.1 Will you be keeping a master list with checksums of certified/correct/canonical/verified data?

Data corruption or mistakes can happen with large amounts of files or large files. Keeping a master list with data checksums can be helpful to prevent expensive mistakes. It can also be helpful to keep the sample list under version control forcing that all changes are well documented.

✓ *b. Yes*

3.a.4.a.2 Will you define a way to detect file or sample swaps, e.g. by measuring something independently?

This will dependent on the applied methods. Examples could include e.g. verifyBamID for known genotypes

✓ *a. No*

3.a.5 Do all datasets you work with have a license?

It is not always clear to everyone in the project (ad outside) what can and can not be done with a data set. It is helpful to associate each data set with a license as early as possible in the project. A data license should ideally be as free as possible: any restriction like 'only for non-commercial use' or 'attribution required' may reduce the reusability and thereby the number of citations. If possible, use a computer-readable and computer actionable license.

✓ *b. Yes*

3.a.5.b.1 Will you store the licenses with the data at all time?

It is very likely that data will be moved and copied. At some point people may lose track of the origins. It can be helpful to have the licenses (of coarse as open as possible) stored in close association with the data.

☰ Data Stewardship for Open Science: [atw](#)

✓ *b. Yes*

3.a.6 How will you keep provenance?

To make your experiments reproducible, all steps in the data processing must be documented in detail. The software you used, including version number, all options and parameters. This information together for every step

of the analysis is part of the so-called data provenance. There are more questions regarding this in the chapter on data processing and curation.

✓ *b. Our work flow system documents the provenance automatically and completely*

3.a.7 How will you do file naming and file organization?

Putting some thoughts into file naming can save a lot of trouble later.

✓ *a. Explore*

3.a.7.a.1 Did you make a SOP (Standard Operating Procedure) for file naming?

It can help if everyone in the project uses the same naming scheme.

✓ *a. No*

3.a.7.a.2 Will you be keeping the relationships between data clear in the file names?

Advice: Use the same identifiers for sample IDs etc throughout the entire project.

✓ *b. Yes*

3.a.7.a.3 Will all the metadata in the file names also be available in the proper metadata?

The file names are very useful as metadata for people involved in the project, but to computers they are just identifiers. To prevent accidents with e.g. renamed files metadata information should always also be available elsewhere and not only through the file name.

✓ *b. Yes, all metadata is also explicitly available elsewhere*

4 Please specify what data you will acquire using measurement equipment

You can use any name for the data set, make sure that it identifies the data set to yourself.

🔖 Tags: *Science Europe DMP*

Answers

4.b.1 Who will do the measurements? And where?

🔖 Tags: *Science Europe DMP*

✓ *b. Users with their own equipment*

4.b.2 Instruments used for data collection

Specify what technical instruments you are using to collect the data.

🔖 Tags: *Science Europe DMP*

Answers

4.b.2.b.1 Instrument name

🔖 Tags: *Science Europe DMP*

✓ *plankton mesh*

4.b.2.b.2 Instrument description

Tags: *Science Europe DMP*

✓ *A plankton net with 63 µm mesh opening was used.*

4.b.3 Is the equipment completely standard and well described?

If the technology is very much under development, you may want to come back later to understand exactly how the measurements have been made. Is the measurement equipment and protocol sufficiently standard that you will be able to explain how it is done or refer to a standard explanation?

Tags: *Science Europe DMP*

✓ *a. Very well described and known*

4.b.4 Is special care needed to get the raw data ready for processing?

Where does the data come from? And who will need it? Sometimes the raw data is measured somewhere else than where the primary processing is taking place. In such cases the ingestion or transport of the primary data may take special planning. You also need to make sure that data is secure and that data integrity is guaranteed.

✓ *a. No, this is all fine*

4.b.5 Will you be using quality processes?

Tags: *Science Europe DMP*

✗ **This question has not been answered yet!**

5 Do you have any non-equipment data capture?

Does the data you collect contain non-equipment captured data such as questionnaires, case report forms, electronic patient records?

Tags: *Science Europe DMP*

Data Stewardship for Open Science: [ybw](#)

✓ *a. No*

6 Is there a data integration tool that can handle and combine all the data types you are dealing with in your RI?

✓ *b. Yes*

6.b.1 What software will you be using to collect all data?

✓ *d. Software not listed here*

7 Will you be storing physical samples?

Data Stewardship for Open Science: [kuz](#)

✓ *a. No*

8 Will you need consent for any newly collected personal data?

Tags: *maDMP, Science Europe DMP*

External Links: [NSD Information and consent](#), [REC Informed consent](#)

✓ *a. No, We do not collect any new personal data*

9 How is the ownership of the collected data arranged?

Tags: *Science Europe DMP*

✓ *e. Other*

9.e.1 How else is the intellectual property for the data collected in the project arranged?

Tags: *Science Europe DMP*

✓ *Dataset owners choose their own sharing license*

IV. Data sensitivity

Ethical and legal issues

adapted from 2019 version of [NSD DMP tool](#) and [Tryggve Checklist on ELSI issues and GDPR compliance](#)

Report

Indications

Answered (current phase)	2 / 2
Answered	2 / 3

Metrics

No metrics for this chapter.

Questions

1 Will you collect or generate data about people?

✓ *b. No*

2 Will the RI follow any institutional policies, codes of conducts or other ethical guidelines?

Each researcher has an independent responsibility for making sure that the research is being carried out in accordance with general scientific and ethical principles and guidelines. For an overview of general and subject-specific research ethics guidelines, see the [Norwegian National Research Ethics Committees](#). Note that in multidisciplinary projects it may be relevant to look to guidelines for several subject areas. In addition, the [Research Ethics Act](#) applies to all research in Norway. Also, check which guidelines apply to your institution.

✓ *b. No*

3 Other ethical / legal issues.

✗ **This question has not been answered yet!**

V. Processing data

In the processing phase, the data will be undergoing the mostly automated steps for processing, before the analysis and interpretation.

Report

Indications

Answered (current phase)	30 / 30
Answered	36 / 36

Metrics

Metric	Score
Accessibility	1
Reusability	0.63
Good DMP Practice	0.87

Questions

1 Will you be providing the data to the user through a shared working space ?

Will you be using a working space that is shared between all the people working on the data in the project? Sometimes such a system is called a *Virtual Research Environment*.

Tags: *Science Europe DMP*

✓ a. No

1.a.1 Are data that users store themselves adequately backed up and traceable?

Tags: *Science Europe DMP*

✓ b. Yes, protected against both equipment failure and human error

2 Data storage systems and file naming conventions

It is a good idea to pre-define how data will be organised in the project work space, and to set conventions for how any data files and folders will be named.

Tags: *Science Europe DMP*

✓ a. Explore

2.a.1 Are you using a filesystem with files and folders?

Are some of the data in the project stored in a filesystem with files and folders?

Tags: *Science Europe DMP*

✓ a. No

2.a.2 Will you be storing data in an "object store" system?

Tags: *Science Europe DMP*

✓ a. No

2.a.3 Will you use a relational database system to store project data?

Tags: *Science Europe DMP*

✓ a. No

2.a.4 Will you use a graph database for data in the project?

Tags: *Science Europe DMP*

✓ a. No

2.a.5 Will you be storing data in a triple store?

Tags: *Science Europe DMP*

✓ b. Yes

3 Workflow development

It is likely that you will be developing or modifying the workflow for data processing. There are a lot of aspects of this workflow that can play a role in your data management, such as the use of an existing work flow engine, the use of existing software vs development of new components, and whether every run needs human intervention or whether all data processing can be run in bulk once the work flow has been defined.

✓ a. *This has been arranged*

4 How will you make sure to know what exactly has been run?

✓ a. *Explore*

4.a.1 Will you keep results together with all processing scripts or workflows including documentation of the versions of the tools that have been run?

✓ b. Yes

4.a.2 Will you make use of the metadata fields in your output data files to register how the data was obtained?

File formats like VCF (for genetics) and TIFF (for images) have possibilities to document metadata in the file header. It is a good idea to use work flow tools that use these fields to document what was done to obtain the data.

✓ b. Yes

4.a.3 Will you use a central repository for all tools and their versions as used in your RI/for each user project?

Especially if analysis and processing of data in the project is done on multiple different computers by different people, it is a good idea to have your own repository of tools and their blessed versions.

Data Stewardship for Open Science: [p2q](#)

✓ b. Yes

4.a.4 Will you use a central repository for reference data used at your RI?

Especially if analysis and processing of data in the project is done on multiple different computers by different people, it is a good idea to have your own repository of reference data versions.

Data Stewardship for Open Science: [p2q](#)

✓ b. Yes

4.a.5 Will you make use of standard workflow engines and automatic workflows for all data analysis at your RI?

It is much easier to guarantee consistency and reproducibility if all data processing is done using automated work flows, especially if the workflow engine automatically keeps adequate provenance data.

✓ *b. Yes*

4.a.6 Are all software tools in the workflow professionally maintained, with version control?

Will you be able to find and reproduce exactly which version was used for any analysis? Not only for the major tools in the workflows, but also for all 'glue' code and small tools you created especially for the project?

✓ *b. Yes*

5 How will you validate the integrity of the results?

✓ *a. Explore*

5.a.1 Will you run a subset of your jobs several times across the different compute infrastructures you are using?

There are surprisingly many complications that can cause (slight) inconsistencies between results when workflows are run on different compute infrastructures. A good way to make sure this does not bite you is to run a subset of all jobs on all different infrastructure to check the consistency.

✓ *a. No*

5.a.2 Will you be instrumenting the tools into pipelines and workflows using automated tools?

Surrounding all tools in your data processing and analysis workflows with the 'boilerplate' code necessary on the computer system you are using is tedious and error prone. Especially if you are using the same tools in multiple different work flows and/or on multiple different computer architectures. Automated instrumentation, e.g. by using a workflow management system, can prevent many mistakes.

✓ *a. No*

5.a.3 Will you use independently developed duplicate tools or workflows for critical steps to reduce or eliminate human errors?

Validation of results without a golden standard is very hard. One way of doing it is to develop two solutions for a problem (two independent workflows or two independently developed tools) to check whether the results are identical or comparable.

✓ *a. No*

5.a.4 Will you run part of data sets repeatedly to catch unexpected changes in results?

Running a small subset of the data repeatedly can be useful to catch unexpected problems that would otherwise be very hard to detect.

 [Data Stewardship for Open Science: *egv*](#)

✓ *a. No*

6 Do you need to do compute capacity planning?

If you require substantial amounts of compute power, amounts that are not trivially absorbed in what you usually have available, some planning is necessary. Do you think you need to do compute capacity planning?

✓ *a. No*

7 Is the risk of information loss, leaks and vandalism acceptably low?

There are many factors that can contribute to the risk of information loss or information leaks. They are often part of the behavior of the people that are involved in the project, but can also be steered by properly planned

infrastructure.

Tags: *Science Europe DMP*

✓ *a. Explore*

7.a.1 Do RI members store data or software on computers in the lab or external hard drives connected to those computers?

When assessing the risk, take into account who has access to the lab, who has (physical) access to the computer hardware itself. Also consider whether data on those systems is properly backed up

Tags: *Science Europe DMP*

✓ *a. No*

7.a.2 Do RI members carry data with them?

Does anyone carry project data on laptops, USB sticks or other external media?

Tags: *Science Europe DMP*

✓ *a. No*

7.a.3 Do RI members store project data in cloud accounts?

Think about services like Dropbox, but also about Google Drive, Apple iCloud accounts, or Microsoft's Office365

✓ *b. Yes*

Make sure your users are aware of the risks of cloud storage (not so much that the cloud is unreliable, but there is no protection against "accidentally" sharing a cloud folder with people outside the project)

7.a.4 Do RI members send project data or reports per e-mail or other messaging services?

✓ *b. Yes*

7.a.5 Do all data centers where RI data is stored carry sufficient certifications?

Tags: *Science Europe DMP*

✓ *b. Yes*

7.a.6 Are all RI web services addressed via secure http (https://)?

Tags: *Science Europe DMP*

✓ *b. Yes*

7.a.7 Have RI members been instructed about the risks (generic and specific to the project)?

RI members may need to know about passwords (not sharing accounts, using different passwords for each service, and two factor authentication), about security for data they carry (encryption, backups), data stored in their own labs and in personal cloud accounts, and about the use of open WiFi and https

Tags: *Science Europe DMP*

✓ *a. No*

7.a.8 Did you consider the possible impact to the RI or organization if information is lost?

Tags: *Science Europe DMP*

✓ *b. Yes; the effect is small*

7.a.9 Did you consider the possible impact to the RI or organization if information leaks?

Tags: *Science Europe DMP*

✓ *b. Yes; the effect is small*

7.a.10 Did you consider the possible impact to the RI or organization if information is vandalized?

Tags: *Science Europe DMP*

✓ *b. Yes; the effect is small*

7.a.11 Are personal data sufficiently protected?

Tags: *Science Europe DMP*

✓ *a. We are not using any personal information*

8 Do you have a contingency plan?

What will you do if the compute facility is down?

✓ *b. We have an alternative*

9 Will you version datasets?

[SEEK](#) which is used in [FAIRDOMHub](#) and can be used together with [NeLS](#) supports versioning by default.

[NeLS](#) can also be used with [Git Large File Storage \(LFS\)](#)

External Links: [FAIRDOMHub](#), [SEEK](#), [NeLS](#), [Git Large File Storage \(LFS\)](#)

✓ *a. Yes*

VI. Interpreting data

The interpretation of the data consists of the last steps of processing (often with manual interventions), visualisation, and data integration. In this chapter many questions about data interoperability will come up.

Report

Indications

Answered (current phase)	24 / 24
Answered	24 / 24

Metrics

Metric	Score
Interoperability	1
Reusability	0.5

Questions

1 How will you be doing the integration of different data sources?

✓ a. Explore

1.a.1 List the data formats you will be using for data integration

Answer some questions for each

Answers

1.a.1.b.1 Data format:

✓ None

1.a.1.b.2 How is the data structured in general?

✓ b. A table or set of tables (consisting of 'data records')

1.a.1.b.2.b.1 Does each column have a header?

In a table, the data items are arranged in columns. Is there a header for each of these describing what is in there?

✓ b. Yes

1.a.1.b.2.b.1.b.1 Are all column headers unambiguous?

A human being quickly 'understands' data items and their relations. For good data reusability, it is necessary that computers can understand your data too.

✓ b. Yes

1.a.1.b.2.b.1.b.2 Do all columns/headers have a data type?

A label like 'temperature' only makes sense to a computer if it is also clear what the units are and what temperature has been measured. In many cases, it is also important how it was measured.

✓ b. Yes

1.a.1.b.2.b.1.b.3 Are the limitations to allowed data values in each column explicit?

If there are reasonable limitations to the values in a column, or even a limited set of allowed values, it is very good for data validation and reusability if these limitations are explicit, and e.g. software used for data entry and editing will not allow anything else.

🔗 External Links: [Rightfield: Template fields in Microsoft Excel](#)

✓ b. Yes

1.a.1.b.2.b.2 Is it clear what a row in the table represents?

✓ b. Yes

1.a.1.b.2.b.3 Does each row have an identifier?

✓ b. Yes

1.a.1.b.2.b.4 Is there a distinguishing way a missing value in the table can be recognized?

Sometimes, an empty field or a zero is indicating a missing value. But is that really unique? Could there be valid empty or zero fields? Has the convention for missing values been made explicit somewhere?

✓ *b. Yes*

1.a.1.b.2.b.5 Is the relation between each of the columns and the record identifier clear?

It may appear that in a table with 'patients' as rows, a column labeled 'disease' coupled to an ontology has a clear meaning. But that is not always explicit enough! A 'disease' could e.g. be the disease that the patient is suffering from, but it could also be an earlier diagnose, a suspected diagnose, or the disease a family member recently died of.

✓ *b. Yes*

1.a.1.b.2.b.6 Are all the relations between the column headers explicit?

For a good understanding of tabular data, you need to make the relationship between each pair of columns explicit. E.g. if one column is 'disease' and another is 'treatment', you want to make sure that this is the chosen treatment that this person is undergoing for the given disease.

✓ *b. Yes*

1.a.2 Will you/your users be using a workflow for data integration, e.g. with tools for database access or conversion?

☰ Data Stewardship for Open Science: [qqb](#)

✓ *a. No*

1.a.3 Will you/your users use a 'linked data' approach?

🔗 External Links: [Linked data \(wikipedia\)](#)

✓ *b. Yes*

1.a.3.b.1 Are your data sources using linked data?

✓ *b. Partly*

1.a.3.b.2 Will you provide your results as semantically interoperable linked data?

☰ Data Stewardship for Open Science: [fxm](#)

✓ *b. Yes*

2 Will you/your users be using common or exchangeable units?

✓ *b. Yes*

3 Will you/your users be using common ontologies?

✓ *b. Yes*

Choose the ontologies before you start

4 Will there be potential issues with statistical normalization?

✓ *a. No*

5 Will you/your users be integrating different data sources to get more samples or more data points?

✓ a. No

6 Will you/your users be integrating different data sources in order to get more information for each sample or data point?

✓ a. No

7 Do you/your users have all tools to couple the necessary data types?

✓ b. Yes

8 Will you/your users be doing (automated) knowledge discovery?

☰ Data Stewardship for Open Science: [bzu](#)

✓ a. No

VII. Preserving data

In this chapter, issues regarding data publication and long term archiving are addressed.

Report

Indications

Answered (current phase)	14 / 14
Answered	43 / 43

Metrics

Metric	Score
Findability	0.6
Accessibility	1
Reusability	0.62
Good DMP Practice	0.43

Questions

1 Will you /your users be archiving data (using so-called 'cold storage') for long term preservation already during the RI runtime/project?

Much of the raw data you have will need to be archived for your own later use somewhere. This is often done off-line on tape, not on the disks of the compute facility. Please note that this does not refer to the data publication.

☰ Data Stewardship for Open Science: [kjp](#)

✓ a. No

1.a.1 Can the original data be regenerated?

☰ Data Stewardship for Open Science: [ixr](#)

✓ a. No

1.a.2 **When is the raw data archived?**

- ✓ *c. All at once with the results at the end of the project*

2 **Specify details of data types which will be produced at your RI**

It is useful to think about a data types as some collection of data that will be ending up in the same place.

Tags: *maDMP, Science Europe DMP*

Answers

2.b.1 **Data type:**

Consider one data set as a collection of data from one set of samples.

Tags: *maDMP, Science Europe DMP*

- ✓ *Occurrence data*

2.b.2 **Description of the data type**

Examples could be "Field observations", "raw instrument data", "genomic variants".

Tags: *Science Europe DMP, maDMP*

- ✓ *biodiversity occurrence data*

2.b.3 **Identifier of the data type**

Please add all "formal" identifiers you have for this data set: these can be handles or DOIs or any other type. One important purpose of these identifiers is to be able to find the dataset back.

A good identifier is *persistent* (i.e. it does not change, and also the same identifier will never be used for another data set), *globally unique* (nobody else uses the same identifier for a different data set) and *resolvable* (you can actually locate the data set if you only know the identifier).

Tags: *Science Europe DMP, maDMP*

Answers

2.b.3.b.1 **What type of identifier?**

Which type of identifier is this?

- ✓ *b. DOI*

2.b.3.b.2 **The actual identifier**

- ✓ *<https://doi.org/10.5281/zenodo.3248925>*

2.b.4 **Will this data types be published?**

Will you publish the data set somewhere? Note that this does not necessarily mean that the data set becomes openly available, conditions for access and use may apply.

Tags: *maDMP, Science Europe DMP*

- ✓ *b. Yes*

2.b.4.b.1 **Specify where you will distribute this data from**

Add each of the locations where this data set will be made available. Give each of these a short name that identifies it to yourself.

Tags: *Science Europe DMP, maDMP*

Answers

2.b.4.b.1.b.1 What repository will this data be stored in?

Domain repositories often have the best functionality to make the data findable and reusable. Many of them are listed in <https://fairsharing.org/>

If a repository offers to give your data set a DOI it is a good idea to use that option.

Tags: *Science Europe DMP*

External Links: [FAIRSharing](#), [Registry of Research data Repositories](#), [ELIXIR deposition repositories](#), [Biolmage Archive](#)

✓ a. *A domain-specific repository*

2.b.4.b.1.b.1.a.1 What repository?

Tags: *Science Europe DMP*

✓ *Global Biodiversity Information Facility*



<https://fairsharing.org/bsg-d000635>

2.b.4.b.1.b.1.a.2 Will you contact the repository beforehand?

Contacting the repository early may be useful to establish conditions, formats, and metadata requirements for submission. It may also be necessary to establish whether the repository can accommodate your data

Tags: *Science Europe DMP*

✓ a. *No, this submission is routine both for us and the repository*

2.b.4.b.1.b.2 Who will the data in this place be shared with?

Tags: *maDMP, Science Europe DMP*

✓ a. *Open: The data will be shared with anyone, as long as they obey conditions of the license*

2.b.4.b.1.b.3 Licenses under which this distribution of the data set will be available

Please add each license that this data set will be available as. For each license you will be able to specify when it starts being applicable, so that you can e.g. specify that the data is restricted for a few months and open afterwards.

Tags: *Science Europe DMP, maDMP*

Answers

2.b.4.b.1.b.3.b.1 Under what license will the data set be made available?

Tags: *maDMP, Science Europe DMP*

External Links: [Guide from the Digital Curation Centre \(UK\) on licenses for research data](#), [Guide from the Open Data Institute on licenses for research data](#)

✓ a. *They will be freely available for any use (public domain or CC0)*

2.b.4.b.1.b.3.b.2 Starting date

From which date will data be available under this license?

Tags: *Science Europe DMP, maDMP*

✓ *Date of publication*

2.b.4.b.2 Will you be adding a reference to the published data to at least one data catalogue?

Data is sometimes difficult to locate, especially if it is not in a domain-specific repository. Data catalogues may increase findability.

Tags: *Science Europe DMP*

✓ *a. No*

2.b.5 How long will this data set be kept?

For optimum reusability data needs to be available for as long as possible. There may be financial reasons why you can't keep the data any longer; there may be legal reasons requiring you to delete the data.

Tags: *Science Europe DMP*

✓ *a. As long as technically possible*

2.b.6 Will the metadata be available even when the data no longer exists?

This is a one of the FAIR principles.

Tags: *Science Europe DMP*

✓ *b. Yes*

2.b.7 Does the data usually contain personal data?

Is there anything in this dataset that could be tied to a person? This could be a physical characteristic, but also behavior of a person, movements, communications. Note that e.g. readouts about the performance of an airplane are considered to contain personal data of the pilot!

Tags: *Science Europe DMP, maDMP*

✓ *a. No*

2.b.8 Does this data contain sensitive information?

Personal information can be sensitive if it is for instance about the health, sexual orientation, religion of a person. But there are also other classes of sensitive information: e.g. locations of rare species in biodiversity could be sensitive and should not leak to poachers.

Tags: *Science Europe DMP, maDMP*

✓ *a. No*

2.b.9 Do you make use of persistent and unique identifiers such as Repository specific Identifiers or Digital Object Identifiers for this ?

✓ *a. Yes*

3 Will any of the repositories you use charge you/your users for their services?

Tags: *Science Europe DMP*

✓ *a. No*

4 Did you budget for the time and effort it will take to help user to prepare the data for publication?

Tags: *Science Europe DMP*

✓ *b. Yes*

5 Will you be making sure that blocks of data deposited by you or by the users in different repositories can be recognized as belonging to the same study?

✓ *b. Yes, all data sets will have links to the related data*

6 Are there any recurring fees to keep data or documents available?

Are you using any commercially licensed products to keep data, software or documents available, for which a regular fee must be paid?

✓ *a. No*

7 Will you be archiving your data after the RI runtime in 'cold storage'?

Will you be storing (in cold storage) copies of your own data for a longer period after the project has ended? Possibly as a continuation of archival as part of data storage strategy during the project? Data archival is distinct from data publishing, an archive is usually limited in who can access the data.

Data Stewardship for Open Science: [fxe](#)

✓ *a. No*

8 Will you also publish data if the results of your study are negative/inconclusive or unpublishable?

Even if you do not obtain the results you had foreseen from your own study, the data can still be valuable for reuse in another context. Also, publishing the data can avoid that someone else collects a similar data set with a similar negative result.

✓ *b. Yes*

9 Specify a list of software packages you will be publishing

Specify a short name for each software package.

Answers

9.b.1 Software package:

✓ *None*

9.b.2 Will you be adding a proper open-source license?

External Links: [Choose an open source license](#), [Open Source Initiative: Licenses](#)

✓ *a. No*

9.b.3 Where will the software package be available?

✓ *Nowhere, it does not exist*

9.b.4 Will this software be listed in a catalogue?

✓ *a. No*

10 How will you be making sure there is good provenance of the data (and analysis)?

Data analysis is normally done manually on a step-by-step basis. It is essential to make sure all steps are properly documented, otherwise results will not be reproducible.

Tags: *Science Europe DMP*

✓ *c. We use other arrangements*

10.c.1 What other arrangements?

Tags: *Science Europe DMP*

✓ *No data analysis is done, this is just a project to publish data*

11 Will reference data be created?

Will any of the data that you will be creating form a reference data set for future research (by others)?

Data Stewardship for Open Science: [rbz](#)

✓ *a. No*

12 How will you document your/the user data?

For reusability, the data should be well documented. In this section of the questionnaire you can specify what kinds of documentation you will be providing.

Tags: *Science Europe DMP*

✓ *a. Explore*

12.a.1 Will you be documenting the data with Dublin Core metadata?

Dublin Core is a standard documenting domain independent aspects of a resource; including who has created it, audience, function, formatting and licensing. Does your documentation follow the Dublin Core standard?

Tags: *Science Europe DMP*

External Links: [Dublin Core Metadata Terms](#), [Dublin Core Initiative](#)

✓ *b. Yes*

12.a.2 Will you be documenting the data with W3C PROV provenance?

The W3C Prov standard documents processes (workflow) that were used to produce a resource. This can be used to document e.g. the software (including version) and parameters you use to analyze the data. Will your documentation follow the W3C Prov standard?

Tags: *Science Europe DMP*

External Links: [W3C Prov primer](#)

✓ *a. No*

13 Will you do systems biology modeling (for users)?

✓ a. No

14 Will you do structural modeling?

✓ a. No

VIII. Giving access to data

This chapter deals with the information needed by people who will re-use your data, and with the access conditions they will need to follow.

Report

Indications

Answered (current phase)	4 / 4
Answered	4 / 4

Metrics

Metric	Score
Openness	1

Questions

1 Will you be working with the philosophy 'as open as possible' for your data/your users data?

Tags: *Science Europe DMP*

Data Stewardship for Open Science: [jvm](#)

✓ b. Yes

2 Are there potential copyright and Intellectual Property Rights (IPR) issues?

✓ b. No

3 Can all of your data at your RI become completely open immediately?

Tags: *maDMP, Science Europe DMP*

✓ b. Yes

4 Will there be valorization or translational returns of the data generated at your RI?

✓ a. No