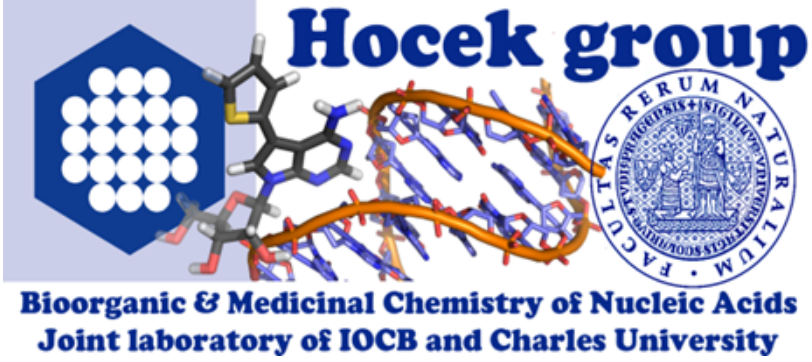

HOCEK GROUP
Data Management Plan

Organization	DSW (researchers)
Created by	Michal Tichy (michal.tichy@uochb.cas.cz)
Based on	Common DSW Knowledge Model, 2.4.4 (dsw:root:2.4.4)
Project Phase	Before Submitting the Proposal
Created at	15 Mar 2024

Summary Report

Summary

Answered (current phase)	60 / 62
Answered	164 / 193

Metric	Score
Findability	1.00
Accessibility	0.82
Interoperability	0.82
Reusability	0.88
Good DMP Practice	0.83
Openness	0.83

I. Administrative information

Summary

Answered (current phase)	0 / 1	
Answered	8 / 15	

Questions

1

Contributors

Horizon 2020 DMP maDMP Science Europe DMP Horizon Europe DMP

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.

Answers

1.a.1

Name

Horizon 2020 DMP maDMP Science Europe DMP Horizon Europe DMP

✔ Michal Hocek

1.a.2

E-mail address

Horizon 2020 DMP maDMP Science Europe DMP Horizon Europe DMP

✔ michal.hocek@uochb.cas.cz

1.a.3

ORCID Identifier

Horizon 2020 DMP | maDMP | Science Europe DMP | Horizon Europe DMP

✔ 0000-0002-1113-2047

1.a.4

Affiliation

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔

Czech Academy of Sciences, Institute of Organic Chemistry and Biochemistry

1.a.5

Role

Horizon 2020 DMP | maDMP | Science Europe DMP | Horizon Europe DMP

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".

- [ ] Contact Person
- [ ] Data Collector
- [ ] Data Curator
- [ ] Data Manager
- [ ] Data Protection Officer
- [ ] Data Steward
- [ ] Distributor
- [ ] Editor
- [ ] Producer
- [x] Project Leader
- [ ] Project Manager
- [ ] Project Member
- [ ] Researcher
- [ ] Rights Holder
- [ ] Sponsor
- [ ] Supervisor

- [ ] Work Package Leader
- [ ] Other

1.b.1

Name

Horizon 2020 DMP maDMP Science Europe DMP Horizon Europe DMP

✖ *This question has not been answered yet!*

1.b.2

E-mail address

Horizon 2020 DMP maDMP Science Europe DMP Horizon Europe DMP

✖ *This question has not been answered yet!*

1.b.3

ORCID Identifier

Horizon 2020 DMP maDMP Science Europe DMP Horizon Europe DMP

✖ *This question has not been answered yet!*

1.b.4

Affiliation

Horizon 2020 DMP Science Europe DMP Horizon Europe DMP

✖ *This question has not been answered yet!*

1.b.5

Role

Horizon 2020 DMP | maDMP | Science Europe DMP | Horizon Europe DMP

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".

✖ *This question has not been answered yet!*

2

Research Project(s)

Horizon 2020 DMP | maDMP | Science Europe DMP | Horizon Europe DMP

Add each of the research project(s) that you are (or 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.

✖ *This question has not been answered yet!*

3

To execute the DMP, is additional specialist expertise required?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ a. No

4

Do you require hardware or software in addition to what is usually available in the institute?

Horizon 2020 DMP | Horizon Europe DMP

✔ a. No

5

Describe national / funder / sectorial / departmental policies and procedures for data management that you will be using.

Horizon Europe DMP

✖ *This question has not been answered yet!*

II. Re-using data

Before you decide to embark on any new study, it is 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.

Summary

Answered (current phase)	5 / 6	
Answered	13 / 17	

Questions

1

Do you need guidance to find existing data?

Research funding organisations more and more demand that you search for existing data sets that could have information that you need, before assuming you need to collect all data yourself. You are asked to list what you have been able to locate, and whether you have found it suitable to use in your research. Do you need pointers to find such existing data sets?

✔ a. No

2

Is there any pre-existing data?

Horizon Europe DMP | Horizon 2020 DMP | maDMP | Science Europe DMP

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

Data Stewardship for Open Science: *atq*

External Links: *Google dataset search*, *Datacite Search*, *RDMkit on Reusing Data*, *RDMkit on Existing Data*

✔ b. Yes

2.b.1

What existing reference data did you consider re-using?

Horizon Europe DMP | Horizon 2020 DMP | Science Europe DMP

Much of todays data is used in comparison with reference data. You may be comparing your own data with a "standard set" which is maintained as a collection by someone else. Or you could be determining differences to a standard (for example in bioinformatics, a genome is often compared with a reference genome to identify genomic variants). If you use reference data, there are several specific issues that you should consider. What are the reference data sets that you will use?

Data Stewardship for Open Science: *quc*

Answers

2.b.1.a.1

Reference database or dataset

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

Give the name of the database or dataset. You will be shown suggestions of data bases from FAIRSharing, but you can also type the name of a dataset that is not in FAIRsharing

✔

The Protein Database

2.b.1.a.2

Where is this reference data available

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

Specify a URL or a persistent identifier (e.g. DOI) for the database or dataset. If possible, refer exactly to the version that you are using.

✔ http://www.ncbi.nlm.nih.gov/protein

2.b.1.a.3

Will you be using this reference data set?

Horizon Europe DMP | Horizon 2020 DMP | Science Europe DMP

✔ b. Yes

2.b.1.a.3.b.1

Name and contact details of the owner of this data

Horizon Europe DMP

✖ *This question has not been answered yet!*

2.b.1.a.3.b.2

What are the conditions of use for this database or dataset?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

Although there is no world-wide rule for the application of copyright on data sets (copyright only applies to things that require a so-called "creative step"), it is wise to check for an explicit permission to use a data set and not to assume that data can be used freely just because you can access it. Note that copyright laws explicitly *forbid* the use of a copyrighted work, *except* if you get permission. Such a permission is called a "licence". So: if you can not find a licence, you have to assume you can not use the data.

External Links: *Wikipedia on Copyright*

✖ *This question has not been answered yet!*

2.b.1.a.3.b.3

Do you know in what format the reference data is available?

Horizon Europe DMP

Do you know the data format of the reference data? Is this suitable for your work? Does it need to be converted?

Data Stewardship for Open Science: *jxb*

✔ a. I can directly use it

2.b.1.a.3.b.4

Is the reference database or dataset versioned?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

Many reference datasets and databases evolve over time. If the reference data changes, this may affect your results. If different versions of a reference data exist, you need to establish your "version policy".

Data Stewardship for Open Science: *rgy*

✔ b. Yes

2.b.1.a.3.b.4.b.1

Which version will you use?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

If there are different versions available, you have to decide with all project partners together which version you will be using. Probably you will go for the latest release as of the date of the start of your research project. However, if you have other data from older projects that need to be merged, you may need to consider using the same release you used for a previous project.

✖ *This question has not been answered yet!*

2.b.1.a.3.b.4.b.2

Will you change version if it updates?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

If the reference changes while you are working on your research project, you need to decide whether you will follow these changes. Most likely that will mean that you have to do some analyses again, so you will need to make sure enough resources are available to do so. You can decide to stay with the version that you started with; this can have the disadvantage that you will not benefit from added information or added consistency.

✔ b. New analyses will be done with the new version

2.b.1.a.3.b.5

How will you make sure the same reference data will be available to reproduce your results?

Horizon Europe DMP

Will the reference data in the version you use be available to others?

✔ a. The provider keeps old versions around

2.b.1.a.3.b.6

What will you use this reference data set for?

Horizon Europe DMP

✔ Obtaining information about proteins.

2.b.2

What existing non-reference data sets did you consider re-using?

Horizon Europe DMP Horizon 2020 DMP maDMP Science Europe DMP

Even if you will be producing your own data, you often will also be relying on existing data sets (e.g. from your own earlier projects). You may need to integrate your new data with an existing data set or retrieve additional information from related data bases. Will you be doing such things?

Data Stewardship for Open Science: *wya*

✘ *This question has not been answered yet!*

2.b.3

Do you need to harmonize different sources of existing data?

Horizon Europe DMP

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.4

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

Horizon Europe DMP

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

In this chapter we describe all the sources of data: they can e.g. come from instruments or from questionnaires; data can be newly collected as part of the current project, but it can also be pre-existing data that may need proper contracts with the maintainer, some pre-processing, and quality checks. It can also be reference data that is part of curated resources and (public) databases.

For more information see Collecting in RDMKit

Summary

Answered (current phase)	24 / 24
Answered	63 / 69

Metric	Score
Findability	1.00
Accessibility	0.50
Interoperability	0.90
Good DMP Practice	1.00

Questions

1

Are you running the project in a collaboration between different groups or institutes?

✔ b. Yes

1.b.1

Is there a collaboration agreement in the project that describes who can have access to what data?

External Links: *See also description of Data Sharing in RDMkit*

✔ b. Yes

2

Will you be collecting physical samples?

Will you be collecting artefacts like specimens, minerals, biological samples?

Data Stewardship for Open Science: *kuz*

✔ a. No

3

How will you do file naming and file organization?

Horizon 2020 DMP Horizon Europe DMP

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

External Links: *RDMkit on data organisation*

✔ a. Explore

3.a.1

Did you make appointments in the project on how to name files and folders?

Horizon 2020 DMP Horizon Europe DMP

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

✔ b. Yes

3.a.1.b.1

How do you name files in the project?

Horizon 2020 DMP Horizon Europe DMP

Describe how everyone in the project will be naming files and folders, and what folder structure you will use.

✘ *This question has not been answered yet!*

3.a.2

Will you be keeping the relationships between data clear in the file names?

Horizon 2020 DMP Horizon Europe DMP

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

✔ b. Yes

3.a.3

Will all the metadata that is embedded in the file names also be available in the proper metadata?

Horizon 2020 DMP Horizon Europe DMP

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.

Also note that if metadata could need to change, embedding it in the file names may require renaming files during the project; and this may have implications for references to those files.

✔ b. Yes, all metadata is also explicitly available elsewhere

3.a.4

Will you be using persistent identifiers to refer to data within the project?

Horizon 2020 DMP

Especially for large projects, referring to data internally via a persistent identifier system can be helpful as such a system can help to keep track of data that moves to a new location.

External Links: *The Handle System*, *Handbook on Persistent Identifiers*

✔ b. Yes

4

Do you need guidance on what data formats/types to use?

In many research fields there are standard file formats and types that have been established by disciplinary standards organisations. Using such standards can help increase the interoperability of your data with other data and tools. Do you need help finding the right standards to use in your research?

✔ a. No

5

What existing data formats/types will you be using?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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 "JPG" and "CSV" [comma separated values]) that help to make these data reusable. Are you using such common data formats?

Please make sure you list all the data types that are important for your project. You should make sure also to list the formats used in any data sets that you are re-using.

Data Stewardship for Open Science: *njy*

Answers

5.a.1

Data format/type

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ graphic data (tiff, PDF/A, PNG, GIF, BMP, SVG

5.a.2

Is this a standard data format widely used by researchers in this field?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ b. Yes

5.a.3

Does this data format enable sharing and long term archiving?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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

Ideally a format should be simple, text only, completely described, not restricted by copyrights, and implemented in different software packages.

✔ b. Yes

5.a.4

What volume of data of this type will you be working with?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ a. So small that it is not a problem

5.a.5

Is this data format completely described?

Formats like XLS or SQL are very flexible; they can be adapted to many different uses, and this makes them good for interoperability. However, their flexibility also makes that it is not immediately obvious from the file structure how it can be used. The data needs a proper *description* in order for others (or yourself at a later time) to be able to unambiguously understand what it contains.

✔ b. Yes

5.b.1

Data format/type

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ Tables/ CSV, TXT, XLS

5.b.2

Is this a standard data format widely used by researchers in this field?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ b. Yes

5.b.3

Does this data format enable sharing and long term archiving?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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

Ideally a format should be simple, text only, completely described, not restricted by copyrights, and implemented in different software packages.

✔ b. Yes

5.b.4

What volume of data of this type will you be working with?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ a. So small that it is not a problem

5.b.5

Is this data format completely described?

Formats like XLS or SQL are very flexible; they can be adapted to many different uses, and this makes them good for interoperability. However, their flexibility also makes that it is not immediately obvious from the file structure how it can be used. The data needs a proper *description* in order for others (or yourself at a later time) to be able to unambiguously understand what it contains.

✔ b. Yes

5.c.1

Data format/type

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ Text/XML, PDF, DOC, HTML

5.c.2

Is this a standard data format widely used by researchers in this field?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ b. Yes

5.c.3

Does this data format enable sharing and long term archiving?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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

Ideally a format should be simple, text only, completely described, not restricted by copyrights, and implemented in different software packages.

✔ b. Yes

5.c.4

What volume of data of this type will you be working with?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ a. So small that it is not a problem

5.c.5

Is this data format completely described?

Formats like XLS or SQL are very flexible; they can be adapted to many different uses, and this makes them good for interoperability. However, their flexibility also makes that it is not immediately obvious from the file structure how it can be used. The data needs a proper *description* in order for others (or yourself at a later time) to be able to unambiguously understand what it contains.

✔ b. Yes

5.d.1

Data format/type

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔

Nuclear Magnetic Resonance Extracted Data Format

5.d.2

Is this a standard data format widely used by researchers in this field?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ b. Yes

5.d.3

Does this data format enable sharing and long term archiving?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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

Ideally a format should be simple, text only, completely described, not restricted by copyrights, and implemented in different software packages.

✔ b. Yes

5.d.4

What volume of data of this type will you be working with?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ a. So small that it is not a problem

5.d.5

Is this data format completely described?

Formats like XLS or SQL are very flexible; they can be adapted to many different uses, and this makes them good for interoperability. However, their flexibility also makes that it is not immediately obvious from the file structure how it can be used. The data needs a proper *description* in order for others (or yourself at a later time) to be able to unambiguously understand what it contains.

✔ b. Yes

6

Do you need guidance on what encodings/terminologies/vocabularies/ontologies to use?

✔ a. No

7

What existing encodings/terminologies/vocabularies/ontologies will you be using?

Horizon 2020 DMP | Horizon Europe DMP

✖ *This question has not been answered yet!*

8

Will you be using new types of data?

Horizon 2020 DMP | Horizon Europe DMP

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

9

How will you be collecting and keeping your metadata?

Science Europe DMP | Horizon 2020 DMP | Horizon Europe DMP

For the re-usability of your data by yourself or others at a later stage, a lot of information about the data – for example 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.

Data Stewardship for Open Science: *rhm*
External Links: *RDMkit on documentation and metadata*, *Metadata Standards Catalogue (RDA)*

✘ *This question has not been answered yet!*

10

Will you be acquiring data using measurement equipment?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ b. Yes

10.b.1

Specify what data sets you will acquire using measurement equipment

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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

Answers

10.b.1.a.1

Data set:

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ NMR

10.b.1.a.2

Who will do the measurements? And where?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

Are there easily accessible specialized service providers for data capture?

✔ a. Experts in the project, with our own equipment

10.b.1.a.3

Instruments used for data collection

Science Europe DMP

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

Answers

10.b.1.a.3.a.1

Instrument name

Science Europe DMP

✔ NMR spectrometer

10.b.1.a.3.a.2

Instrument description

Science Europe DMP

✔ 500 MHz

10.b.1.a.4

Is the equipment completely standard and well described?

Horizon 2020 DMP Science Europe DMP Horizon Europe DMP

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?

✔ a. Very well described and known

10.b.1.a.5

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

10.b.1.a.6

Will you be using quality processes?

Science Europe DMP | Horizon Europe DMP

Specify how you guarantee that measurements measure what you think they should measure, and how you establish completeness and correctness of data and metadata.

External Links: *RDMkit on Data Quality*

✔ a. No

10.b.1.a.7

Who else could be interested in using this data?

Horizon 2020 DMP | Horizon Europe DMP

✔ a. Just us

10.b.1.b.1

Data set:

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ MS data

10.b.1.b.2

Who will do the measurements? And where?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

Are there easily accessible specialized service providers for data capture?

✔ a. Experts in the project, with our own equipment

10.b.1.b.3

Instruments used for data collection

Science Europe DMP

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

✗ *This question has not been answered yet!*

10.b.1.b.4

Is the equipment completely standard and well described?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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?

✓ a. Very well described and known

10.b.1.b.5

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

10.b.1.b.6

Will you be using quality processes?

Science Europe DMP | Horizon Europe DMP

Specify how you guarantee that measurements measure what you think they should measure, and how you establish completeness and correctness of data and metadata.

External Links: *RDMkit on Data Quality*

✔ b. Yes

10.b.1.b.6.b.1

Are you calibrating measurements?

Science Europe DMP Horizon Europe DMP

✔ b. Yes

10.b.1.b.6.b.2

Are you running repeat samples or are you repeating measurements in order to monitor consistency of the results?

Science Europe DMP Horizon Europe DMP

✔ b. Yes

10.b.1.b.6.b.3

Are you running standardized data capture or recording?

Science Europe DMP Horizon Europe DMP

✘ *This question has not been answered yet!*

10.b.1.b.6.b.4

Are you doing Data Entry validation?

Science Europe DMP Horizon Europe DMP

Are you checking that data that is entered e.g. in forms is satisfying criteria of reasonability?

✔ a. No

10.b.1.b.6.b.5

Are you using data peer review?

Science Europe DMP Horizon Europe DMP

Do you make sure all collected is cross-checked by colleagues?

✔ a. No

10.b.1.b.6.b.6

Are you using controlled vocabularies where possible?

Science Europe DMP Horizon Europe DMP

Controlled vocabularies should be used to limit what can be entered in most text fields.

✖ *This question has not been answered yet!*

10.b.1.b.6.b.7

Are you using any other quality processes?

Science Europe DMP Horizon Europe DMP

✔ a. No

10.b.1.b.6.b.8

Are you measuring samples for which the outcome is known in order to monitor consistency?

Horizon Europe DMP

✔ b. Yes

10.b.1.b.7

Who else could be interested in using this data?

Horizon 2020 DMP Horizon Europe DMP

✔ a. Just us

11

Do you have any non-equipment data capture?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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

Data Stewardship for Open Science: *ybw*

✔ a. No

12

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

✔ a. No

12.a.1

Can all data be brought into the same format, e.g. RDF?

✔ a. No

13

Will you collect any data connected to a person, "personal data"?

Horizon Europe DMP

Very many kinds of data are connected to people. If there could be a way for someone, including yourself, to find out who that person is, that is considered personal data.

Simple examples are name, birth day or address; but there are many other data that can be personal: for example a voice recording, a combination of a location and a time (traffic flow), genetic information, or an X-ray of the skull.

External Links: *RDMkit about Sensitive Data*, *RDMkit on Human Data*, *RDMkit on Data Protection*

✔ a. No

14

How is the ownership of the collected data arranged?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ c. All data will be owned by the institute

15

Will you monitor data integrity once it has been collected?

Working with large amounts of heterogenous data in a larger research group can have 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*
External Links: *RDMkit on Data Quality*

✔ b. Yes

15.b.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, because it will help detect early when data files are damaged or mixed up. It can also be helpful to keep the list under version control so that all changes are well documented.

✔ b. Yes

15.b.2

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

✔ a. No

IV. Processing data

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

In this chapter, many questions are focusing on the compute environment that is used to process the data and make it available for interpretation by project partners. Some of those questions (e.g. on workflow systems and data provenance) are also relevant for the work in the interpretation phase.

Summary

Answered (current phase)	11 / 11
Answered	33 / 39

Metric	Score
Accessibility	0.50
Reusability	0.80
Good DMP Practice	0.85

Questions

1

Will you be using a shared working space to work with your data?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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

✔ a. No, participants in the project each have different collections of data and tools

1.a.1

Are data that project members and partners store themselves adequately backed up and traceable?

Science Europe DMP

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

2

Data storage systems and file naming conventions

Science Europe DMP

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.

External Links: *RDMkit on data organisation*, *RDMkit on data storage*

✔ a. Explore

2.a.1

How much storage space will the project require for all data and software, including temporary storage?

✔ a. So little that it is not a problem

2.a.2

Are you using a filesystem with files and folders?

Science Europe DMP

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

External Links: *RDMkit on Data Organisation*

✔ b. Yes

2.a.2.b.1

Will you use a folder for each sample/subject?

Science Europe DMP

✔ b. Yes

2.a.2.b.1.b.1

What is the naming convention for this folder?

What appointment have you made for the naming of the folders? Make sure names are relatively short, and avoid spaces and special characters.

✖ *This question has not been answered yet!*

2.a.2.b.2

Will you use a (sub)folder for each (repeated) analysis?

Science Europe DMP

✔ b. Yes

2.a.2.b.2.b.1

What are the naming conventions for the analysis folders?

What appointment have you made for the naming of the folders? Make sure names are relatively short, and avoid spaces and special characters.

✖ *This question has not been answered yet!*

2.a.2.b.3

Will you use a (sub)folder for each step in the analysis workflow?

Science Europe DMP

✔ b. Yes

2.a.2.b.3.b.1

What are the naming conventions for these folders?

What appointment have you made for the naming of the folders? Make sure names are relatively short, and avoid spaces and special characters.

✖ *This question has not been answered yet!*

2.a.2.b.4

What appointments have you made about the naming of files?

Science Europe DMP

Make sure names are relatively short, and avoid spaces and special characters. You can use underscore characters, and consider using unique identifiers for the samples/experiments. You can consider to add versioning using the date in YYYYMMDD format.

✖ *This question has not been answered yet!*

2.a.2.b.5

Did you document how you manage file versioning?

✔ b. Yes

2.a.3

Will you be storing data in an "object store" or a "document store" system?

Science Europe DMP

Some "file" storage systems do not have a tree structure like we know in a file system, but rather have direct pointers to any file in the system. Such systems are called "object stores" or "document stores". Examples are Amazon S3 and CEPH, or MongoDB.

External Links: *Wikipedia on object storage*, *RDMkit on Data Storage*

✔ a. No

2.a.4

Will you use a database system to store project data?

Science Europe DMP

✔ a. No

2.a.5

Are you storing (some of your) data in an application specific manner?

Are you using e.g. an Electronic Lab Notebook (ELN) or Electronic Data Capture (EDC) application? Such applications often have their own data structure that can only be accessed through the application.

✔ b. Yes

Note that using the data that is stored in such a system will need to be exported in some standard way before it becomes available for other data analysis tools.

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?

External Links: *RDMkit on Data Analysis*

✔ 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.

✘ *This question has not been answered yet!*

4.a.3

Will you use a central repository for all tools and their versions as used in your 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: *pzq*
External Links: *RDMkit on Package management systems*

✔ b. Yes

4.a.4

Will you use a central repository for reference data used in your 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 reference data versions.

Data Stewardship for Open Science: *pzq*

✔ b. Yes

4.a.5

Will you make use of standard workflow engines and automatic work flows for all data analysis in the project?

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.

✔ a. No

4.a.6

Are all software tools in the work flow 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?

Horizon 2020 DMP Horizon Europe DMP

✖ *This question has not been answered yet!*

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 abailable, 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?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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.

✔ a. Explore

7.a.1

Do project members store data or software on computers in the lab or external hard drives connected to those computers?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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

✔ b. Yes

7.a.2

Do project members carry data with them?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

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

✔ a. No

7.a.3

Do project members store project data in cloud accounts?

Think about services like Dropbox, but also about Google Drive, Apple iCloud accounts, or Microsoft Office 365.

✔ 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 project members send project data or reports per e-mail or other messaging services?

✔ b. Yes

7.a.5

Do all data centers where project data is stored carry sufficient certifications?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ b. Yes

7.a.6

Are all project web services addressed via secure http (https://)?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ b. Yes

7.a.7

Have project members been instructed about the risks (generic and specific to the project)?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

Project 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.

✔ b. Yes

7.a.8

Did you consider the possible impact to the project or organization if information is lost?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ c. Yes; the risk is acceptably low

7.a.9

Did you consider the possible impact to the project or organization if information leaks?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

Your institute may have a generic risk assessment; you should be informed about this.

✔ c. Yes; the risk is acceptably low

7.a.10

Did you consider the possible impact to the project or organization if information is vandalized?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ d. Yes; we will need to work on this.

7.a.11

Are personal data sufficiently protected?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

External Links: *RDMkit on Sensitive Data*, *RDMkit on Human Data*

✔ 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?

✔ a. We will wait until the problem is fixed

V. 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.

Summary

Answered (current phase)	7 / 7	
Answered	10 / 14	

Metric	Score
Interoperability	0.00

Questions

1

List the data formats you will be using for interpretation and describe their structure

Give each type of data a name that you recognise.

If you have data in many different structures, integrating the data may be more challenging.

External Links: *RDMkit on Machine Actionability*, *RDMkit on Data Processing*

Answers

1.a.1

Data type

✔ NMR

1.a.2

How is this data structured?

✔ a. A structured domain specific file with data and metadata fields

1.a.2.a.1

Can all of the data you want to couple be captured in that format?

✔ b. Yes

1.a.2.a.1.b.1

Will you be doing it that way?

✔ b. Yes

1.a.2.a.2

Does the domain specific format come with its own suite of integration tools that you will use?

✔ a. No

2

Will you be doing integration or linking of different data types?

If you are getting different types of data from different sources and want to use them together it is likely that you will need to match items and glue everything together. This can be done with traditional table database technology, but it is also possible to use Linked Data and RDF.

✔ a. No

3

Will you be using common ontologies?

✔ a. No

4

Will there be potential issues with statistical normalization?

✔ a. No

5

Will you be integrating different data sources to get more samples or more data points?

✔ a. No

6

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

✘ *This question has not been answered yet!*

7

Do you have all tools to couple the necessary data types?

✘ *This question has not been answered yet!*

8

Will you be using a federated analysis approach?

In some cases it is not practical to bring all data together:

- It may be legally hard to collect the data in one place for analysis
- It may be technically hard to transport data to a single place for analysis

In such cases, a Federated analysis approach may be applicable. Examples of such techniques are DataShield and the Personal Health Train. Secure multi-party computation may be useful too to prevent information leaking between parties.

✖ *This question has not been answered yet!*

9

Will you be doing (automated) knowledge discovery?

Data Stewardship for Open Science: *bzu*

✖ *This question has not been answered yet!*

VI. Preserving data

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

Summary

Answered (current phase)	10 / 10
Answered	31 / 33

Metric	Score
Findability	1.00
Accessibility	0.92
Reusability	1.00
Good DMP Practice	0.75

Questions

1

Specify a list of data sets you will be producing

Horizon 2020 DMP | maDMP | Science Europe DMP | Horizon Europe DMP

Add all the data sets you will be producing. Give each a short name, sufficient for yourself to know what data it is about. It is useful to think about a data set as some collection of data that will be ending up in the same place.

External Links: *RDMkit on Collecting Data*, *RDMkit on Data Preserving*

✖ *This question has not been answered yet!*

2

Will you be archiving data (using so-called 'cold storage') for long term preservation already during your project?

Horizon 2020 DMP | Horizon Europe DMP

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*

✔ b. Yes

2.b.1

Is the archived data changing over time, needing re-archival?

Data Stewardship for Open Science: *tgk*

✔ a. No

2.b.2

Will the archive be stored on disk or on tape?

✔ a. Disk

2.b.3

Will the archive be stored in a remote location, protecting the data against disasters?

Horizon 2020 DMP Horizon Europe DMP

✔ b. Yes

2.b.4

Will the archive need to be protected against loss or theft?

Horizon 2020 DMP Horizon Europe DMP

✔ b. Yes

2.b.4.b.1

Will the archive be encrypted?

Horizon 2020 DMP Horizon Europe DMP

✔ c. Yes

2.b.4.b.1.c.1

Is it clear who has access to the key? Also in case of a required data restore?

Horizon 2020 DMP Horizon Europe DMP

✔ b. Yes

2.b.4.b.2

Is it clear who has physical access to the archives?

Horizon 2020 DMP Horizon Europe DMP

✔ b. Yes

2.b.5

Will your project require the archives to be available on-line?

Data Stewardship for Open Science: *ybd*

✔ b. Yes

2.b.5.b.1

Will data integrity be guaranteed?

If the 'master copy' of the data is available on line, it should probably be protected against being tampered with.

✔ b. Yes

2.b.5.b.2

Is there an interface and a defined process for people to request access to the data?

✔ a. No

2.b.6

Has it been established who has access to the archive, and how fast?

✔ b. Yes

2.b.6.b.1

Has it been established who can ask for a restore during the project?

✔ b. Yes

2.b.6.b.2

If the data is voluminous, will the project be able to cope with the time needed for a restore?

✔ b. Yes

2.b.6.b.3

Has authority over the data been arranged for when the project is finished (potentially long after)?

✔ b. Yes

2.b.7

Has it been established how long the archived data need to be kept? For each of the different parts of the archive (raw data / results)?

Data Stewardship for Open Science: *kdp*

✔ b. Yes

2.b.8

Will the data still be understandable and reusable after a long time?

Horizon 2020 DMP Horizon Europe DMP

See also all questions about keeping metadata and data formats. Make sure the metadata is kept close to the data in the archive, and that community supported data formats are used for all long term archiving.

Data Stewardship for Open Science: *zmu*

✔ b. Yes

3

Will you be archiving your data in 'cold storage' after the project finishes?

Horizon 2020 DMP Horizon Europe DMP

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 strictly limited in who can access the data.

Data Stewardship for Open Science: *fxe*

✔ b. Yes

3.b.1

Who will be paying for the long term storage?

Horizon Europe DMP

✔ c. This is a service budgeted by one or more of the participating institutes

3.b.2

What is the minimum lifetime of the archive?

Horizon Europe DMP

✔ b. 10 years

3.b.3

Can the archival period be extended?

Horizon Europe DMP

✔ b. Yes

3.b.3.b.1

Who will decide whether the archival period is extended?

Horizon Europe DMP

✔ a. One of the principle investigators involved in the project will decide

3.b.3.b.2

What will the decision whether or not to extend the renewal be based on?

Horizon Europe DMP

- ☐ It will be based on the actual use of the archived data
- ☑ It will be based on the predicted use of the archived data
- ☐ It will be based on available budget

3.b.4

Will data formats of data in cold storage be upgraded if they become obsolete?

Horizon Europe DMP

✔ b. Yes

3.b.5

Will data be migrated regularly to more modern storage media (e.g. newer tapes)?

Horizon Europe DMP

✔ b. Yes

4

Will any of the repositories you use charge you for their services?

Horizon 2020 DMP Science Europe DMP Horizon Europe DMP

✔ b. Yes

4.b.1

How will you be paying for these services?

Horizon 2020 DMP Science Europe DMP Horizon Europe DMP

✔ c. These costs will be carried by (one of) the institutes involved in the project

5

Are there any other 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

6

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

Horizon 2020 DMP Science Europe DMP Horizon Europe DMP

✔ a. No

7

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

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

8

Specify a list of software packages you will be publishing

Specify a short name for each software package.

✖ *This question has not been answered yet!*

9

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)?

Much of todays data is used in comparison with reference data. You may be comparing your own data with a "standard set" which is maintained as a collection by someone else. Or you could be determining differences to a standard (for example in bioinformatics, a genome is often compared with a reference genome to identify genomic variants). Will you be creating any data that will be reference data for other researchers?

Data Stewardship for Open Science: *rbz*

✔ a. No

VII. 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.

Summary

Answered (current phase)	3 / 3
Answered	6 / 6

Metric	Score
Reusability	1.00
Openness	0.83

Questions

1

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

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

The FAIR principles do not contain any direction towards "Openness". This is done on purpose, because there can be compelling reasons not to make data "Open", such as privacy, other sensitive data, or intellectual property protection.

The true goal of funding agencies is to create the maximum value for society from their investments. They therefore often add "As open as possible, as closed as necessary" to the requirements for funding.

Data Stewardship for Open Science: *jvm*

✔ b. Yes

2

Can all of your data become completely open over time?

Horizon 2020 DMP | maDMP | Science Europe DMP | Horizon Europe DMP

Some data may be subject to a temporary embargo, or need to stay closed for specific reasons.

✔ b. Yes

3

Will you use temporary restrictions on the reuse of the data (embargo)?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ c. No, all data will be opened as soon as legal restrictions are falling away

4

Will metadata be available openly?

Horizon Europe DMP

Horizon Europe demands that metadata is completely open, e.g. provided under a CC0 license, and that any deviation from this is clarified.

✔ b. Yes

4.b.1

Will metadata contain instructions how to get access to the data?

Horizon Europe DMP

✔ b. Yes

4.b.2

Will the metadata be available in a form that can be harvested and indexed?

Horizon Europe DMP

Repositories often allow search engines and catalogues to index the metadata in an automated way. Will this be the case for your data?

✔ b. Yes, by the repository / repositories