Sequencing data at MMCI

Organization	DSW (researchers)
Created by	Radoslava Kacová (465089@muni.cz)
Based on	Life Sciences DSW Knowledge Model, 2.3.0 (dsw:lifesciences:2.3.0)
Project Phase	Before Submitting the Proposal
Created at	29 Jun 2022

Report generated by Data Stewardship Wizard <https://ds-wizard.org>

Summary

Answered (current phase)	67 / 67
Answered	201 / 213

Metric	Score
Findability	0.86
Accessibility	0.71
Interoperability	1.00
Reusability	0.65
Good DMP Practice	0.79
Openness	1.00

I. Administrative information

Summary

Answered (current phase)	4 / 4
Answered	17 / 24

Questions

1

Contributors

Horizon 2020 DMP **maDMP** **Science 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**

✓ Radoslava Kacová

1.a.2

E-mail address

Horizon 2020 DMP **maDMP** **Science Europe DMP**

✓ 465089@muni.cz

1.a.3

ORCID Identifier

Horizon 2020 DMP   maDMP   Science Europe DMP

✗ *This question has not been answered yet!*

1.a.4

Affiliation

Horizon 2020 DMP   Science Europe DMP

✓ Masaryk Memorial Cancer Institute

https://ror.org/0270ceh40

1.a.5

Role

Horizon 2020 DMP   maDMP   Science 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".

✓ e. Data Steward

1.b.1

Name

Horizon 2020 DMP   maDMP   Science Europe DMP

✓ Zdenka Dudová

1.b.2

E-mail address

Horizon 2020 DMP maDMP Science Europe DMP

✓ dudova@ics.muni.cz

1.b.3

ORCID Identifier

Horizon 2020 DMP maDMP Science Europe DMP

✗ *This question has not been answered yet!*

1.b.4

Affiliation

Horizon 2020 DMP Science Europe DMP

✓ Masaryk Memorial Cancer Institute

https://ror.org/0270ceh40

1.b.5

Role

Horizon 2020 DMP maDMP Science 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".

✓ e. Data Steward

2
Research Project(s)

Horizon 2020 DMP maDMP Science Europe DMP

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

Answers

2.a.1
Project name

Horizon 2020 DMP maDMP Science Europe DMP

✓ Design of Sequence data management and FAIRification at an oncological institution

2.a.2
Project acronym

Horizon 2020 DMP

✗ *This question has not been answered yet!*

2.a.3
Project abstract

Horizon 2020 DMP maDMP Science Europe DMP

✓ Masaryk Memorial Cancer Institute (MMCI) in Brno is a leading institution treating cancer in the Czech Republic. One of the examination techniques used to diagnose the disease is the pa-tient’s DNA/RNA sequencing. These sequencing data have been stored at the MMCI, but since they are not properly described and curated, it is diffi-cult to share them with potential reques-tors, and in addition, their value may de-crease over time. This diploma thesis aims to develop a strategy for bringing the se-quencing data to the researchers who can beneficially reuse them with accordance to FAIR principles.

2.a.4

Link to a project proposal or another description of the methods used in the project

✗ *This question has not been answered yet!*

2.a.5

Date the project will start

Horizon 2020 DMP Science Europe DMP maDMP

✓ 1. 10. 2021

2.a.6

Date the project will end

Horizon 2020 DMP Science Europe DMP maDMP

✓ 30. 6. 2022

2.a.7

Funding

Horizon 2020 DMP maDMP Science Europe DMP

Add all the funding that are part of this project.

Answers

2.a.7.a.1

Funder

Horizon 2020 DMP Science Europe DMP maDMP

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.

✗ *This question has not been answered yet!*

2.a.7.a.2

Funding status

Horizon 2020 DMP Science Europe DMP maDMP

✗ *This question has not been answered yet!*

2.a.7.a.3

Grant number

Horizon 2020 DMP maDMP Science Europe DMP

✗ *This question has not been answered yet!*

3

To execute the DMP, is additional specialist expertise required?

Horizon 2020 DMP Science Europe DMP

✓ a. No

4

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

Horizon 2020 DMP

✓ a. No

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)	4 / 4
Answered	5 / 5

Metric	Score
Reusability	1.00

Questions

1

Is there any pre-existing data?

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*

✓ b. Yes

1.b.1

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

Horizon 2020 DMP maDMP Science Europe DMP

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?

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.

1.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*

✓ b. Yes

1.b.2.b.1

Will you be making your harmonization results available to others?

By publishing either exactly what you did or (better) make sure that the harmonized data is available for reuse, you may save others the effort

✓ b. Yes

1.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 coming together in the project, when it will be coming. We also need to make sure that we have adequate storage space to deal with it, and that all the responsibilities have been taken care of.

Summary

Answered (current phase)	23 / 23
Answered	59 / 62

Metric	**Score**
Findability	1.00
Accessibility	0.00
Interoperability	1.00
Reusability	0.72
Good DMP Practice	1.00

Questions

1

What existing data formats/types will you be using?

Horizon 2020 DMP **Science 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

1.a.1

Data format/type

Horizon 2020 DMP Science Europe DMP

✓ FASTQ Sequence and Sequence Quality Format

FAIRsharing https://fairsharing.org/bsg-s000229

1.a.2

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

Horizon 2020 DMP Science Europe DMP

✓ b. Yes

1.a.3

Does this data format enable sharing and long term archiving?

Horizon 2020 DMP Science 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

1.a.4

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

Horizon 2020 DMP Science Europe DMP

✓ c. I can specify the number of files/subjects and the size of each

Specify an approximation of the expected data volume

1.a.4.c.1

Number of files/subjects

Horizon 2020 DMP Science Europe DMP

✓ 10000

1.a.4.c.2

Rough average size of each file/subject in gigabytes

Horizon 2020 DMP Science Europe DMP

✓ 0.5

1.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

2

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

Horizon 2020 DMP

Answers

2.a.1

Name

Horizon 2020 DMP

✓ NCI Thesaurus

FAIRsharing https://fairsharing.org/10.25504/FAIRsharing.4cvwxa

2.a.2

If you use a standard that is not in FAIRsharing, please specify its PID or URL

Horizon 2020 DMP

✗ *This question has not been answered yet!*

2.b.1

Name

Horizon 2020 DMP

✓ The Data Use Ontology

FAIRsharing https://fairsharing.org/10.25504/FAIRsharing.5dnjs2

2.b.2

If you use a standard that is not in FAIRsharing, please specify its PID or URL

Horizon 2020 DMP

✗ *This question has not been answered yet!*

2.c.1

Name

Horizon 2020 DMP

✓ EDAM Bioimaging Ontology

FAIRsharing https://fairsharing.org/10.25504/FAIRsharing.g593w1

> 2.c.2
>
> ### If you use a standard that is not in FAIRsharing, please specify its PID or URL
>
> **Horizon 2020 DMP**
>
> ✗ *This question has not been answered yet!*

3

Will you be using new types of data?

Horizon 2020 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

4

How will you be collecting and keeping your metadata?

Science Europe DMP **Horizon 2020 DMP**

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

Data Stewardship for Open Science: *rhm*

✓ a. Explore

There are many kinds of metadata, each serving their own purpose. Some key metadata that you should consider:

- There is metadata that helps identify where the data is coming from (e.g. who created it, title). For this the Dublin Core is often used.
- There are different ways of adding metadata to make the data "discoverable" for other researchers. This requires either keywords or ontology terms describing what is in the data.
- There is metadata describing how the data can be re-used, such as license information and, for data about people, the extent of their consent for data reuse.

- There is metadata that makes the data understandable, e.g. linking to the exact processes used to collect them (is a *body temperature* measured under the tongue or in the rectum?) and units (is a temperature given in Celsius or Fahrenheit?).
- There is metadata describing where the data comes from and what it is useful for. For frequently used data types, there are often very well defined metadata standards, in other cases you may need to think about this yourself. For each of these kinds of metadata there are specific standards. There is no single standard that will get you all the metadata needed to make the data as FAIR as possible.

4.a.1

What standard(s) will you use to specify author/title/keyword information?

Horizon 2020 DMP **Science Europe DMP**

There are a few different standards that are often used to give basic information about your dataset. Which ones of these will you be using?

✓ a. Explore

4.a.1.a.1

Will you document the data with Dublin Core metadata?

Horizon 2020 DMP **Science Europe DMP**

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?

External Links: *Dublin Core Metadata Terms*, *Dublin Core Initiative*

✓ a. No

4.a.1.a.2

Will you document the data with DataCite metadata

Science Europe DMP **Horizon 2020 DMP**

External Links: *DataCite metadata schema*

✓ a. No

4.a.1.a.3

Will you document the data with DDI metadata

Horizon 2020 DMP Science Europe DMP

DDI metadata is more extensive than Dublin Core and DataCite, it details more of what is in the data and really can help other researchers locate your data set as an interesting source.

External Links: *DDI metadata documentation*

✓ b. Yes

4.a.2

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

Horizon 2020 DMP

Many research fields have worked together to define what kind of metadata should really be collected when an experiment of a certain kind is performed and described. That information is described in a Minimal Metadata Standard. Often, these standards describe both what kind of information needs to be collected as well as the format in which it is expected.

External Links: *FAIRsharing repository of standards*

✓ b. Yes

4.a.2.b.1

Which "Minimal Information" standards will you use?

Horizon 2020 DMP

Answers

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

Minimal Information Standard

Horizon 2020 DMP

✓ Minimum Information about a (Meta)Genome Sequence

FAIRsharing https://fairsharing.org/10.25504/FAIRsharing.va1hck

4.a.3

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

4.a.4

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 add "optional" metadata where I can

4.a.4.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

4.a.4.b.2

Do you need to exchange your data with others?

✓ b. Yes

4.a.5

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

4.a.5.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, 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.

✓ a. No

4.a.5.a.2

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

✓ a. No

4.a.6

Do all datasets you work on have a license?

It is not always clear to everyone in the project (and 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

4.a.6.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

4.a.7

How will you do file naming and file organization?

Horizon 2020 DMP

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

✓ a. Explore

> 4.a.7.a.1
>
> **Did you make a SOP (Standard Operating Procedure) for file naming?**
>
> **Horizon 2020 DMP**
>
> It can help if everyone in the project uses the same naming scheme.
>
> ✓ b. Yes
>
> > 4.a.7.a.1.b.1
> >
> > **Describe your SOP (Standard Operating Procedure) for file naming**
> >
> > **Horizon 2020 DMP**
> >
> > Describe how everyone in the project will be naming files and folders, and what folder structure you will use.
> >
> > ✓ -
>
> 4.a.7.a.2
>
> **Will you be keeping the relationships between data clear in the file names?**
>
> **Horizon 2020 DMP**
>
> Advice: Use the same identifiers for sample IDs etc throughout the entire project.

✓ b. Yes

4.a.7.a.3

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

Horizon 2020 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

4.a.7.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*

✓ a. No

4.a.8

How will you be keeping track of the "provenance" of the data?

Horizon 2020 DMP Science Europe DMP

Data analysis is normally done step-by-step. It is essential to make sure all steps are properly documented, otherwise results will not be reproducible. Re-users of the data also need this information to decide whether the data can be used for their purpose.

✓ c. We use other arrangements

4.a.8.c.1

What other arrangements?

Horizon 2020 DMP Science Europe DMP

✓ -

4.a.9

Will you be documenting the data with W3C PROV provenance?

Horizon 2020 DMP Science Europe DMP

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?

External Links: *W3C Prov primer*

✓ a. No

4.a.10

Will you use a workflow system that automatically keeps track of the steps in the analysis?

Some workflow systems automatically keep track of which steps were done in data analysis and what options were selected. This can help document the data for reproducibility.

✓ a. No

5

Will you be acquiring data using measurement equipment?

Horizon 2020 DMP Science Europe DMP

✓ a. No

6

Do you have any non-equipment data capture?

Horizon 2020 DMP Science 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*

✓ b. Yes

6.b.1

Will you be collecting questionnaires?

Horizon 2020 DMP Science Europe DMP

✓ a. No

6.b.2

Will you be collecting case report forms?

Horizon 2020 DMP Science Europe DMP

Data Stewardship for Open Science: *hfg*

✓ a. No

6.b.3

Will you be collecting data from electronic patient records?

Horizon 2020 DMP Science Europe DMP

✓ b. Yes

6.b.3.b.1

Has access to the electronic patient records been arranged?

Horizon 2020 DMP

✓ a. Not yet

6.b.4

Please list all non-equipment data sets you will collect

Horizon 2020 DMP Science Europe DMP

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

Answers

6.b.4.a.1

Data set:

Horizon 2020 DMP Science Europe DMP

✓ -

6.b.4.a.2

Description

Horizon 2020 DMP Science Europe DMP

Briefly describe the contents of this data set

✓ -

6.b.4.a.3

How will the data be captured?

✓ a. All immediately in digital form

6.b.4.a.4

Will all data be collected by a single person?

✓ a. This data will be collected by one person

6.b.4.a.5

Who else could be interested in using this data?

Horizon 2020 DMP

✓ b. Other researchers working in the same field of research

7

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

✓ a. No

7.a.1

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

✓ b. Yes

8

Will you be storing physical samples?

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

Data Stewardship for Open Science: *kuz*

✓ b. Yes

8.b.1

How will the samples be identified?

✓ a. Samples will get an internal code

> 8.b.2
>
> **Where will the samples be stored**
>
> ✓ at MMCI biobank

9

Will you need consent for any newly collected personal data?

Horizon 2020 DMP maDMP Science Europe DMP

✓ d. Yes, we will collect consent for our use as well as reuse of the data

10

How is the ownership of the collected data arranged?

Horizon 2020 DMP Science Europe DMP

✓ c. All data will be owned by the institute

IV. Processing data

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

Summary

Answered (current phase)	19 / 19
Answered	60 / 60

Metric	Score
Accessibility	1.00
Reusability	0.58
Good DMP Practice	0.77

Questions

1

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

Horizon 2020 DMP **Science Europe DMP**

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*.

✓ a. No

> 1.a.1
>
> **Are data that project members 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.

✓ a. Explore

2.a.1

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?

✓ b. Yes

2.a.1.b.1

Will you use a folder for each sample/subject?

Science Europe DMP

✓ b. Yes

2.a.1.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.

✓ -

2.a.1.b.2

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

Science Europe DMP

✓ b. Yes

2.a.1.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.

✓ -

2.a.1.b.3

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

Science Europe DMP

✓ a. No

2.a.1.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.

✓ -

2.a.2

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

Science Europe DMP

✓ a. No

2.a.3

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

Science Europe DMP

✓ b. Yes

2.a.3.b.1

How will you handle changes in the data?

Science Europe DMP

Database systems can be configured to keep all data, so that it is possible to reconstruct any past state of the data. How are changes in the data handled by your database?

✓ c. Modifications will be made by Expiring the existing data and Adding updated data

2.a.4

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

Science Europe DMP

✓ a. No

2.a.5

Will you be storing data in a triple store?

Science Europe DMP

✓ a. No

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.

✓ b. More guidance is desired

3.b.1

Will you be exploring parameters to the workflow, or run in bulk?

What will be the operational mode for your workflows? Will you be exploring options by changing tools and tweaking parameters? Of will you be running the same exact workflow on a large number of data files?

Data Stewardship for Open Science: *qzt*

✓ b. We will be running in bulk

3.b.2

What data will the workflow developers or implementers use?

The people implementing the data analysis work flow for your project probably need test data that they can use to see whether what they build works. How will this be arranged?

✓ c. They can use data from our project

3.b.2.c.1

When will they have access?

✓ a. Data will be available at the start of the project, no waiting needed

3.b.2.c.2

How will data security be dealt with?

✓ d. We have made other arrangements

3.b.2.c.2.d.1

What other arrangements?

✓ pseudonymization

3.b.3

List existing software components you will use in the analysis workflow

Your workflow may be available in components from different sources. Specify the different parts that you recognize and that you will each acquire in a different way

Answers

3.b.3.a.1

Software component:

✓ -

3.b.3.a.2

Where are you getting this software from? Please specify a web address if available.

✓ -

3.b.3.a.3

What version of this software will you use?

✓ a. The exact version that we point to

3.b.3.a.4

How is your experience with this software?

✓ d. We will use the manual

3.b.4

List new software components you will develop for the analysis workflow

Not all components you need may be available already. Please list here what you will be developing yourself. Do not underestimate the time needed to integrate components into a work flow!

Answers

3.b.4.a.1

Software component:

✓ -

3.b.4.a.2

Please specify the software repository you use for development

Preferably use a direct URL other users could use

✓ -

3.b.4.a.3

Did you consider existing options?

✓ d. This will be one of the prime distinctive outcomes of our project

3.b.4.a.4

What license will you use for your tool?

Make sure the license is compatible with all components you use, and also make sure the license is made explicit in the repository.

✓ d. LGPL 3.0 or later

3.b.5

Did you choose the workflow engine you will be using?

Data Stewardship for Open Science: *ydj*

✓ b. Yes, we will be using what we always use

Make sure that you are not missing out on alternatives that would have better properties for the project

3.b.6

Do you plan taking special measures to guaranty the integrity of tools in the workflow?

✓ a. No

Consider changing this!

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 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*

✓ a. No

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*

✓ a. No

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

✓ a. Explore

5.a.1

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

Horizon 2020 DMP

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.

✓ b. Yes

5.a.2

Will you be instrumenting the tools into pipelines and workflows using automated tools?

Horizon 2020 DMP

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.

✓ b. Yes

5.a.3

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

Horizon 2020 DMP

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 the data set repeatedly to catch unexpected changes in results?

Horizon 2020 DMP

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 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

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
>
> 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
>
> ✓ a. No
>
> **7.a.2**
>
> **Do project members carry data with them?**
>
> Horizon 2020 DMP Science 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'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 project members send project data or reports per e-mail or other messaging services?

✓ a. No

7.a.5

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

Horizon 2020 DMP **Science Europe DMP**

✓ b. Yes

7.a.6

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

Horizon 2020 DMP **Science 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**

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**

✓ b. Yes; the effect is small

7.a.9

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

Horizon 2020 DMP **Science Europe DMP**

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

7.a.10

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

Horizon 2020 DMP **Science 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

✓ b. Yes, all personal information will be processed in pseudonymized form only

7.a.11.b.1

How is pseudonymization handled?

Horizon 2020 DMP Science Europe DMP

✓ a. We pseudonymize inside the project, only limited people can access the keys

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)	5 / 5
Answered	14 / 16

Metric	Score
Interoperability	1.00
Good DMP Practice	1.00

Questions

1

List the data formats you will be using and 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.

Answers

1.a.1

Data type

✓ FASTQ Sequence and Sequence Quality Format

FAIRsharing https://fairsharing.org/10.25504/FAIRsharing.r2ts5t

1.a.2

How is this data structured?

> ✓ e. Differently
>
> > **1.a.2.e.1**
> >
> > **Describe the structure of the data**
> >
> > ✓ -

2

Will you be doing integration 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.

This is an advanced subject that you may want to skip if this is not an issue for you. On the other hand, if this is your expertise we would like your help in improving the questions in this section.

✓ a. No

3

Will you be using common or exchangeable units?

✓ b. Yes

4

Will you be using common ontologies?

✗ *This question has not been answered yet!*

5

Will there be potential issues with statistical normalization?

✓ a. No

6

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

✗ *This question has not been answered yet!*

7

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

✓ b. Yes

7.b.1

Did you already select the variables on which you will join the data sets?

✓ a. No

7.b.2

Will you make sure that you do not inadvertently create a biased subset?

Some parameters you select on may have been collected only for a subset of the subjects or data points. An obvious example is if you match on secondary education type, you will bias to people over 18 years old because younger people do not have this field. In many cases the selection bias may be a lot less obvious and special measures exist to verify that the diversity of the sample is not reduced by the integration step.

✓ b. Yes

7.b.3

Could the coupling of data create a danger of re-identification of anonymized privacy sensitive data?

✓ a. No

7.b.4

Did you make a conscious decision to be either accurate or complete?

If the coupling parameters are lenient, you will find more connections than when they are strict. But you may find that they are less accurate. This is a balance.

✓ c. Completeness of the mapping is most important

8

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

✓ a. No

9

Will you be doing (automated) knowledge discovery?

Data Stewardship for Open Science: *bzu*

✓ a. No

VI. Preserving data

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

Summary

Answered (current phase)	10 / 10
Answered	44 / 44

Metric	**Score**
Findability	0.67
Accessibility	0.77
Reusability	0.65
Good DMP Practice	0.75

Questions

1

Specify a list of data sets you will be producing

Horizon 2020 DMP maDMP Science 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.

Answers

1.a.1

Data set:

Horizon 2020 DMP maDMP Science Europe DMP

✓ Sequencing Data

1.a.2

Description of the data set

Science Europe DMP maDMP

What type of data is in this data set? Examples could be "Field observations", "raw instrument data", "genomic variants".

✓ raw instrument data

1.a.3

Identifier of the data set

Horizon 2020 DMP Science Europe DMP maDMP

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).

Answers

1.a.3.a.1

What type of identifier?

Science Europe DMP maDMP Horizon 2020 DMP

Which type of identifier is this?

✓ e. Other

1.a.3.a.1.e.1

What is the identifier type?

maDMP Science Europe DMP Horizon 2020 DMP

✓ UUID

1.a.3.a.2

The actual identifier

Science Europe DMP maDMP Horizon 2020 DMP

✓ mmci_GDH6jR02jkP2

1.a.4

Will this data set be published?

Horizon 2020 DMP maDMP Science Europe DMP

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.

✓ a. No

1.a.5

How long will this data set be kept?

Horizon 2020 DMP Science Europe DMP

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.

✓ a. As long as technically possible

1.a.6

Will the metadata be available even when the data no longer exists?

Horizon 2020 DMP Science Europe DMP

This is a one of the FAIR principles.

✓ b. Yes

1.a.7

Does this dataset contain personal data?

Horizon 2020 DMP · Science Europe DMP · maDMP

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!

✓ b. Yes

1.a.8

Does this dataset contain sensitive information?

Horizon 2020 DMP · Science Europe DMP · maDMP

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.

✓ b. Yes

2

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

Horizon 2020 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

✓ b. Yes

2.b.4
Will the archive need to be protected against loss or theft?

Horizon 2020 DMP

✓ b. Yes

> **2.b.4.b.1**
> **Will the archive be encrypted?**
>
> Horizon 2020 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
> >
> > ✓ b. Yes

2.b.4.b.2

Is it clear who has physical access to the archives?

Horizon 2020 DMP

✓ b. Yes

2.b.5

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

Data Stewardship for Open Science: *ybd*

✓ 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?

✓ a. No

2.b.6.b.2

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

✓ a. No

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

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

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

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

✓ a. No

3.b.2

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

✓ b. Yes

4

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

Horizon 2020 DMP Science Europe DMP

✓ b. Yes

4.b.1

How will you be paying for these services?

Horizon 2020 DMP Science 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**

✓ a. No

7

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

8

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

✓ c. Yes, all data sets will be linked from a single catalog entry

9

Specify a list of software packages you will be publishing

Specify a short name for each software package.

Answers

> 9.a.1
>
> **Software package:**
>
> ✓ -

9.a.2

Will you be adding a proper open-source license?

✓ c. Yes, we will decide on an open source license

9.a.3

Where will the software package be available?

✓ -

9.a.4

Will this software be listed in a catalogue?

✓ a. No

10

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

11

Will you do systems biology modeling?

✓ a. No

12

Will you do structural modeling?

✓ 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)	2 / 2
Answered	2 / 2

Metric	Score
Openness	1.00

Questions

1

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

Horizon 2020 DMP Science 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 immediately?

Horizon 2020 DMP maDMP Science Europe DMP

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

✓ b. Yes