NOPD_Misconduct_Complaints

Organization	DSW (researchers)
Created by	Zhiyuan Pan (panzy3272824@gmail.com)
Based on	Common DSW Knowledge Model, 2.6.0 (dsw:root:2.6.0)
Project Phase	Before Submitting the Proposal
Created at	26 Jul 2023

Summary Report

Summary

Answered (current phase)	61 / 62
Answered	144 / 187

Metric	Score
Findability	1.00
Accessibility	1.00
Interoperability	0.92
Reusability	0.93
Good DMP Practice	1.00
Openness	1.00

I. Administrative information

Summary

Answered (current phase)	5 / 5	
Answered	19 / 25	

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

✔ Zhiyuan Pan

1.a.2

E-mail address

Horizon 2020 DMP maDMP Science Europe DMP Horizon Europe DMP

✔ zhiyuan.pan@uni-jena.de

1.a.3

ORCID Identifier

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

✔ https://orcid.org/0009-0007-8074-0533

1.a.4

Affiliation

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✖ *This question has not been answered yet!*

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".

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

☐ Other

2

Research Project(s)

ELSI Horizon 2020 DMP maDMP Science 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.

Answers

2.a.1

Project name

Horizon 2020 DMP maDMP Science Europe DMP Horizon Europe DMP

✔ Analysis on NOPD Misconduct Complaints

2.a.2

Project acronym

Horizon 2020 DMP Horizon Europe DMP

✘ *This question has not been answered yet!*

2.a.3

Project abstract

Horizon 2020 DMP maDMP Science Europe DMP Horizon Europe DMP

✔ This dataset represents complaints of misconduct originated by a citizen either directly to NOPD or through the IPM or by an employee of the Police Department per NOPD Misconduct Complaint Intake and Investigation policy.

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

Starting date of the project

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

✔ 2016-11-02

2.a.6

Ending date of the project

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

✔ 2023-07-25

2.a.7

Funding

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

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 | Horizon Europe DMP

✖ *This question has not been answered yet!*

2.a.7.a.3

Grant number

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

✖ *This question has not been answered yet!*

2.a.8

Does the project require ethical approval?

ELSI

If the project involves testing on animals or processing of personal data, an approval is likely needed. It is the responsibility of the Researcher to make sure that the project and associated data is in agreement with current legislation.

✔ a. No

3

To execute the DMP, is additional specialist expertise required?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ b. Yes, trained support staff is available

4

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

Horizon 2020 DMP Horizon Europe DMP

✔ b. Yes

4.b.1

What specific hard/software do you need, and why?

Horizon 2020 DMP Science Europe DMP Horizon Europe DMP

✔ Python, Excel, Collectica, Collectica Reader

5

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

Horizon Europe DMP

Answers

5.a.1

Name of the policy or process

Horizon Europe DMP

✔ Fachkollegium "Sozialwissenschaften" zum Umgang mit Forschungsdaten in der Soziologie, der Politikwissenschaft und der Kommunikationswissenschaft (2020)

5.a.2

Give a link/reference to the policy or process

Horizon Europe DMP

✔ https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/forschungsdaten/empfehlungen/index.html

5.a.3

Description of how and why you are using this policy or process

Horizon Europe DMP

✔ All research institutions have to implement these guidelines in order to be eligible to receive DFG funding.

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	26 / 27

Metric	Score
Reusability	1.00

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?

✔ b. Yes

There are data set search engines that can help you find diverse data on the web, such as:

- Google Dataset Search
- OpenAIRE
- FAIRsharing

Also, many governments run their own "government data portals".

1.b.1

Will you need existing data from Natural Sciences?

✔ a. No

1.b.2

Will you need existing data from Engineering and Technology

✔ a. No

1.b.3

Will you need existing data from Medical and Health Sciences

✔ a. No

1.b.4

Will you need existing data from Agricultural Sciences

✔ a. No

1.b.5

Will you need existing data from Social Sciences

✔ b. Yes

We are interested to receive your expertise on this topic. You can use the feedback button if you have resources that we could list here

1.b.6

Will you need existing data from Humanities

✔ a. No

2

Is there any pre-existing data?

ELSI 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 openly available standard reference data did you consider re-using?

ELSI 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

✔

NOPD_Misconduct_Complaints

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.

✔ https://data.nola.gov/Public-Safety-and-Preparedness/NOPD-Misconduct-Complaints/gz2m-ef5u

2.b.1.a.3

Will you be using this reference data set?

ELSI | Horizon 2020 DMP | Horizon Europe 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

✔ City of New Orleans Police Department

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

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

ELSI | Horizon 2020 DMP | Science 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*

✔ a. They are freely available for any use (public domain or CC0)

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.

✔ latest version

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?

✔ b. I will keep a copy and make it available with my results

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

What will you use this reference data set for?

Horizon Europe DMP

✔ Do some analysis of the influence of relevant factors.

2.b.2

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

ELSI | 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*

✓ b. Yes

2.b.4.b.1

Will you be making the data itself available in computer readable form to others?

Horizon Europe DMP

It is advisable to make the result of your efforts available in a standard repository. If you answer 'yes' here, do not forget to mention the dataset under the chapter 'giving access to data'.

✔ b. Yes, through a standard repository

2.b.4.b.2

Will you provide machine readable, standardised metadata to others?

Horizon Europe DMP

✔ b. Yes

If the data is published in a standard repository, that repository will tell you which metadata standards to use.

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

Which metadata standards will you use?

Horizon Europe DMP

Answers

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

Metadata Standard

Horizon Europe DMP

✔

Data Documentation Initiative Lifecycle (DDI-Lifecycle) type model and format

Data Documentation Initiative (DDI) Lifecycle (DDI-Lifecycle, DDI-L) is designed to document and manage data across the entire life cycle, from conceptualization to data publication, analysis and beyond. The freely available international DDI standards describe data that result from observational methods in the social, behavioral, economic, and health sciences. DDI is used to document data in over 80 countries of the world. It encompasses all of the DDI-Codebook specification and extends it. Based on XML Schemas, DDI-Lifecycle is modular and extensible.

FAIRsharing https://fairsharing.org/10.25504/FAIRsharing.1t5ws6

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)	13 / 13
Answered	27 / 39

Metric	Score
Findability	1.00
Interoperability	0.75
Reusability	0.85

Questions

1

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

ELSI

✔ a. No

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*

✖ *This question has not been answered yet!*

4

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

4.a.1

Data format/type

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔

Comma-separated Values (CSV) type: model and format

A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A

CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

FAIRsharing https://fairsharing.org/10.25504/FAIRsharing.1943d4

4.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

4.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

4.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

4.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.

✖ *This question has not been answered yet!*

5

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

Horizon 2020 DMP Horizon Europe DMP

If you want to facilitate interoperability of your data with data you get from elsewhere, and facilitate re-use of your own data later, it is handy if you use standard ways to encode certain data. Such standard encodings are sometimes called terminologies or ontologies.

External Links: *FAIRCookbook introduction to terminologies and ontologies*, *FAIRCookbook on selecting terminologies and ontologies*, *FAIRCookbook on Selecting an ontology lookup service*

✖ *This question has not been answered yet!*

6

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

7

How will you be collecting and keeping your metadata?

ELSI | 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 – 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)*, *FAIRCookbook on FAIR and the notion of Metadata*

✔ 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.

7.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

7.a.1.a.1

Will you document the data with Dublin Core metadata?

Horizon 2020 DMP | Science Europe DMP | Horizon 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*

✔ b. Yes

7.a.1.a.2

Will you document the data with DataCite metadata

Science Europe DMP | Horizon 2020 DMP | Horizon Europe DMP

External Links: *DataCite metadata schema*

✖ *This question has not been answered yet!*

7.a.1.a.3

Will you document the data with DDI metadata

Horizon 2020 DMP | Science Europe DMP | Horizon 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

7.a.1.a.4

Will you be including keywords or relevant ontology references to optimise the possibility for discovery and potential reuse?

Horizon Europe DMP

✔ b. Yes

7.a.2

Will you be creating a data/variable dictionary?

If anyone wants to see whether they can re-use your data, it must be possible to see what information in contains. For this, a data dictionary can play an important role.

External Links: *FAIRCookbook on creating a data/variable dictionary*

✔ b. Yes

7.a.3

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

Horizon 2020 DMP Horizon Europe 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*

✔ a. No

Did you really check a service like fairsharing.org to verify this?

7.a.3.a.1

Do you have a good idea of what metadata is needed to make it possible for others to read and interpret your data in the future?

Horizon 2020 DMP Horizon Europe DMP

Imagine yourself trying to interpret a similar data set from a colleague in a few years. Is there any way in which the data could be misinterpreted? Does everyone in the field do things in exactly the same way? And will that still be known in 10 years? How can such misinterpretation be avoided?

External Links: *FAIRsharing repository of standards*

✔ b. Yes

7.a.4

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

7.a.5

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

7.a.5.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

7.a.5.b.2

Do you need to exchange your data with others?

✔ b. Yes

7.a.6

Will a license be assigned to your datasets?

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.

External Links: *RDMkit on Licensing*

✔ b. Yes

7.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 course *as open as possible*) stored in close association with the data.

Data Stewardship for Open Science: *atw*

✔ b. Yes

7.a.7

Where do you keep any applicable restrictions on the usage of the data?

External Links: *Data Use Ontology*

✖ *This question has not been answered yet!*

7.a.8

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

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

Data analysis is normally done step-by-step. It is important that for all data the origin and all processing and filtering steps are 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.

In computing, systems like Galaxy and (Jupyter) notebooks often automatically keep provenance information.

✖ *This question has not been answered yet!*

7.a.9

Will you be documenting the data with W3C PROV provenance?

Horizon 2020 DMP | Science Europe DMP | Horizon 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: *FAIRCookbook on Provenance information*, *W3C Prov primer*

✖ *This question has not been answered yet!*

7.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.

✖ *This question has not been answered yet!*

8

Will you be acquiring data using measurement equipment?

ELSI | Horizon 2020 DMP | Science Europe DMP

✔ a. No

9

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

10

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

✔ a. No

10.a.1

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

External Links: *FAIRCookbook on tools to convert data to RDF*

✔ a. No

11

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

ELSI

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

12

Is the data collection subject to ethical legislation?

ELSI

✖ *This question has not been answered yet!*

13

Data use restrictions

ELSI

Are there any limitations on the data use e.g. only for research on certain types of diseases, sharing only within certain geographical boundaries, etc.?

Describing data use in a formalised way by using the data use ontology (DUO) greatly improves the data reusability.

External Links: *Data Use Ontology*, *FAIRCookbook on permitted uses of data*

✖ *This question has not been answered yet!*

14

How are the rights of the collected data arranged?

Horizon 2020 DMP Science Europe DMP

✖ *This question has not been answered yet!*

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*

✖ *This question has not been answered yet!*

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)	19 / 19
Answered	29 / 37

Metric	Score
Accessibility	1.00
Reusability	1.00
Good DMP Practice	1.00

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 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*.

✔ b. Yes

1.b.1

Will this work space be run by dedicated specialists?

If your work space is run and maintained by specialists, e.g. the ICT department of one of the institutes involved in the projects, this means that backup and restore as well as access management is properly addressed.

✔ b. Yes

1.b.2

How will you work with your data?

Science Europe DMP

There are several questions regarding the dynamics of the data in the working area, who works with it, the software that is run on it, etc.

External Links: *RDMkit on Data Processing*, *RDMkit on Data Analysis*

✔ a. Explore

1.b.2.a.1

What kind of data will you have in your work space?

When making the work space, it helps to know whether you expect to work with very many small files, a few very large files, whether you will use a (SQL) database to store most of the data. Maybe your data is suitable for a system like Hadoop? Such information can be collected here.

✔ .csv .xml

1.b.2.a.2

Do you need the work space to be close to the compute capacity?

If you have large volumes of data that are intensely and repeatedly used by the computing work flow, it may be needed to keep the storage in the same place as where the computing takes place.

Data Stewardship for Open Science: *wia*

✔ a. No

1.b.2.a.3

Will you be working with your data in another form than the way it will be archived?

Archival and working with data have different requirements. You want archived information to be in a form that others could read and in a format that is also understandable in a number of years. When working with the data, you need to be able to address it efficiently. If the two differ, you need to plan for conversions.

✔ a. No, data format will be archived in the same way we work with it

1.b.2.a.4

How does the storage need change over time?

Science Europe DMP

To perform capacity planning, it is important to know what the need for storage capacity at the beginning and the end of the project will be.

✔ c. Storage needs are small at the beginning and will grow later

1.b.2.a.5

Will you need to temporarily archive data sets (e.g. to tape)?

Usually, data sets will be archived if it is unlikely you need them in the short term, but it would be hard to create them again, and/or they are essential for reproducing your work. Archival storage of large volumes can be significantly cheaper than keeping it in the working area for an extensive period.

✔ a. No

1.b.2.a.6

How will your first data come in?

External Links: *RDMkit on Data Transfer*

✔ c. We will need a high-speed network connection to copy the initial data

1.b.2.a.7

How will project partners access the work space?

✔ a. Explore

1.b.2.a.7.a.1

Who will arrange access control?

✔ c. The work space should be connected to a single-sign-on system

1.b.2.a.7.a.2

Will the work space storage need to be remote mounted?

✔ a. No

1.b.2.a.7.a.3

Will data be copied out and in to the workspace storage by remote users?

✘ *This question has not been answered yet!*

1.b.3

How available/reliable should must the work space be?

Science Europe DMP

There are a number of questions that can help you to decide whether your work space will be reliable enough for your project.

✔ a. Explore

1.b.3.a.1

How do you prevent a total loss of data in the work space?

Science Europe DMP

✔ b. All essential data is also stored elsewhere

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

Is there software in the work space? Can it also be restored quickly?

Data Stewardship for Open Science: *cbq*

✔ c. Special care will be taken for the software and configurations

1.b.3.a.2

Can you handle it when the work space is off line for a while?

✔ c. Problems during the evenings and weekends can not wait for work hours to be repaired

1.b.3.a.3

How long can you wait for a restore if the storage fails?

✔ c. No waiting is possible, a hot copy must be ready to take over

1.b.3.a.4

How long can you wait for a restore if you accidentally damage a file?

✔ c. Any user needs to be able to restore an old copy instantaneously

1.b.3.a.5

Will you make backup copies of project data that is not in the work space?

Science Europe DMP

Are there any data files e.g. on laptops of project members? Also: supercomputing centers and other high performance computer centers often write in their terms of use that you need to take care of your own backups.

✔ d. We make (automated) backups of all data stored outside of the working area

1.b.4

How will access to the work space be controlled?

Science Europe DMP

✔ c. Only project members will have read access; only selected project members will be able to write data

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*

✖ *This question has not been answered yet!*

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*

✖ *This question has not been answered yet!*

2.a.4

Will you use a database system to store project data?

Science Europe DMP

✖ *This question has not been answered yet!*

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.

✖ *This question has not been answered yet!*

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*

✖ *This question has not been answered yet!*

5

How will you validate the integrity of the results?

Horizon 2020 DMP Horizon Europe 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 Horizon Europe 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 Horizon Europe 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 Horizon Europe 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.

✔ b. Yes

5.a.4

Will you run part of the data set repeatedly to catch unexpected changes in results?

Horizon 2020 DMP Horizon Europe 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*

✔ b. Yes

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?

ELSI 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.

✖ *This question has not been answered yet!*

8

Do you have a contingency plan?

What will you do if the compute facility is down?

✖ *This question has not been answered yet!*

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)	4 / 4	
Answered	13 / 20	

Metric	Score	
Interoperability	1.00	

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

✔

Comma-separated Values (CSV) type model and format

A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.

1.a.2

How is this data structured?

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

1.a.2.b.1

Does each column have a header?

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

✔ b. Yes

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

Are all column headers unambiguous?

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

✔ b. Yes

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

Do all columns/headers have a data type?

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

✔ b. Yes

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

Are the limitations to allowed data values in each column explicit?

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

External Links: *Rightfield: Template fields in Microsoft Excel*

✔ b. Yes

1.a.2.b.2

Is it clear what a row in the table represents?

✔ b. Yes

1.a.2.b.3

Does each row have an identifier?

✔ b. Yes

1.a.2.b.4

Is there a distinguishing way a missing value in the table can be recognized?

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

✔ b. Yes

1.a.2.b.5

Is the relation between each of the columns and the record identifier clear?

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

✔ b. Yes

1.a.2.b.6

Are all the relations between the column headers explicit?

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

✔ b. Yes

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.

You may also need to use "mappings" to match data together.

External Links: *FAIRCookbook on Interlinking data using mappings*

✔ a. No

3

Will you be using common ontologies?

External Links: *FAIRCookbook on selecting an ontology lookup service*

✖ *This question has not been answered yet!*

4

Will there be potential issues with statistical normalization?

✖ *This question has not been answered yet!*

5

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

✖ *This question has not been answered yet!*

6

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

ELSI

✖ *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)	12 / 12
Answered	24 / 32

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

Questions

1

Specify a list of data sets you will be producing

ELSI 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.

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

✖ *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?

ELSI 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*

✔ b. Yes

2.b.1.b.1

Do you need frequent backups?

The general term 'backup' is used for protection against two different kinds of problems: equipment failure and human error. Protections against these two may need different solutions. Both are considered backup for this and subsequent questions.

✔ b. Yes, data changes frequently

2.b.1.b.2

Will you be relying on these backups to recover from human error (accidental changes or deletions)?

✔ b. Yes

This puts other demands on the possibility to restore files. Make sure this is covered

2.b.2

Will the archive be stored on disk or on tape?

✗ *This question has not been answered yet!*

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

✓ 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

✗ *This question has not been answered yet!*

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?

✓ b. Yes

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

✘ *This question has not been answered yet!*

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

✘ *This question has not been answered yet!*

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

✖ *This question has not been answered yet!*

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?

✔ b. Yes

Make sure this will be kept running by the department or institute. It is best to also have a backup plan, being able to move data and documents to a different place if a service is discontinued. For this, you may need to arrange permission from all project partners beforehand.

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

✗ *This question has not been answered yet!*

7

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

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 / 7

Metric	Score
Reusability	1.00
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 | 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?

ELSI | 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

3

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

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

✔ d. Yes, data without legal restrictions will be released openly after a fixed time period

3.d.1

What embargo period are you using?

Horizon 2020 DMP | Science Europe DMP | Horizon Europe DMP

After what period will restrictions on the reuse of data (except ethical and legal restrictions) fall away?

✖ *This question has not been answered yet!*

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