Project name

|  |  |
| --- | --- |
| Organization | My Organization |
| Created by | Albert Einstein ([albert.einstein@example.com](mailto:albert.einstein@example.com)) |
| Based on | Common DSW Knowledge Model, 2.3.0 (dsw:root:2.3.0) |
| Project phase | Before Submitting the DMP |
| Created at | 20 Aug 2021 |

### Summary

#### Answered indication

|  |  |  |
| --- | --- | --- |
| Answered (current phase) | 57 / 80 | 71.25 % |
| Answered | 64 / 99 | 64.65 % |

#### Metrics

| Metric | Score |
| --- | --- |
| Findability | 0.50 |
| Accessibility | 0.00 |
| Interoperability | 0.50 |
| Reusability | 0.78 |
| Good DMP Practice | 0.75 |
| Openness | 0.66 |

# I. Administrative information

### Summary

#### Answered indication

|  |  |  |
| --- | --- | --- |
| Answered (current phase) | 3 / 6 | 50.00 % |
| Answered | 4 / 9 | 44.44 % |

#### 1. Contributors

🏷️ **Tags:** *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 items)

#### 1.a.1. Name

🏷️ **Tags:** *Horizon 2020 DMP*, *maDMP*, *Science Europe DMP*

✔️ Marek Suchánek

#### 1.a.2. E-mail address

🏷️ **Tags:** *Horizon 2020 DMP*, *maDMP*, *Science Europe DMP*

✔️ marek.suchanek@ds-wizard.org

#### 1.a.3. ORCID Identifier

🏷️ **Tags:** *Horizon 2020 DMP*, *maDMP*, *Science Europe DMP*

❗ *This question has not been answered yet!*

#### 1.a.5. Role

🏷️ **Tags:** *Horizon 2020 DMP*, *maDMP*, *Science Europe DMP*

Roles in a project should be given as they are defined by [datacite](https://schema.datacite.org/meta/kernel-4.1/doc/DataCite-MetadataKernel_v4.1.pdf).

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".

✔️ a. Contact Person

# 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 indication

|  |  |  |
| --- | --- | --- |
| Answered (current phase) | 8 / 17 | 47.06 % |
| Answered | 10 / 23 | 43.48 % |

#### 1. Is there any pre-existing data?

🏷️ **Tags:** *Horizon 2020 DMP*, *maDMP*, *Science Europe DMP*

📖 **Data Stewardship for Open Science:** [*atq*](http://localhost:8080/book-references/atq)

🔗 **External Links:** [*Google dataset search*](https://datasetsearch.research.google.com), [*Datacite Search*](https://search.datacite.org)

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

✔️ b. Yes

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

🏷️ **Tags:** *Horizon 2020 DMP*, *maDMP*, *Science Europe DMP*

📖 **Data Stewardship for Open Science:** [*ezi*](http://localhost:8080/book-references/ezi)

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?

✔️ b. Yes

#### 1.b.1.b.1. What reference data will you use?

🏷️ **Tags:** *Horizon 2020 DMP*, *Science Europe DMP*

📖 **Data Stewardship for Open Science:** [*quc*](http://localhost:8080/book-references/quc)

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?

#### Answers (1 items)

#### 1.b.1.b.1.a.1. Reference database or dataset

🏷️ **Tags:** *Horizon 2020 DMP*, *Science 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

✔️ Very interesting dataset

#### 1.b.1.b.1.a.2. Where is this data available

🏷️ **Tags:** *Horizon 2020 DMP*, *Science 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.

✔️ RandomDB

#### 1.b.1.b.1.a.3. Name and contact details of the owner of this data

❗ *This question has not been answered yet!*

#### 1.b.1.b.1.a.4. What are the conditions of use for this database or dataset?

🏷️ **Tags:** *Horizon 2020 DMP*, *Science Europe DMP*

🔗 **External Links:** [*Wikipedia on Copyright*](https://en.wikipedia.org/wiki/Copyright)

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.

✔️ b. They are freely available with obligation to quote the source (e.g. CC-BY)

# 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 indication

|  |  |  |
| --- | --- | --- |
| Answered (current phase) | 10 / 12 | 83.33 % |
| Answered | 10 / 13 | 76.92 % |

#### Metrics

| Metric | Score |
| --- | --- |
| Reusability | 0.71 |
| Good DMP Practice | 1.00 |

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

🏷️ **Tags:** *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?

🏷️ **Tags:** *Science Europe DMP*

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

#### 2. Data storage systems and file naming conventions

🏷️ **Tags:** *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.

❗ *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?

❗ *This question has not been answered yet!*

#### 5. How will you validate the integrity of the results?

🏷️ **Tags:** *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?

🏷️ **Tags:** *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?

🏷️ **Tags:** *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.

✔️ a. No

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

🏷️ **Tags:** *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.

✔️ b. Yes

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

🏷️ **Tags:** *Horizon 2020 DMP*

📖 **Data Stewardship for Open Science:** [*egv*](http://localhost:8080/book-references/egv)

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

✔️ 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?

🏷️ **Tags:** *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?

✔️ 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 indication

|  |  |  |
| --- | --- | --- |
| Answered (current phase) | 8 / 9 | 88.89 % |
| Answered | 8 / 9 | 88.89 % |

#### Metrics

| Metric | Score |
| --- | --- |
| Interoperability | 0.50 |

#### 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.

❗ *This question has not been answered yet!*

#### 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?

✔️ a. No

#### 5. Will there be potential issues with statistical normalization?

✔️ b. Yes

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

✔️ a. No

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

✔️ a. No

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

✔️ b. Yes

#### 9. Will you be doing (automated) knowledge discovery?

📖 **Data Stewardship for Open Science:** [*bzu*](http://localhost:8080/book-references/bzu)

✔️ a. No

# VI. Preserving data

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

### Summary

#### Answered indication

|  |  |  |
| --- | --- | --- |
| Answered (current phase) | 13 / 15 | 86.67 % |
| Answered | 17 / 21 | 80.95 % |

#### Metrics

| Metric | Score |
| --- | --- |
| Findability | 0.50 |
| Accessibility | 0.00 |
| Reusability | 1.00 |
| Good DMP Practice | 0.67 |

#### 1. Specify a list of data sets you will be producing

🏷️ **Tags:** *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.

❗ *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?

🏷️ **Tags:** *Horizon 2020 DMP*

📖 **Data Stewardship for Open Science:** [*kjp*](http://localhost:8080/book-references/kjp)

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.

✔️ b. Yes

#### 2.b.1. Is the archived data changing over time, needing re-archival?

📖 **Data Stewardship for Open Science:** [*tgk*](http://localhost:8080/book-references/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

**Advice:**

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?

✔️ b. Tape

#### 2.b.3. Will the archive be stored in a remote location, protecting the data against disasters?

🏷️ **Tags:** *Horizon 2020 DMP*

✔️ a. No

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

🏷️ **Tags:** *Horizon 2020 DMP*

✔️ a. No

#### 2.b.5. Will your project require the archives to be available on-line?

📖 **Data Stewardship for Open Science:** [*ybd*](http://localhost:8080/book-references/ybd)

✔️ a. No

#### 2.b.6. Has it been established who has access to the archive, and how fast?

❗ *This question has not been answered yet!*

#### 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*](http://localhost:8080/book-references/kdp)

❗ *This question has not been answered yet!*

#### 2.b.8. Will the data still be understandable and reusable after a long time?

🏷️ **Tags:** *Horizon 2020 DMP*

📖 **Data Stewardship for Open Science:** [*zmu*](http://localhost:8080/book-references/zmu)

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.

✔️ b. Yes

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

🏷️ **Tags:** *Horizon 2020 DMP*

📖 **Data Stewardship for Open Science:** [*fxe*](http://localhost:8080/book-references/fxe)

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.

✔️ a. No

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

🏷️ **Tags:** *Horizon 2020 DMP*, *Science Europe DMP*

✔️ b. Yes

#### 4.b.1. How will you be paying for these services?

🏷️ **Tags:** *Horizon 2020 DMP*, *Science Europe DMP*

✔️ b. These costs will be carried by (one of) the departments 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?

🏷️ **Tags:** *Horizon 2020 DMP*, *Science Europe DMP*

✔️ b. Yes

#### 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.

✔️ a. No

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

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

#### 9. 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!*

#### 10. Will reference data be created?

📖 **Data Stewardship for Open Science:** [*rbz*](http://localhost:8080/book-references/rbz)

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?

✔️ 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 indication

|  |  |  |
| --- | --- | --- |
| Answered (current phase) | 7 / 7 | 100.00 % |
| Answered | 7 / 7 | 100.00 % |

#### Metrics

| Metric | Score |
| --- | --- |
| Openness | 0.66 |

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

🏷️ **Tags:** *Horizon 2020 DMP*, *Science Europe DMP*

📖 **Data Stewardship for Open Science:** [*jvm*](http://localhost:8080/book-references/jvm)

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.

✔️ b. Yes

#### 2. Can all of your data become completely open immediately?

🏷️ **Tags:** *Horizon 2020 DMP*, *maDMP*, *Science Europe DMP*

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

✔️ a. No

#### 2.a.1. Are there legal reasons why (some of your) data can not be completely open?

🏷️ **Tags:** *Horizon 2020 DMP*, *maDMP*, *Science Europe DMP*

✔️ a. No

#### 2.a.2. Are there business reasons why (some of your) data can not be completely open?

🏷️ **Tags:** *Horizon 2020 DMP*, *Science Europe DMP*

✔️ b. Yes, patents still need to be applied for

#### 2.a.3. Are there other reasons why (some of your) data can not be completely open?

🏷️ **Tags:** *Horizon 2020 DMP*, *Science Europe DMP*

✔️ b. Yes, papers need to be submitted first

#### 2.a.4. Will you use a limited embargo?

🏷️ **Tags:** *Horizon 2020 DMP*, *Science Europe DMP*

✔️ c. Yes, data that is not legally restrained will be released after a fixed time period, unconditionally

#### 2.a.4.c.1. What embargo period are you using?

🏷️ **Tags:** *Horizon 2020 DMP*, *Science Europe DMP*

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

✔️ 10 days