NOR-OpenScreen Gap Analysis

Core DS Knowledge Model

Organization	ELIXIR Norway
Created by	Alexandra Gade (alexga@uio.no)
Based on	RI gap analysis, 0.0.1 (elixir.no:ri-elixir-norway:0.0.1)
Project Phase	Before Finishing the Project
Created at	20 Jan 2021

I. Administrative details

Report

Indications

Answered (current phase)	29 / 29
Answered	44 / 59

Metrics

No metrics for this chapter.

Questions

1 **Contributors**

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

Tags: *maDMP, Science Europe DMP*

Answers

1.b.1 **Name**

Tags: *maDMP, Science Europe DMP*

✔ *Alex Gade*

1.b.2 **E-mail address**

Tags: *maDMP, Science Europe DMP*

✔ *alexandra.gade@ncmm.uio.no*

1.b.3 **ORCID Identifier**

Tags: *maDMP, Science Europe DMP*

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

1.b.4 **Affiliation**

Tags: *Science Europe DMP*

✔ *NCMM*

1.b.5 **Role**

Roles in a project should be given as they are defined by datacite.

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

Tags: *maDMP, Science Europe DMP*

✔ *d. Data Manager*

1.c.1 **Name**

Tags: *maDMP, Science Europe DMP*

✔ *Johannes Landskron*

1.c.2 **E-mail address**

Tags: *maDMP, Science Europe DMP*

✔ *johannes.landskron@ncmm.uio.no*

1.c.3 **ORCID Identifier**

Tags: *maDMP, Science Europe DMP*

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

1.c.4 **Affiliation**

Tags: *Science Europe DMP*

✔ *NCMM*

1.c.5 **Role**

Roles in a project should be given as they are defined by datacite.

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

Tags: *maDMP, Science Europe DMP*

✔ *p. Work Package Leader*

1.d.1 **Name**

Tags: *maDMP, Science Europe DMP*

✔ *Ruth Brenk*

1.d.2 **E-mail address**

Tags: *maDMP, Science Europe DMP*

This question has not been answered yet!

1.d.3 ORCID Identifier

Tags: *maDMP, Science Europe DMP*

This question has not been answered yet!

1.d.4 Affiliation

Tags: *Science Europe DMP*

UiB

1.d.5 Role

Roles in a project should be given as they are defined by datacite.

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

Tags: *maDMP, Science Europe DMP*

p. Work Package Leader

1.e.1 Name

Tags: *maDMP, Science Europe DMP*

Torkild Visnes

1.e.2 E-mail address

Tags: *maDMP, Science Europe DMP*

This question has not been answered yet!

1.e.3 ORCID Identifier

Tags: *maDMP, Science Europe DMP*

This question has not been answered yet!

1.e.4 Affiliation

Tags: *Science Europe DMP*

SINTEF

1.e.5 Role

Roles in a project should be given as they are defined by datacite.

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

Tags: *maDMP, Science Europe DMP*

✔ *p. Work Package Leader*

1.f.1 **Name**

Tags: *maDMP, Science Europe DMP*

✔ *Jeanette Andersen*

1.f.2 **E-mail address**

Tags: *maDMP, Science Europe DMP*

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

1.f.3 **ORCID Identifier**

Tags: *maDMP, Science Europe DMP*

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

1.f.4 **Affiliation**

Tags: *Science Europe DMP*

✔ *UiT*

1.f.5 **Role**

Roles in a project should be given as they are defined by datacite.

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

Tags: *maDMP, Science Europe DMP*

✔ *p. Work Package Leader*

1.g.1 **Name**

Tags: *maDMP, Science Europe DMP*

✔ *Aurora Martinez*

1.g.2 **E-mail address**

Tags: *maDMP, Science Europe DMP*

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

1.g.3 **ORCID Identifier**

Tags: *maDMP, Science Europe DMP*

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

1.g.4 **Affiliation**

Tags: *Science Europe DMP*

✔ *UiB*

1.g.5 **Role**

Roles in a project should be given as they are defined by datacite.

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

Tags: *maDMP, Science Europe DMP*

✔ *p. Work Package Leader*

1.h.1 **Name**

Tags: *maDMP, Science Europe DMP*

✔ *Espen Hansen*

1.h.2 **E-mail address**

Tags: *maDMP, Science Europe DMP*

✔ *espen.hansen@uit.no*

1.h.3 **ORCID Identifier**

Tags: *maDMP, Science Europe DMP*

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

1.h.4 **Affiliation**

Tags: *Science Europe DMP*

✔ *UiT*

1.h.5 **Role**

Roles in a project should be given as they are defined by datacite.

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

Tags: *maDMP, Science Europe DMP*

✔ *p. Work Package Leader*

1.i.1 **Name**

Tags: *maDMP, Science Europe DMP*

✔ *Geir Klinkenberg*

1.i.2 **E-mail address**

Tags: *maDMP, Science Europe DMP*

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

1.i.3 **ORCID Identifier**

Tags: *maDMP, Science Europe DMP*

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

1.i.4 **Affiliation**

Tags: *Science Europe DMP*

✓ *SINTEF*

1.i.5 **Role**

Roles in a project should be given as they are defined by datacite.

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

Tags: *maDMP, Science Europe DMP*

✓ *p. Work Package Leader*

1.j.1 **Name**

Tags: *maDMP, Science Europe DMP*

✓ *Janna Saarela*

1.j.2 **E-mail address**

Tags: *maDMP, Science Europe DMP*

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

1.j.3 **ORCID Identifier**

Tags: *maDMP, Science Europe DMP*

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

1.j.4 **Affiliation**

Tags: *Science Europe DMP*

✓ *NCMM*

1.j.5 **Role**

Roles in a project should be given as they are defined by datacite.

You should specify at least one "Contact Person". If your project has a work package for data management,

identify the leader of that work package as "Data Curator".

Tags: *maDMP, Science Europe DMP*

✔ *p. Work Package Leader*

2 **RI**

Add each of the 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.

Tags: *maDMP, Science Europe DMP*

Answers

2.b.1 **RI name**

Tags: *maDMP, Science Europe DMP*

✔ *NOR-OpenScreen*

2.b.2 **Project short discription**

Tags: *maDMP, Science Europe DMP*

✔ *NOR-OPENSCREEN II aims to continue to provide: - National availability to cutting-edge technology + expertise in CB and marine bioprospecting - Open access to users from academia + SMEs - Competence building and advanced training - A Norwegian node of the EU-OPENSCREEN ERIC - Identification of bioactive compounds is expensive and highly demanding and coordination at a national and European level is necessary*

2.b.3 **Date the RI will started**

Tags: *Science Europe DMP, maDMP*

✔ *01.09.2021*

2.b.4 **Date the RI funding will end**

Tags: *Science Europe DMP, maDMP*

✔ *31.08.2026*

2.b.5 **Funding**

Add all the funding that are part of this project.

Tags: *maDMP, Science Europe DMP*

Answers

2.b.5.b.1 **Funder**

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

Tags: *Science Europe DMP, maDMP*

✔ *Norges Forskningsråd*

Crossref http://dx.doi.org/10.13039/501100005416

2.b.5.b.2 **Funding status**

Tags: *Science Europe DMP, maDMP*

✔ *a. Planned*

2.b.5.b.3 **Grant number**

Tags: *maDMP, Science Europe DMP*

✔

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

Tags: *Science Europe DMP*

✔ *b. Yes, we will be training existing staff*

3.b.1 **What kind of training?**

Tags: *Science Europe DMP*

✔ *Training within BioMedData to better assist customers in their data management (improved internal data stewardship)*

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

✔ *b. Yes*

4.b.1 **What specific hard/software do you need, and why?**

Tags: *Science Europe DMP*

✔ *Centralized data repository for entire infrastructure*

II. Re-using data

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

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

Report

Indications

Answered (current phase)	24 / 24
Answered	34 / 36

Metrics

No metrics for this chapter.

Questions

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

✔ *Researchers interested in chemical biology tools*

2 **Is there pre-existing data?**

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

Tags: *maDMP, Science Europe DMP*
Data Stewardship for Open Science: *atq*

✔ *b. Yes*

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

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

Tags: *maDMP, Science Europe DMP*
Data Stewardship for Open Science: *ezi*

✔ *b. Yes*

2.b.1.b.1 **What reference data will you use?**

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 stanndard (in bioinformatics, a genome is often compared with a reference genome to identify genomic variants). If you use reference data, there are several other issues that you should consider. What are the reference data sets that you will use?

Tags: *Science Europe DMP*
Data Stewardship for Open Science: *quc*

Answers

2.b.1.b.1.b.1 **Source of reference data set**

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

Tags: *Science Europe DMP*

✔ *PubChem*

FAIRsharing https://fairsharing.org/bsg-d000455

2.b.1.b.1.b.2 **What are the conditions of use for this data sets?**

Tags: *Science Europe DMP*

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

2.b.1.b.1.b.3 **Do you know in what format the reference data is available?**

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*

✔ *b. I need to convert it before using*

2.b.1.b.1.b.4 **Is the reference data resource versioned?**

Many reference data sets evolve over time. If the reference data set changes, this may affect your

results. If different versions of a reference data set exist, you need to establish your "version policy".

Tags: *Science Europe DMP*
Data Stewardship for Open Science: *rgy*

✔ *a. No*

2.b.1.b.1.b.5 **How will you make sure the same reference data will be available to reproduce results of your users?**

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

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

2.b.1.b.1.c.1 **Source of reference data set**

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

Tags: *Science Europe DMP*

✔ *ChEMBL*

FAIRsharing https://fairsharing.org/bsg-d000015

2.b.1.b.1.c.2 **What are the conditions of use for this data sets?**

Tags: *Science Europe DMP*

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

2.b.1.b.1.c.3 **Do you know in what format the reference data is available?**

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*

✔ *b. I need to convert it before using*

2.b.1.b.1.c.4 **Is the reference data resource versioned?**

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

Tags: *Science Europe DMP*
Data Stewardship for Open Science: *rgy*

✔ *a. No*

2.b.1.b.1.c.5 **How will you make sure the same reference data will be available to reproduce results of your users?**

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

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

2.b.1.b.2 **Will you use non-reference data sets?**

Tags: *Science Europe DMP*

✔ *a. Yes*

2.b.1.b.2.a.1 **What existing non-reference data sets will you use?**

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?

Tags: *maDMP, Science Europe DMP*
Data Stewardship for Open Science: *wya*

Answers

2.b.1.b.2.a.1.b.1 **Non-reference data set**

Give a name for the data set. You will get suggestions listing databases listed in FAIRSharing, but you can also specify data sets that are not listed.

Tags: *Science Europe DMP*

✔ *PubChem*

FAIRsharing https://fairsharing.org/bsg-d000455

2.b.1.b.2.a.1.b.2 **Where is this data set available**

Specify a URL or a persistent identifier (e.g. DOI) for the data set.

Tags: *Science Europe DMP*

✔ *various pubchem assay data*

2.b.1.b.2.a.1.b.3 **What are the conditions of use for this data set?**

Tags: *Science Europe DMP*

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

2.b.1.b.2.a.1.b.4 **Will the owners of this data set work with you on this study?**

Data Stewardship for Open Science: *dcy*

✔ *a. No*

2.b.1.b.2.a.1.b.4.a.1 **Do you need to request access to the data**

✔ *a. No*

2.b.1.b.2.a.1.b.5 **Is extension of any consent for privacy sensitive data needed?**

If the data that you will re-use is coupled to people, the informed consent that was originally obtained from those people may not be covering your current research. In that case re-consent may be necessary.

Tags: *maDMP, Science Europe DMP*
Data Stewardship for Open Science: *bqy*
External Links: *Legal bases for personal data processing under GDPR*

✔ *a. Not applicable: none of the data is personal data*

2.b.1.b.2.a.1.b.6 **Can you use the data in a format that is available?**

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

✔ *b. I need to convert it before using*

2.b.1.b.2.a.1.b.7 **How will you be accessing the data?**

Tags: *Science Europe DMP*

✔ *b. Will download or get a copy*

2.b.1.b.2.a.1.b.8 **Is the data set fixed, or will it be updated in the future?**

Is the data set you will reuse a fixed data set (with a persistent identifier), or is it a data set that is being worked on (by others) and may be updated during your project or after?

✔ *a. It is a fixed data set, this will not influence reproducibility of my results*

2.b.1.b.2.a.1.b.9 **Can you and will you use the complete existing data set?**

If you use any filtering, how will you make sure that your results will be reproducible by yourself and others at a later time?

✔ *c. I will make sure the selected subset will be available together with my results*

2.b.1.b.3 **Will you couple existing (biobank) data sets?**

✔ *b. Yes*

2.b.1.b.3.b.1 **Will you use deterministic couplings?**

Data sets that have exactly identical fields that are well filled can be coupled using deterministic methods. Will you be using such methods?

✔ *a. No*

2.b.1.b.3.b.2 **Will you be using a trusted third party for coupling?**

What will be the procedure that is followed? Where will what data be sent? Did a legal advisor look at the procedures?

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

2.b.1.b.3.b.3 **Is consent available for the couplings?**

✔ *a. No*

2.b.1.b.3.b.4 **How will you check whether your coupled data are representative of your goal population?**

Sometimes, through the nature of the data sets that are coupled, the coupled set is no longer representative for the whole population (e.g. some fields may only have been filled for people with high blood pressure). This can disturb your research if undetected. How will you make sure this is not happening?

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

2.b.1.b.3.b.5 **What is the goal of the coupling?**

✔ *a. More data about the same subjects (intersection)*

2.b.1.b.3.b.6 **What variable(s) will you be using for the coupling?**

✔ *Assay type, cell line*

2.b.1.b.3.b.7 **Will you use probabilistic couplings?**

Data sets that have similar but not identical fields or with identical fields that are not consistently filled can be coupled using probabilistic methods. Will you be using such methods?

✔ *a. No*

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

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

Data Stewardship for Open Science: *wht*

✔ *a. No*

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

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

Data Stewardship for Open Science: *pth*

✔ *a. No*

III. Creating and collecting data

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

Report

Indications

Answered (current phase)	51 / 51
Answered	54 / 55

Metrics

Metric	Score
Findability	0
Accessibility	0
Interoperability	1
Reusability	0.36
Good DMP Practice	0

Questions

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

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

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

Tags: *Science Europe DMP*

Data Stewardship for Open Science: *njy*

Answers

1.b.1 **Data format/type**

Tags: *Science Europe DMP*

✔ *Tabular Data Package*

FAIRsharing https://fairsharing.org/bsg-s001302

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

Tags: *Science Europe DMP*

✔ *b. Yes*

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

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

Tags: *Science Europe DMP*

✔ *b. Yes*

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

Tags: *Science Europe DMP*

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

1.c.1 **Data format/type**

Tags: *Science Europe DMP*

✔ *Tagged Image File Format*

FAIRsharing https://fairsharing.org/bsg-s000268

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

Tags: *Science Europe DMP*

✔ *b. Yes*

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

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

Tags: *Science Europe DMP*

✔ *b. Yes*

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

Tags: *Science Europe DMP*

✔ *b. I can specify the total amount per year*

1.c.4.b.1 **Data volume in Gigabytes per year**

Specify an approximation of the expected data volume

Tags: *Science Europe DMP*

✔ *1000*

1.d.1 **Data format/type**

Tags: *Science Europe DMP*

✔ *Portable Network Graphics*

FAIRsharing https://fairsharing.org/bsg-s000206

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

Tags: *Science Europe DMP*

✔ *b. Yes*

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

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

Tags: *Science Europe DMP*

✔ *b. Yes*

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

Tags: *Science Europe DMP*

✔ *b. I can specify the total amount per year*

1.d.4.b.1 **Data volume in Gigabytes per year**

Specify an approximation of the expected data volume

Tags: *Science Europe DMP*

✔ *1000*

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

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

Data Stewardship for Open Science: *ikk*

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

3 **How will you/your users be storing metadata?**

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

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

Data Stewardship for Open Science: *rhm*
External Links: *SEEK*

✔ *a. Explore*

> 3.a.1 **Do suitable 'Minimal Metadata About ...' (MIA...) standards exist for your experiments?**
>
> External Links: *FAIRsharing repository of standards*
>
> ✔ *a. No*
>
> *Did you really check a service like fairsharing.org to verify this?*
>
> > 3.a.1.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?**
> >
> > External Links: *FAIRsharing repository of standards*
> >
> > ✔ *b. Yes*
>
> 3.a.2 **Do you know how and when you will be collecting the necessary metadata?**
>
> Often it is easiest to make sure you collect the metadata as early as possible.
>
> External Links: *FAIRsharing repository of standards*
>
> ✔ *a. No*
>
> 3.a.3 **Will you consider re-usability of your data beyond your original purpose?**
>
> Adding more than the strict minimum metadata about your experiment will possibly allow more wide re-use of your data, with associated higher data citation rates. Please note that it is not easy for yourself to see all other ways in which others could be reusing your data.
>
> ✔ *a. No, I will just document the bare minimum*
>
> 3.a.4 **Did you consider how to monitor data integrity?**
>
> Working with large amounts of heterogenous data in a larger research group has implications for the data integrity. How do you make sure every step of the workflow is done with the right version of the data? How do you handle the situation when a mistake is uncovered? Will you be able to redo the strict minimum data handling?
>
> Data Stewardship for Open Science: *spg*
>
> ✔ *a. Explore*
>
> > 3.a.4.a.1 **Will you be keeping a master list with checksums of certified/correct/canonical/verified**

data?

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

✔ *a. No*

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

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

✔ *a. No*

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

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

✔ *a. No*

3.a.6 **How will you keep provenance?**

To make your experiments reproducible, all steps in the data processing must be documented in detail. The software you used, including version number, all options and parameters. This information together for every step of the analysis is part of the so-called data provenance. There are more questions regarding this in the chapter on data processing and curation.

✔ *a. All steps will be documented in an (electronic) lab notebook*

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

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

✔ *a. Explore*

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

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

✔ *a. No*

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

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

✔ *a. No*

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

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

✔ *a. No, the file names in the project are an essential part of the metadata*

4 **Please specify what data you will acquire using measurement equipment**

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

Tags: *Science Europe DMP*

Answers

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

Tags: *Science Europe DMP*

✔ *c. Experts in the RI, at a our infrastructure*

4.b.2 **Instruments used for data collection**

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

Tags: *Science Europe DMP*

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

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

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

Tags: *Science Europe DMP*

✔ *a. Very well described and known*

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

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

✔ *b. Yes, lets explore this*

4.b.4.b.1 **Is the data format established?**

Has the storage and transport format of the primary data been established between the people responsible for the measurement and the people responsible for the processing?

✔ *a. No*

4.b.4.b.2 **How will the raw data be transported?**

✔ *c. Via the network*

4.b.4.b.2.c.1 **Is sufficient network capacity available?**

Can the volume of data be accommodated by the standard network connection? Has a special network connection (e.g. light path) that is needed been reserved?

✔ *a. Yes, has been taken care of*

4.b.4.b.3 **Is data integrity guaranteed during this stage?**

Do you have any means of identifying whether the raw data has been transferred error free and has not been tampered with?

✔ *a. No*

4.b.4.b.4 **Is data security guaranteed during this stage?**

Are the raw data encrypted or otherwise protected from theft or leaks at either site or during transport? You could e.g. use a light path or a virtual private network if you transport the data over the net.

✔ *b. Yes*

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

Tags: *Science Europe DMP*

✔ *b. Yes*

4.b.5.b.1 **Are you calibrating measurements?**

Tags: *Science Europe DMP*

✔ *b. Yes*

4.b.5.b.2 **Are you running repeat samples or are you repeating measurements?**

Tags: *Science Europe DMP*

✔ *b. Yes*

4.b.5.b.3 **Are you running standardized data capture or recording?**

Tags: *Science Europe DMP*

✔ *b. Yes*

4.b.5.b.4 **Are you doing Data Entry validation?**

Tags: *Science Europe DMP*

✔ *a. No*

4.b.5.b.5 **Are you using data peer review?**

Tags: *Science Europe DMP*

✔ *a. No*

4.b.5.b.6 **Are you using controlled vocabularies?**

Tags: *Science Europe DMP*

✔ *a. No*

4.b.5.b.7 **Are you using any other quality processes?**

Tags: *Science Europe DMP*

✔ *b. Yes*

4.b.5.b.7.b.1 **What other quality processes do you use?**

✔ *identity and purity testing of compounds via MS; assessment and comparison of assay results*

5 **Do you have any non-equipment data capture?**

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

Tags: *Science Europe DMP*
Data Stewardship for Open Science: *ybw*

✔ *a. No*

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

✔ *b. Yes*

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

✔ *d. Software not listed here*

7 **Will you be storing physical samples?**

Data Stewardship for Open Science: *kuz*

✔ *b. Yes*

You might want to contact Biobank Norway for advice

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

Tags: *maDMP, Science Europe DMP*
External Links: *NSD Information and consent, REC Informed consent*

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

9 **How is the ownership of the collected data arranged?**

Tags: *Science Europe DMP*

✔ *b. All data will be owned by the Principle Investigator/user*

IV. Data sensitivity

Ethical and legal issues

adapted from 2019 version of NSD DMP tool and Tryggve Checklist on ELSI issues and GDPR compliance

Report

Indications

Answered (current phase)	22 / 22
Answered	25 / 26

Metrics

Metric	Score

Metric	Score
Good DMP Practice	0

Questions

1 **Will you collect or generate data about people?**

✔ *a. Yes*

1.a.1 **Will you collect and/or process personally identifiable data?**

What is personally identifiable data?

Personal data is any information that can be connected to a person e.g. name, address, phone number, e-mail address, IP-address, car registration number, images, fingerprints, iris patterns, head shape (for facial recognition) and birth number, or through a combination of background information. Information about behavioral patterns may also be considered as personal data.

Sensitive personal data is information relating to racial or ethnic origin, political, philosophical or religious beliefs, that a person has been suspected, charged or convicted of a crime, health, sex life, and union membership.

Read more about personal and sensitive data at the Data Inspectorate and at the NSD - Data Protection Services.

Sensitive data has to be stored and analysed using appropriate measures and infrastructure. (such as TSD) - You can apply for quotas through: contact@bioinfo.no

✔ *a. Yes*

1.a.1.a.1 **Please specify**

Answers

1.a.1.a.1.b.1 **Give a short description of the data and the contained personal information**

✔ *Readouts from optical data on patient samples indicating the samples responsiveness to different compounds*

1.a.1.a.1.b.2 **Have you been in contact with a data protection official or other inspectorates?**

If your are to archive personally identifiable data, you must document that you are allowed to archive the data in accordance with current regulations. If you do not have permission to store personally identifiable data, data must be anonymised so that anonymous version can be archived before the original data is deleted.

Consider using EGA or a local branch of it.

✔ *b. No, but I will*

1.a.1.a.1.b.3 **Where are you going to process active personally identifiable research data?**

ELIXIR Norway provides sustainable free storage to the Norwegian life sciences community in collaboration with Sigma2 on TSD upon application.

External Links: *TSD, HUNTCloud, SAFE, EUTRO*

✔ *e. other*

1.a.1.a.1.b.3.e.1 **Where?**

✔ *Currently locally archived*

1.a.1.a.1.b.4 **Do you have a consent form the research subjects already?**

The content of informed consents needed to be valid under different laws (e.g. GDPR, the Health

Research Act or other ethical legislation) might differ.

External Links: *GDPR consent*, *The Health Research Act - Participant consent*

✔ *b. No, but I will*

1.a.1.a.1.b.5 **Who will be the data controllers of the personal data processed in the dataset?**

External Links: *See also*

Answers

1.a.1.a.1.b.5.b.1 **Data controller**

GDPR Article 4 (7): "'controller' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; […]." For cross-border collaborative projects the controllers of different datasets should be identified. Also, if joint controllership is considered, make sure that all parties understand their obligations, and it is probably good to define the terms for this in an agreement between the parties.

External Links: *Art. 4 - GDPR Definitions*

Answers

1.a.1.a.1.b.5.b.1.b.1 **List the data controller(s)**

✔ *Up to the customer*

1.a.1.a.1.b.5.b.1.b.2 **Is there more than one controller of the personal data in the dataset?**

✔ *c. Needs to be investigated*

1.a.1.a.1.b.6 **What is the legal basis for processing the personal data?**

GDPR - Article 6 (1) lists under what conditions the processing is considered lawful. Of these, *Consent* or *Public interest* are relevant when it comes to research. You should determine what legal basis (or bases) you have for processing the personal data in your project. Traditionally, consent has been the basis for processing personal data for research, but under the GDPR there cannot be an imbalance between the processor and the data subject for it to be considered to be freely given. In some countries the use of consent as the legal basis for processing by universities for research purposes is therefore not recommended. In those cases, public interest should probably be your legal basis. Note that if your legal basis for processing is consent, a number of requirements exists for the consent to be considered valid under the GDPR. Consents given before the GDPR might not live up to this. Also note that even if public interest is the legal basis, other laws and research ethics standards might still require you to have consent from the subjects for performing the research. Please consult with the Data Protection Officer of your organisation on which legal basis to apply to your data.

External Links: *Art. 6 GDPR - Lawfulness of processing*

✔ *b. Consent*

1.a.1.a.1.b.6.b.1 **Are consents in compliance with the GDPR?**

✔ *b. Needs to be investigated*

1.a.1.a.1.b.7 **What are the exemptions for the prohibition for processing of special categories of data (such as health and genetic data)?**

Processing of certain categories of personal data is not allowed unless there are exemptions in law to allow this. Among these categories ("sensitive data") are "'[…] data revealing racial or ethnic origin, […] genetic data, […] data concerning health'". Most types of personal data collected in biomedical research

will fall under these categories. Article 9 (2) lists a number of exemptions that apply, of which consent and scientific research are most likely to be relevant for research. Please consult with your Data Protection Officer of your organisation.

External Links: *Art. 9 GDPR - Processing of special categories of personal data*

✔ *c. Needs to be investigated*

1.a.1.a.1.b.8 **Have data processing agreements been established between the data controller(s) and any data processors?**

A data processing agreement has to contain the obligatory clauses specified in Art 28.3 of the GDPR. The agreement should also regulate the use of any sub-processors.

External Links: *Art. 28 - GDPR Processor*

✔ *b. Needs to be investigated*

1.a.1.a.1.b.9 **Have Data Protection Impact Assessments (DPIA) been performed for the personal data?**

List DPIAs done and for which parts of the data. *Note:* All Nordic Data Protection Authorities have identified that most types of research projects on health or genetic data require a DPIA.

Where a type of processing is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data, a so called Data Protection Impact Assessment (DPIA) - Article 35. To clarify when this is necessary, the Data Protection Authorities (DPAs) in Denmark, Finland, Norway and Sweden have issued guidance of when an impact assessment is required. Large-scale processing of sensitive data such as genetic or other health related data is listed by all DPAs as requiring DPIAs. The French DPA has made a PIA tool (endorsed by several other DPAs) available that can help in performing these impact assessments. Please also consult your Data Protection Officer of your organisation.

External Links: *Art. 35 GDPR - Data protection impact assessment , Datatilsynet: When must an impact assessment be carried out?, The open source PIA software helps to carry out data protection impact assesment*

✔ *b. Needs to be investigated*

1.a.1.a.1.b.10 **What technical and procedural safeguards have been established for processing the data?**

To ensure that the personal data that you process in the project is protected at an appropriate level, you should apply technical and procedural safeguards to ensure that the rights of the data subjects are not violated. Examples of such measures include, but are not limited to, pseudonymisation end encryption of data, the use of computing and storage environments with heightened security, and clear and documented procedures for project members to follow.

✔ *Needs to be investigated*

1.a.1.a.1.b.11 **What happens with the dataset after project completion?**

The GDPR states that the processing (including storing) of personal data should stop when the intended purpose of the processing is done. There are, however, exemptions to this e.g. when the processing is done for research purposes. Also, from a research ethics point of view, research data should be kept to make it possible for others to validate published research findings and reuse data for new discoveries. This is also governed by what the data subjects have been informed about regarding how you will treat the data after project completion. The recommendation is to deposit the sensitive data in the appropriate controlled access repositories if such are available, but this requires that the data subjects are informed and have agreed to this. Other considerations

✔ *b. The dataset will be archived at the controller(s)*

1.a.1.a.1.b.11.b.1 **In what form will the data sets be stored?**

✔ *b. Pseudonymised*

1.a.1.a.1.b.12 **Are there other relevant national legislation considerations that has to be taken into account?**

External Links: *Lov om helseregistre og behandling av helseopplysninger (helseregisterloven), Lov om medisinsk og helsefaglig forskning (helseforskningsloven), Forskrift om organisering av medisinsk og helsefaglig forskning, Merknader til forskrifter til Helseforskningsloven, Veileder til Helseforskningsloven, Forskrift om befolkningsbaserte helseundersøkelser, Lov om helsepersonell mv (helsepersonelloven), Lov om pasient- og brukerrettigheter (pasient- og brukerrettighetsloven), Lov om legemidler mv (legemiddelloven), Forskrift om klinisk utprøving av legemidler til mennesker, Norm for helsedata, Lov om humanmedisinsk bruk av bioteknologi mm (bioteknologiloven), Lov om arkiv [arkivlova], Lov om organisering av forskningsetisk arbeid (forskningsetikkloven)*

✔ *c. Needs to be investigated*

1.a.1.a.1.b.13 **Are there other Terms & conditions for data access (in particular if presenting obstacles for cross-border processing of health data)?**

E.g. register data access policies (requirement of PI in the same country, moving data to other secure services)

✔ *b. Needs to be investigated*

1.a.2 **Other comments regarding processing of personal data**

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

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

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

✔ *a. Yes*

2.a.1 **Provide names and links below.**

✔ *Whatever is dictated by the customers; also following relevant ethical guidelines to each of node's host institutions*

3 **Other ethical / legal issues.**

✔ *Some data relating to patients for personalized medicine have to stay in Norway*

V. Processing data

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

Report

Indications

Answered (current phase)	86 / 86

Answered	100 / 100

Metrics

Metric	Score
Accessibility	0.5
Reusability	0.18
Good DMP Practice	0.35

Questions

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

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

Tags: *Science Europe DMP*

✔ *a. No*

> 1.a.1 **Are data that users store themselves adequately backed up and traceable?**
>
> Tags: *Science Europe DMP*
>
> ✔ *a. No*

2 **Data storage systems and file naming conventions**

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

Tags: *Science Europe DMP*

✔ *a. Explore*

> 2.a.1 **Are you using a filesystem with files and folders?**
>
> Are some of the data in the project stored in a filesystem with files and folders?
>
> Tags: *Science Europe DMP*
>
> ✔ *b. Yes*
>
> > 2.a.1.b.1 **Will you use a folder for each sample/subject?**
> >
> > Tags: *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.
> > >
> > > ✔ *Project #/Experiment Type/Date -- not standardized across RI because RI does not have centralized policies/data (yet)*
>
> > 2.a.1.b.2 **Will you use a (sub)folder for each (repeated) analysis?**
> >
> > Tags: *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.

✔ *date*

2.a.1.b.3 **Will you use a (sub)folder for each step in the analysis workflow?**

Tags: *Science Europe DMP*

✔ *a. No*

2.a.1.b.4 **What appointments have you made about the naming of files?**

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.

Tags: *Science Europe DMP*

✔ *Project dependent!*

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

Tags: *Science Europe DMP*

✔ *a. No*

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

Tags: *Science Europe DMP*

✔ *b. Yes*

2.a.3.b.1 **How will you handle changes in the data?**

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?

Tags: *Science Europe DMP*

✔ *b. We will be allowing Create, Update and Delete*

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

Tags: *Science Europe DMP*

✔ *a. No*

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

Tags: *Science Europe DMP*

✔ *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*

NeLS provides access and computing for Galaxy workflows

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*

✔ *c. A bit of both*

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?**

✔ *b. They can start with a subset as soon as our first data comes*

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?**

✔ *Nothing outside of University standard practices*

3.b.3 **List existing software components you will use in the analysis/processing workflows**

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.b.1 **Software component:**

✔ *R*

3.b.3.b.2 **Where are you getting this software from? Please specify a web address if available.**

✔ *CRAN*

3.b.3.b.3 **What version of this software will you use?**

✔ *c. Whatever is the latest version at the time the analysis is run*

3.b.3.b.3.c.1 **Will you re-run any analysis for users when a new version comes out?**

✔ *a. No*

3.b.3.b.4 **How is your experience with this software?**

✔ *a. We know the software well*

3.b.3.c.1 **Software component:**

✔ *KNIME*

3.b.3.c.2 **Where are you getting this software from? Please specify a web address if available.**

✔ *knime.com*

3.b.3.c.3 **What version of this software will you use?**

✔ *c. Whatever is the latest version at the time the analysis is run*

3.b.3.c.3.c.1 **Will you re-run any analysis for users when a new version comes out?**

✔ *a. No*

3.b.3.c.4 **How is your experience with this software?**

✔ *a. We know the software well*

3.b.3.d.1 **Software component:**

✔ *DSF Analyzer*

3.b.3.d.2 **Where are you getting this software from? Please specify a web address if available.**

✔ *Developed in-house*

3.b.3.d.3 **What version of this software will you use?**

✔ *a. One exact version per user data set*

3.b.3.d.4 **How is your experience with this software?**

✔ *b. We know the authors well*

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.b.1 **Software component:**

✔ *Various in house analysis functions*

3.b.4.b.2 **Please specify the software repository you use for development**

Preferably use a direct URL other users could use

✔ *N/A*

3.b.4.b.3 **Did you consider existing options?**

✔ *c. Alternatives exist, but we prefer to develop our own*

3.b.4.b.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.

External Links: *Apache 2.0*, *GPL 3.0*, *AGPL 3.0*, *LGPL 3.0*

✔ *e. We are using a different license*

3.b.4.b.4.e.1 **What other license?**

Preferably give a web pointer to the license.

✔ *Licenses not specified*

3.b.4.b.4.e.2 **Is this a recognised reusable open source license?**

✔ *a. No*

3.b.5 **Did you choose the workflow engine you will be using?**

Data Stewardship for Open Science: *ydj*

✔ *d. Explore*

3.b.5.d.1 **Do you need the workflow engine to produce provenance information automatically?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.1.b.1 **Does the provenance need to be stored or converted to some standard format?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.1.b.2 **Can the workflow be annotated to make it understandable?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.2 **Do you need the workflow engine to be run high-throughput?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.3 **Is ease of development of the workflow engine itself an issue for you?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.3.b.1 **Can you reach out to the developers? Is there a contact?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.3.b.2 **Does the workflow engine need to support a particular compute back end you will use?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *a. No*

3.b.5.d.3.b.3 **Does the workflow engine need standard tools for administrators?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *a. No*

3.b.5.d.4 **Is ease of development of the workflows an issue for you?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.4.b.1 **Does the workflow engine need a developer GUI?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *a. No*

3.b.5.d.4.b.2 **Does it need to be easy to support new tools in a workflow?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.4.b.3 **Does it need support for specific kinds of data processing or data integration pipelines?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.4.b.4 **Does it need support for complex control structures like conditionals and/or loops?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these

questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.4.b.5 **Does it need native support for specific data types?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.4.b.6 **Does it need support for nested workflows?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.4.b.7 **Does it need support collaborative editing of workflows?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *a. No*

3.b.5.d.5 **Is ease of use an issue for you?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.5.b.1 **Who are the people that will run the workflows?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

Data Stewardship for Open Science: *jrw*

✔ *b. Subject matter experts with a strong computer knowledge*

3.b.5.d.5.b.2 **Does the running of workflows need to be controlled via a GUI?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.5.b.3 **Do workflows need to be run on remote computers?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

Data Stewardship for Open Science: *grt*

✔ *a. No*

3.b.5.d.6 **Is sustainability of workflows important for you?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these

questions can lead you to think about the most important features for the project

Data Stewardship for Open Science: *xyf*

✔ *b. Yes*

3.b.5.d.6.b.1 **Do you need the same workflow to run next year? Durability against 'workflow decay'?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.6.b.2 **Are all versions of all tools, including built-in tools, under total control of the project?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *a. No*

3.b.5.d.6.b.3 **Do you need to be able to export/import workflows (e.g. in the Common Workflow Language, CWL)?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *b. Yes*

3.b.5.d.7 **Do you want your workflow engine to be professionally hosted by a specialized party?**

There is no concrete map of these features to suitable workflow engines at this moment, but filling in these questions can lead you to think about the most important features for the project

✔ *a. No*

3.b.5.d.8 **Can you make a decision for the workflow engine based on the criteria deemed important?**

External Links: *snakemake*, *nextflow*, *Common Workflow Language*, *galaxy*, *taverna*, *knime*, *molgenis*, *moteur*, *clcbio*, *chipster*

✔ *f. We will use 'knime'*

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?**

✔ *a. No*

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.

✔ *a. No*

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

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

Data Stewardship for Open Science: *pzq*

✔ *a. No*

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

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

Data Stewardship for Open Science: *pzq*

✔ *a. No*

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

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

✔ *a. No*

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

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

✔ *a. No*

5 **How will you validate the integrity of the results?**

✔ *a. Explore*

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

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

✔ *a. No*

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

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

✔ *a. No*

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

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

✔ *a. No*

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

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

Data Stewardship for Open Science: *egv*

✔ *a. No*

6 **Do you need to do compute capacity planning?**

If you require substantial amounts of compute power, amounts that are not trivially absorbed in what you usually have 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?**

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

Tags: *Science Europe DMP*

✔ *a. Explore*

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

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

Tags: *Science Europe DMP*

✔ *b. Yes*

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

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

Tags: *Science Europe DMP*

✔ *b. Yes*

7.a.2.b.1 **Are all data carriers encrypted? Are accounts on the laptop password protected?**

Tags: *Science Europe DMP*

✔ *a. No*

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

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

✔ *b. Yes*

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

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

✔ *b. Yes*

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

Tags: *Science Europe DMP*

✔ *b. Yes*

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

Tags: *Science Europe DMP*

✔ *b. Yes*

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

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

Tags: *Science Europe DMP*

✔ *a. No*

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

Tags: *Science Europe DMP*

✔ *a. No*

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

Tags: *Science Europe DMP*

✔ *a. No*

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

Tags: *Science Europe DMP*

✔ *a. No*

7.a.11 **Are personal data sufficiently protected?**

Tags: *Science Europe DMP*

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

7.a.11.b.1 **How is pseudonymization handled?**

Tags: *Science Europe DMP*

✔ *b. Pseudonymization is handled by an independent party in order to allow data coupling*

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*

9 **Will you version datasets?**

SEEK which is used in FAIRDOMHub and can be used together with NeLS supports versioning by default.

NeLS can also be used with Git Large File Storage (LFS)

External Links: *FAIRDOMHub, SEEK, NeLS, Git Large File Storage (LFS)*

✔ *b. No*

VI. Interpreting data

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

Report

Indications

Answered (current phase)	27 / 27
Answered	27 / 27

Metrics

Metric	Score
Interoperability	0.62
Reusability	0.5
Good DMP Practice	0

Questions

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

✔ *a. Explore*

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

Answer some questions for each

Answers

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

✔ *Table Schema*

FAIRsharing https://fairsharing.org/bsg-s001301

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

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

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

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

✔ *b. Yes*

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

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

✔ *a. No*

Check whether you can find an ontology for each of your data items

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

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

✔ *b. Yes*

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

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

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

✔ *b. Yes*

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

✔ *b. Yes*

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

✔ *b. Yes*

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

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

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

✔ *a. No*

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

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

✔ *a. No*

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

Data Stewardship for Open Science: *qqb*

✔ *b. Yes*

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

External Links: *Linked data (wikipedia)*

✔ *a. No*

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

✔ *b. Yes*

3 **Will you/your users be using common ontologies?**

✔ *a. No*

4 **Will there be potential issues with statistical normalization?**

✔ *b. Yes*

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

✔ *b. Yes*

5.b.1 **Have these been collected with sufficiently identical protocols?**

✔ *a. No*

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

✔ *b. Yes*

6.b.1 **Did you already select the variables on which you will join the data sets?**

✔ *b. Yes*

6.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.

✔ *a. No*

6.b.3 **Could the coupling of data create a danger of re-identification of anonymized privacy sensitive data?**

✔ *a. No*

6.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*

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

✔ *a. No*

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

Data Stewardship for Open Science: *bzu*

✔ *b. Yes*

VII. Preserving data

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

Report

Indications

Answered (current phase)	24 / 24
Answered	34 / 35

Metrics

Metric	Score
Findability	0
Accessibility	0.15
Reusability	0.12
Good DMP Practice	0

Questions

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

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

Data Stewardship for Open Science: *kjp*

✔ *a. No*

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

Data Stewardship for Open Science: *ixr*

✔ *a. No*

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

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

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

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

Tags: *maDMP, Science Europe DMP*

Answers

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

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

Tags: *maDMP, Science Europe DMP*

✔ *optical readout data*

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

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

Tags: *Science Europe DMP, maDMP*

✔ *raw instrument data*

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

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

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

Tags: *Science Europe DMP, maDMP*

Answers

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

Which type of identifier is this?

✔ *e. Other*

2.b.3.b.1.e.1 **What is the identifier type?**

✔ *Do not have standardized*

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

✔ *Do not have standardized*

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

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

Tags: *maDMP, Science Europe DMP*

✔ *a. No*

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

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

Tags: *Science Europe DMP*

✔ *c. For a fixed period (prepaid)*

2.b.5.c.1 **How long will the data be kept?**

Specify the period, and optionally add how this relates to the minimum period that you are required to keep the data and/or a reason for choosing this period.

Tags: *Science Europe DMP*

✔ *Usually 10 years minimum*

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

This is a one of the FAIR principles.

Tags: *Science Europe DMP*

✔ *a. No*

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

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

Tags: *Science Europe DMP, maDMP*

✔ *a. No*

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

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

Tags: *Science Europe DMP, maDMP*

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

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

✔ *b. No*

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

Tags: *Science Europe DMP*

✔ *a. No*

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

Tags: *Science Europe DMP*

✔ *a. No*

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

✔ *a. No*

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

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

✔ *a. No*

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

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

Data Stewardship for Open Science: *fxe*

✔ *a. No*

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

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

✔ *a. No*

9 **Specify a list of software packages you will be publishing**

Specify a short name for each software package.

Answers

> **9.b.1** **Software package:**
>
> ✔ *DSF Analyzer from Bergen*
>
> **9.b.2** **Will you be adding a proper open-source license?**
>
> External Links: *Choose an open source license*, *Open Source Initiative: Licenses*
>
> ✔ *a. No*
>
> **9.b.3** **Where will the software package be available?**
>
> ✔ *On Demand to Users*

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

✔ *a. No*

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

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

Tags: *Science Europe DMP*

✔ *a. We use lab notebooks*

Make sure to make the notes available in electronic form along with your data

11 **Will reference data be created?**

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

Data Stewardship for Open Science: *rbz*

✔ *a. No*

12 **How will you document your/the user data?**

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

Tags: *Science Europe DMP*

✔ *a. Explore*

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

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

Tags: *Science Europe DMP*
External Links: *Dublin Core Metadata Terms, Dublin Core Initiative*

✔ *a. No*

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

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

Tags: *Science Europe DMP*
External Links: *W3C Prov primer*

✔ *a. No*

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

✔ *a. No*

14 **Will you do structural modeling?**

✔ *a. No*

VIII. Giving access to data

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

Report

Indications

Answered (current phase)	23 / 23
Answered	23 / 23

Metrics

Metric	Score
Accessibility	0
Good DMP Practice	1
Openness	0.2

Questions

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

Tags: *Science Europe DMP*
Data Stewardship for Open Science: *jvm*

✔ *a. No*

You will need to explain!

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

✔ *a. Yes*

> 2.a.1 **How will you manage copyright and Intellectual Property Rights (IPR) issues?**
>
> ✔ *only publish metadata for IPR data*

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

Tags: *maDMP, Science Europe DMP*

✔ *a. No*

> 3.a.1 **Are there legal reasons why (some of your) data can not be completely open?**
>
> Tags: *maDMP, Science Europe DMP*
>
> ✔ *b. Yes*
>
> > 3.a.1.b.1 **Are there privacy reasons why data can not be open?**
> >
> > Tags: *maDMP*
> >
> > ✔ *b. Yes*
> >
> > > 3.a.1.b.1.b.1 **Are there restrictions on where the data need to be stored?**
> > >
> > > Tags: *maDMP*

✔ *c. Yes, they must stay in the same country*

3.a.1.b.1.b.1.c.1 **Are you going to use a national platform for your data?**

External Links: *NSD Archive*, *EGA*, *HUNT DB*

✔ *d. other*

3.a.1.b.1.b.1.c.1.d.1 **Which platfom?**

✔ *local*

3.a.1.b.1.b.2 **Could pseudonymization be used to make the data more openly available?**

Legally, pseudonymous data (which means that someone has the key to reverse the process) is still considered privacy sensitive information. However, the EU is working on special cases where the data can still be opened as long as the key availability is sufficiently limited.

Tags: *maDMP*

✔ *a. No*

3.a.1.b.1.b.3 **Could anonymization be used to make the data more openly available?**

Different anonymization techniques exist. Disadvantage of anonymization is that data integration becomes virtually impossible, but it may be the only way to open up your data for other research

Tags: *maDMP*

✔ *b. Yes*

3.a.1.b.1.b.4 **Could you use data aggregation to make the data openly available?**

Aggregated data, where typically at least 15 individuals are in any data point, are considered sufficiently anonymous. This is an alternative way of making data openly available for future research

Tags: *maDMP*

✔ *a. No*

3.a.1.b.2 **Are there IP reasons why data can not be open?**

✔ *b. Yes*

3.a.1.b.2.b.1 **Is it clear who owns data and documents?**

✔ *b. Yes*

3.a.1.b.2.b.1.b.1 **Who will own the intellectual property rights (copyrights) of the data that you will collect or create?**

✔ *PI*

3.a.1.b.2.b.2 **Will someone be given decision power to move documents or data to a new place after the project has finished?**

In one case in the past, all documents that had been assembled by a project in a documentation system had to be deleted because not a single person could decide to move them to a new platform when the documentation system was going off-line.

✔ *a. No*

3.a.1.b.3 **Will you/your users be allowing authenticated access to the data?**

Tags: *Science Europe DMP*

✔ *a. No*

3.a.2 **Are there business reasons why (some of) the data at your RI can not be completely open?**

Tags: *Science Europe DMP*

✔ *c. Yes, other business reasons*

3.a.2.c.1 **What other business reasons are there not to open all data immediately?**

Tags: *Science Europe DMP*

✔ *IP is owned by the user*

3.a.3 **Are there other reasons why (some of) the data at your RI can not be completely open?**

Tags: *Science Europe DMP*

✔ *a. No*

3.a.4 **Will you use a limited embargo?**

Tags: *Science Europe DMP*

✔ *a. No, some restricted data will be embargoed indefinitely*

3.a.4.a.1 **What is the maximum embargo period?**

✔ *indefinite*

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

✔ *b. Yes*