NorSeq_gap_analysis

Core DS Knowledge Model

Organization	ELIXIR Norway
Created by	Federico Bianchini (fredebi@uio.no)
Based on	RI gap analysis, 0.0.1 (elixir.no:ri-elixir-norway:0.0.1)
Project Phase	Before Submitting the DMP
Created at	11 Mar 2021

I. Administrative details

Report

Indications

Answered (current phase)	13 / 15
Answered	20 / 24

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*

✓ *Federico Bianchini*

1.b.2 E-mail address

Tags: *maDMP, Science Europe DMP*

✓ *federico.bianchini@mn.uio.no*

1.b.3 ORCID Identifier

Tags: *maDMP, Science Europe DMP*

✓ *0000-0002-9016-4820*

1.b.4 Affiliation

Tags: *Science Europe DMP*

✓ *Centre for Bioinformatics, University of Oslo*

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*

✓ *Arvind Sundaram*

1.c.2 E-mail address

Tags: *maDMP, Science Europe DMP*

✓ *arvind.sundaram@medisin.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*

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

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*

✓ *j. RI Member*

1.d.1 Name

Tags: *maDMP, Science Europe DMP*

✓ *Robert Lyle*

1.d.2 E-mail address

Tags: *maDMP, Science Europe DMP*

✓ *Robert.Lyle@medisin.uio.no*

1.d.3 ORCID Identifier

Tags: *maDMP, Science Europe DMP*

✓ *0000-0001-6317-732X*

1.d.4 Affiliation

Tags: *Science Europe DMP*

✓ *Department of Medical Genetics, Oslo University Hospital, Oslo, Norway*

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*

✓ *j. RI Member*

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*

✓ *NorSeq*

2.b.2 Project short discription

Tags: *maDMP, Science Europe DMP*

✓ *The National Consortium for Sequencing and Personalized Medicine is a Norwegian consortium with partners at the universities and university hospitals in Oslo, Bergen, Trondheim and Tromsø. Our aim is to provide cost-efficient high-throughput DNA sequencing analyses to researchers, and to facilitate the development and implementation of personalized medicine in Norway.*

2.b.3 Date the RI will started

Tags: *Science Europe DMP, maDMP*

✓ *2017*

2.b.4 Date the RI funding will end

Tags: *Science Europe DMP, maDMP*

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

> 2.b.5 **Funding**
>
> Add all the funding that are part of this project.
>
> Tags: *maDMP, Science Europe DMP*
>
> ✗ **This question has not been answered yet!**

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

Tags: *Science Europe DMP*

✓ *a. No*

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

✓ *a. No*

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)	1 / 2
Answered	1 / 2

Metrics

No metrics for this chapter.

Questions

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

✓ *NorSeq provides cost-efficient high throughput DNA sequencing to researchers with the aim to facilitate the development and implementation of personalized medicine.*

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*

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

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)	39 / 42
Answered	52 / 55

Metrics

Metric	Score
Findability	1
Accessibility	0.5
Interoperability	1
Reusability	0.8

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*

✓ *FASTQ Sequence and Sequence Quality Format*

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

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*

✓ *d. The user will have to take care of data storage immediately*

1.c.1 Data format/type

Tags: *Science Europe DMP*

✓ *Binary Alignment Map Format*

FAIRsharing https://fairsharing.org/bsg-s000210

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*

✓ *d. The user will have to take care of data storage immediately*

1.d.1 Data format/type

Tags: *Science Europe DMP*

✓ *Variant Call Format*

FAIRsharing https://fairsharing.org/bsg-s000270

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*
>
> ✓ *d. The user will have to take care of data storage immediately*

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*
>
> ✓ *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*
>
> ✓ *b. Yes*
>
> 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.

✓ *b. Yes*

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.

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

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.

✓ *b. Our work flow system documents the provenance automatically and completely*

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.

✓ *b. Yes*

3.a.7.a.1.b.1 **Describe your SOP (Standard Operating Procedure) for file naming**

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

✓ *File names always include the sample name and read number. The former is provided by the user upon submission of the sample and can be modified by the CF to avoid ambiguity with other samples. The read number is always “R1” for single-read runs and “R1/R2”NextSeq for paired-end runs (first/second read). Other strings depend on the Ilumina instrumentation used or on the run mode. Result files from MiSeq, NextSeq and Novaseq contain a sample number string which marks the order of appearance in the run and it is set to 0 when only one sample is analysed. In the case of HiSeq, a “Barcode Sequence” string is used instead. This is an index sequence used for the sample, which is set to 'NoIndex' if the samples are not multiplexed. Lanes are independent subunits of the flow cell: machines can have 1, 2 or 4 lanes depending on the model. Knowledge of the lane number becomes important if lanes are loaded with different pools of libraries, which is the case for HiSeq and for NovaSeq (Xp mode only). When relevant, a different file is provided for each lane and the lane number is explicitly indicated in the file name. If the user indicated bioinformatics analyses to be performed, additional files will be delivered depending on the specific service.*

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.

✓ *b. Yes*

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.

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

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*

Answers

4.b.2.b.1 **Instrument name**

Tags: *Science Europe DMP*

✓ *Illumina MiSeq*

4.b.2.b.2 **Instrument description**

Tags: *Science Europe DMP*

✓ *Sequencing platform https://www.illumina.com/systems/sequencing-platforms.html*

4.b.2.c.1 **Instrument name**

Tags: *Science Europe DMP*

✓ *Illumina NextSeq*

4.b.2.c.2 **Instrument description**

Tags: *Science Europe DMP*

✓ *Sequencing platform https://www.illumina.com/systems/sequencing-platforms.html*

4.b.2.d.1 **Instrument name**

Tags: *Science Europe DMP*

✓ *HiSeq 2500*

4.b.2.d.2 **Instrument description**

Tags: *Science Europe DMP*

✓ *Sequencing platform https://www.illumina.com/systems/sequencing-platforms.html*

4.b.2.e.1 **Instrument name**

Tags: *Science Europe DMP*

✓ *HiSeq 3000*

4.b.2.e.2 **Instrument description**

Tags: *Science Europe DMP*

✓ *Sequencing platform https://www.illumina.com/systems/sequencing-platforms.html*

4.b.2.f.1 **Instrument name**

Tags: *Science Europe DMP*

✓ *HiSeq 4000*

4.b.2.f.2 **Instrument description**

Tags: *Science Europe DMP*

✓ *Sequencing platform https://www.illumina.com/systems/sequencing-platforms.html*

4.b.2.g.1 **Instrument name**

Tags: *Science Europe DMP*

✓ *HiSeq X*

4.b.2.g.2 **Instrument description**

Tags: *Science Europe DMP*

✓ *Sequencing platform https://www.illumina.com/systems/sequencing-platforms.html*

4.b.2.h.1 **Instrument name**

Tags: *Science Europe DMP*

✓ *PacBio RSII*

4.b.2.h.2 **Instrument description**

Tags: *Science Europe DMP*

✓ *https://www.pacb.com/products-and-services/sequel-system/*

4.b.2.i.1 **Instrument name**

Tags: *Science Europe DMP*

✓ *PacBio Sequel*

4.b.2.i.2 **Instrument description**

Tags: *Science Europe DMP*

✓ *https://www.pacb.com/products-and-services/sequel-system/*

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.

✓ *a. No, this is all fine*

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

Tags: *Science Europe DMP*

✓ *a. No*

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?**

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

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)	1 / 2
Answered	1 / 3

Metrics

No metrics for this chapter.

Questions

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

✓ *b. No*

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.

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

3 **Other ethical / legal issues.**

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

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)	43 / 48
Answered	51 / 59

Metrics

Metric	Score
Accessibility	1
Reusability	0.83
Good DMP Practice	1

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*

✓ *b. Yes*

ELIXIR Norway offers NeLS a multi tiered shared storage for collaborating on data sets

> ### 1.b.1 Will this work space be run by dedicated specialists?
>
> If your work space is run and maintained by specialists, e.g. the ICT department of one of the institutes involved in the projects, this means that backup and restore as well as access management is properly addressed.
>
> ✓ *b. Yes*
>
> ### 1.b.2 How will you/your users work with the data?
>
> There are several questions regarding the dynamics of the data in the working area, who works with it, the software that is run on it, etc.
>
> Tags: *Science Europe DMP*
>
> ✓ *a. Explore*
>
> > ### 1.b.2.a.1 What kind of data will you/your users have in the work space?
> >
> > When making the work space, it helps to know whether you expect to work with very many small files, a few very large files, whether you will use a (SQL) database to store most of the data. Maybe your data is suitable for a system like Hadoop? Such information can be collected here.
> >
> > ✓ *Directory with FASTQ data*
> >
> > ### 1.b.2.a.2 Do you/your users need the work space to be close to the compute capacity?
> >
> > If you have large volumes of data that are intensely and repeatedly used by the computing work flow, it may be needed to keep the storage in the same place as where the computing takes place.
> >
> > Data Stewardship for Open Science: *wia*
> >
> > ✓ *a. No*

1.b.2.a.3 **Will you/your users be working with your data in another form than the way it will be archived?**

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

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

1.b.2.a.4 **How does the storage need change over time?**

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

Tags: *Science Europe DMP*

✓ *a. Storage needs will be the same during the whole RI runtime*

1.b.2.a.5 **Will you need to temporarily archive data sets (e.g. to tape)?**

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

✓ *a. No*

1.b.2.a.6 **How will your first data come in?**

✓ *a. No special planning is needed for the initial data*

1.b.2.a.7 **How will the RI parterns/ the users access the work space?**

✓ *a. Explore*

> **1.b.2.a.7.a.1** **Who will arrange access control?**
>
> ✓ *b. RI management will need to be able to give people access*
>
> **1.b.2.a.7.a.2** **Will the work space storage need to be remote mounted?**
>
> ✓ *c. No, all data processing will be done in the same environment (virtual research environment)*
>
> **1.b.2.a.7.a.3** **Will data be copied out and in to the workspace storage by remote users?**
>
> ✓ *b. No, all data processing will be done in the same environment (virtual research environment)*

1.b.3 **How available/reliable should must the work space be?**

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

Tags: *Science Europe DMP*

✓ *a. Explore*

> **1.b.3.a.1** **How do you prevent a total loss of data in the work space?**
>
> Tags: *Science Europe DMP*

✓ *b. All essential data is also stored elsewhere*

1.b.3.a.1.b.1 **Is there software in the work space? Can it also be restored quickly?**

Data Stewardship for Open Science: *cbq*

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

1.b.3.a.2 **Can you/your users handle it when the work space is off line for a while?**

✓ *a. We could handle a few days of offline time per year*

1.b.3.a.3 **How long can you/your users wait for a restore if the storage fails?**

✓ *b. A spare copy needs to be deployed quickly*

1.b.3.a.4 **How long can you wait for a restore if you accidentally damage a file?**

✓ *b. Hours*

1.b.3.a.5 **Will you make backup copies of your/your users data that is not in the work space?**

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

Tags: *Science Europe DMP*

✓ *b. There is no data elsewhere*

1.b.4 **How will access control to the work space be controlled?**

Tags: *Science Europe DMP*

✓ *e. Only specfic members of user project and of the RI have read/write access to the data*

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*

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

2.a.1.b.2 Will you use a (sub)folder for each (repeated) analysis?

Tags: *Science Europe DMP*

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

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

Tags: *Science Europe DMP*

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

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*

✓

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*

✓ *a. No*

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.

✓ *a. This has been arranged*

4 How will you make sure to know what exactly has been run?

✓ *a. Explore*

4.a.1 **Will you keep results together with all processing scripts or workflows including documentation of the versions of the tools that have been run?**

✓ *b. Yes*

4.a.2 **Will you make use of the metadata fields in your output data files to register how the data was obtained?**

File formats like VCF (for genetics) and TIFF (for images) have possibilities to document metadata in the file header. It is a good idea to use work flow tools that use these fields to document what was done to obtain the data.

✓ *b. Yes*

4.a.3 **Will you use a central repository for all tools and their versions as used in your 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*

✓ *b. Yes*

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*

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

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.

✓ *b. Yes*

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?

✓ *b. Yes*

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.

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

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.

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

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*

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

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

If you require substantial amounts of compute power, amounts that are not trivially absorbed in what you usually have abailable, some planning is necessary. Do you think you need to do compute capacity planning?

✓ *a. No*

7 **Is the risk of information loss, leaks and vandalism acceptably low?**

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*

✓ *a. No*

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*

✓ *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

✓ *a. No*

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

✓ *a. No*

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*

✓ *b. Yes*

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

Tags: *Science Europe DMP*

✓ *b. Yes; the effect is small*

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

Tags: *Science Europe DMP*

✓ *b. Yes; the effect is small*

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

Tags: *Science Europe DMP*

✓ *b. Yes; the effect is small*

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

Tags: *Science Europe DMP*

✓ *a. We are not using any personal information*

8 **Do you have a contingency plan?**

What will you do if the compute facility is down?

✓ *b. We have an alternative*

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)*

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

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)	1 / 8
Answered	1 / 8

Metrics

Metric	Score
Interoperability	0

Questions

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

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

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

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

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

✓ *a. No*

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

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

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

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

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

✗ This question has not been answered yet!

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

✗ This question has not been answered yet!

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

Data Stewardship for Open Science: *bzu*

✗ This question has not been answered yet!

VII. Preserving data

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

Report

Indications

Answered (current phase)	7 / 12
Answered	15 / 29

Metrics

Metric	Score
Accessibility	1
Good DMP Practice	0.5

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?**
>
> ✓ *b. As soon as it has all arrived, in one session*

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*

✓ *FASTQ*

2.b.2 Description of the data type

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

Tags: *Science Europe DMP, maDMP*

✓ *FASTQ format is a text-based format for storing a biological sequence and its quality scores, both encoded with a single ASCII character.*

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?
>
> ✗ **This question has not been answered yet!**
>
> #### 2.b.3.b.2 The actual identifier
>
> ✗ **This question has not been answered yet!**

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*

✓ *a. As long as technically possible*

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*

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

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*

✓ *a. No*

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

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

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

Tags: *Science Europe DMP*

✓ *b. Yes*

3.b.1 **How will you/your users be paying for these services?**

Tags: *Science Europe DMP*

✓ *d. Costs will be paid by other means*

3.b.1.d.1 **How will these costs be paid?**

Tags: *Science Europe DMP*

✓ *Arranged by the user/PI, not by the RI*

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*

✓ *b. Yes*

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?**

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

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?

✗ This question has not been answered yet!

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*

✗ This question has not been answered yet!

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.

✗ This question has not been answered yet!

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 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*

✗ This question has not been answered yet!

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*

✗ This question has not been answered yet!

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*

✗ This question has not been answered yet!

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

✗ This question has not been answered yet!

14 **Will you do structural modeling?**

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

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)	3 / 9
Answered	3 / 9

Metrics

Metric	Score
Openness	0

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*

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

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

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

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*
>
> ✗ **This question has not been answered yet!**
>
> 3.a.2 **Are there business reasons why (some of) the data at your RI can not be completely open?**
>
> Tags: *Science Europe DMP*
>
> ✗ **This question has not been answered yet!**
>
> 3.a.3 **Are there other reasons why (some of) the data at your RI can not be completely open?**
>
> Tags: *Science Europe DMP*
>
> ✓ *c. Yes, other reasons*
>
> > 3.a.3.c.1 **What other reasons are there not to open all data immediately?**

> > Tags: *Science Europe DMP*
> >
> > ✓ *Data belong to the user.*
>
> 3.a.4 **Will you use a limited embargo?**
>
> Tags: *Science Europe DMP*
>
> ✗ **This question has not been answered yet!**

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

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