Core DS Knowledge Model

Organization: *DSW Global*
Created by: *Gerald W (e0125536@student.tuwien.ac.at)*
Based on: *Common ELIXIR Knowledge Model, 1.0.0*
Project Phase: *Before Submitting the Proposal*
Created at: *22.04.2019*

I. Design of experiment

Before you decide to embark on any new study, it is nowadays good practice to consider all options to keep the data generation part of your study as limited as possible. 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.

Report

Indications

Answered	0
Unanswered	0

Metrics

Metric	Score
Findability	0
Accessibility	0
Interoperability	0
Reusability	0
Good DMP Practice	0
Openness	0

1

Is there any pre-existing data?

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

Data Stewardship for Open Science: *atq*
External Links:

✔

b. Yes

1.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?

Data Stewardship for Open Science: *ezi*
External Links:

✔

b. Yes

1.b.1.b.1

What reference data will you use?

Much of todays data is used in comparison with reference data. A genome for instance is 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?

Data Stewardship for Open Science: *quc*
External Links:

Answers

✔
Item 1.b.1.b.1.a Fremont Bridge Hourly Bicycle Counts by Month October 2012 to present

1.b.1.b.1.a.1
Do you know where and how is it available?

Do you know where the reference data is available, what the conditions for use are, and how to reference it?

Data Stewardship for Open Science: *ckt*
External Links:

✔
b. Yes

1.b.1.b.1.a.2
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*
External Links:

✔
a. I can directly use it

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

Data Stewardship for Open Science: *rgy*
External Links:

✔
a. No

1.b.1.b.1.a.4
How will you make sure the same reference data will be available to reproduce your results?

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

1.b.1.b.2
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 earlier . You may need to integrate your new data with an existing data set or retrieve additional information from related data bases. Will you be doing such things?

Data Stewardship for Open Science: *wya*
External Links:

Answer

✖

This question has not been answered yet!

1.b.2

Do you need to harmonize different sources of existing data?

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

Data Stewardship for Open Science: *wht*
External Links:

✔

a. No

1.b.3

Will you be using data that needs to be (re-)made computer readable first?

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

Data Stewardship for Open Science: *pth*
External Links:

✔

a. No

2

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*
External Links:

✔

b. Yes

2.b.1

What will the Intellectual Property be like?

Who will own the rights to the reference data set? Who will be able to use it?

Data Stewardship for Open Science: *hct*
External Links:

Answer

✖

This question has not been answered yet!

2.b.2

How will you maintain it?

How will maintenance be paid for in the long run? Will you host it yourself or deposit it with a repository? How will you deal with requests for help? And with requests for adding data?

Data Stewardship for Open Science: *usx*
External Links:

Answer

✖

This question has not been answered yet!

2.b.3

How will the release schedule be?

Will you be updating the reference data at regular intervals?

Answer

This question has not been answered yet!

3

Will you be storing samples?

Data Stewardship for Open Science: *kuz*

External Links:

a. No

4

Will you be collecting experimental data?

Data Stewardship for Open Science: *csx*

External Links:

a. No

II. Data design and planning

In the data design and planning phase, we will make sure that we know what data comes when, that we have enough storage space and compute power to deal with it, and that all the responsibilities have been taken care of.

Report

Indications

Answered	0
Unanswered	0

Metrics

Metric	Score
Findability	0.5
Accessibility	0.5
Interoperability	1
Reusability	0.5
Good DMP Practice	0.5
Openness	0

1

What data formats/types will you be using?

Have you identified types of data that you will use that are used by others too? Some types of data (e.g. genetic variants in the life sciences) are used by many different projects. For such data, often common standards exist that help to make these data reusable. Are you using such common data formats?

Data Stewardship for Open Science: *njy*
External Links:

Answers

✔

Item 1.a CSV

1.a.1
Is this a standard data format used by others too?

✔

b. Yes

1.a.2
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

✔

b. Yes

2
Will you 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*
External Links:

✔

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

3
How will you 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

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

✔

a. Explore

3.a.1
Do suitable 'Minimal Metadata About ...' (MIA...) standards exist for your experiments?

✔

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?

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.

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*

External Links:

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?

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.

b. Yes

3.a.5.b.1

Will you store the licenses with the data at all time?

It is very likely that data will be moved and copied. At some point people may lose track of the origins. It can be helpful to have the licenses (of coarse as open as possible) stored in close association with the data.

Data Stewardship for Open Science: *atw*

External Links:

✔

b. Yes

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.

Answer

✖

This question has not been answered yet!

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.

✔

b. Yes, all metadata is also explicitly available elsewhere

4

During the project, will you be archiving data (using so-called 'cold storage') for long term preservation?

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*
External Links:

✔

a. No

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

Data Stewardship for Open Science: *ixr*
External Links:

✔

a. No

4.a.2
When is the raw data archived?

✔

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

5
Will you need a shared working space to work with your data?

✔

a. No

5.a.1
Are data all project members store adequately backed up and traceable?

✔

a. No

6
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.

✔

a. Explore

6.a.1
Do project 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

✔

a. No

6.a.2
Do project members carry data with them?

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

✔

a. No

6.a.3

Do project members store project data in cloud accounts?

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

✔

a. No

6.a.4

Do project members send project data or reports per e-mail or other messaging services?

✔

a. No

6.a.5

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

✔

b. Yes

6.a.6

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

✔

b. Yes

6.a.7

Have project members been instructed about the risks (generic and specific to the project)?

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

✔

a. No

6.a.8

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

✔

b. Yes; the effect is small

6.a.9

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

✔

b. Yes; the effect is small

6.a.10

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

✔

a. No

7

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

III. Data Capture/Measurement

Report

Indications

Answered	0
Unanswered	0

Metrics

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

1

Please specify what data sets you will acquire using measurement equipment

Answers

Item 1.a Bicycle count automatically via induction loops by Seattle Department of Transportation

1.a.1

Who will do the measurements? And where?

Are there easily accessible specialized service providers for data capture?

c. External party

consider making them partner in the project

1.a.1.c.1

Has formal ownership of the data been established?

a. The party measuring the data owns it

1.a.1.c.2

Has responsibility for long term safe keeping of the raw data been established? Who will deal with data publication?

d. We have made other arrangements

1.a.1.c.2.d.1

What other arrangements?

The measuring party will periodically release new data under public domain licence.

1.a.2

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?

✔

b. Less well described or not completely standard

Take pictures of the instruments for documentation, and keep copies of any documentation.

1.a.3

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

2

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?

Data Stewardship for Open Science: *ybw*

External Links:

✔

a. No

3

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

✔

a. No

3.a.1

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

✔

b. Yes

IV. Data processing and curation

Report

Indications

Answered	0

Unanswered	0

Metrics

Metric	Score
Findability	0
Accessibility	0
Interoperability	0
Reusability	0.21052631578947367
Good DMP Practice	0
Openness	0

1

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

2

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

✔

a. Explore

2.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

2.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

2.a.3

Will you use a central repository for all tools and their versions as used in your project?

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

Data Stewardship for Open Science: *pzq*

External Links:

✔

a. No

2.a.4

Will you use a central repository for reference data used in your project?

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

Data Stewardship for Open Science: *pzq*

External Links:

✔

b. Yes

2.a.5

Will you make use of standard workflow engines and automatic work flows for all data analysis in the project?

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

✔

a. No

2.a.6

Are all software tools in the work flow professionally maintained, with version control?

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

✔

b. Yes

3

How will you validate the integrity of the results?

✔

a. Explore

3.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

3.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

3.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

3.a.4

Will you run part of the data set 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*

External Links:

a. No

4

Do you have a contingency plan?

What will you do if the compute facility is down?

a. We will wait until the problem is fixed

V. Data integration

Report

Indications

Answered	0
Unanswered	0

Metrics

Metric	Score
Findability	0
Accessibility	0
Interoperability	0.38095238095238093
Reusability	0
Good DMP Practice	0
Openness	0

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

Item 1.a.1.a CSV

1.a.1.a.1

How is the data structured?

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

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

Are all column headers unambiguous?

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

b. Yes

1.a.1.a.1.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.

a. No

Check whether you can find a data type (potentially in a data type registry) for each of your data items

1.a.1.a.1.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.

a. No

Check whether there is a way to explicitly limit data values.

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

Is it clear what a row in the table represents?

b. Yes

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

Does each row have an identifier?

a. No

1.a.1.a.1.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.a.1.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.a.1.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 be using a workflow for data integration, e.g. with tools for database access or conversion?

Data Stewardship for Open Science: *qqb*

External Links:

✔

a. No

1.a.3

Will you use a 'linked data' approach?

✔

a. No

2

Will you be using common or exchangeable units?

✔

a. No

3

Will you be using common ontologies?

✔

a. No

4

Will there be potential issues with statistical normalization?

✔

a. No

5

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

✔

a. No

6

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

✔

a. No

7

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

✔

a. No

VI. Data interpretation

Report

Indications

Answered	0
Unanswered	0

Metrics

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

1

Will data interpretation and modeling require significant compute infrastructure capacity?

✔

a. No

2

How will you be making sure there is good provenance of the data 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.

✔

b. We use an electronic lab notebook

3

Will you be doing (automated) knowledge discovery?

Data Stewardship for Open Science: *bzu*
External Links:

a. No

VII. Information and insight

Report

Indications

Answered	0
Unanswered	0

Metrics

Metric	Score
Findability	0.4
Accessibility	0
Interoperability	0
Reusability	1
Good DMP Practice	0
Openness	0.6666666666666666

1

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

Data Stewardship for Open Science: *jvm*
External Links:

b. Yes

2

Can all of your data become completely open immediately?

b. Yes

3

Specify a list of data sets you will be publishing

Specify a short name for each data set, sufficient for yourself to know what data it is about. It is useful to think about a data set as some collection of data that will be ending up in the same place.

Answers

Item 3.a Fremont Bridge Data

3.a.1

What kind of repository will this data be stored in?

Domain repositories often have the best functionality to make the data findable and reusable. Many of them are listed in https://fairsharing,org/

✔

a. A domain-specific repository

3.a.1.a.1

What repository?

✔

Zendoo

3.a.1.a.2

Will you contact the repository beforehand?

Contacting the repository early may be useful to establish conditions, formats, and metadata requirements for submission. It may also be necessary to establish whether the repository can accommodate your data

✔

a. No, this submission is routine both for us and the repository

3.a.2

Will you be adding a reference to the published data to at least one data catalogue?

Data is sometimes difficult to locate, especially if it is not in a domain-specific repository. Data catalogues may increase findability.

✔

a. No

4

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

✔

a. No

5

Did you work out the financial aspects of making the data available?

✔

a. Explore

6

Will your publications be open access?

Data Stewardship for Open Science: *mjf*

External Links:

✔

a. No

7

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

8

Will you be archiving your data after the project 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*

External Links:

a. No

9

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

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

b. Yes

10

Specify a list of software packages you will be publishing

Specify a short name for each software package.

Answers

Item 10.a Jupyter Notebook

10.a.1

Will you be adding a proper open-source license?

b. Yes, we have decided on an open source license

10.a.1.b.1

What is the license?

CC0

10.a.2

Where will the software package be available?

Github/Zendoo

10.a.3

Will this software be listed in a catalogue?

a. No

11

Will there be planning of valorization or translational returns?

✔

a. No