Hypothetical Next-Generation Cancer Cure Development

Horizon Europe
Data Management Plan

21 June 2022

Data Management Plan created in Data Stewardship Wizard «ds-wizard.org»
using Common DSW Knowledge Model v2.4.2 (dsw:root:2.4.2).

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
v1.0	19 Mar 2022	DMP ready for project proposal
v0.2	19 Mar 2022	Overview of data re-use and collecting
v0.1	19 Feb 2022	First draft with basic project information

Contributors

The following contributors are related to the project of this DMP:

- **Jana Freeman**
  jana.freeman@ds-wizard.org, iD 0000-0005-0002-0001
  Roles: *Data Manager*, *Project Manager*
  Affiliation: *DSW*
- **Marek Suchanek**
  marek.suchanek@ds-wizard.org, iD 0000-0002-0002-0031
  Roles: *Data Manager*
  Affiliation: *Czech Technical University in Prague*

Projects

We will be working on the following project and for those are the data and work described in this DMP.

Optimalization of data collection in HPLC - ICP MS analysis

Acronym: *HPLC - ICP MS collection*

Start date: *2022-01-01*

End date: *2022-12-31*

Funding: *Grantová Agentura České Republiky: GA22-123-456-789 (applied)*

Objective of this project is to optimize data collection from HPLC - ICP MS analysis.

1. Data Summary

Instrument datasets

The following instrument datasets will be acquired in the project:

- **HPLC - ICP MS**

  This dataset will be collected by an external party. For the ownership of the data we have made the following arrangements: both parties own the data.

  The equipment is very well described and known.

  Researchers working in other fields of research could be interested in using this data. We think that other researchers can use this data as follows: to compare the results.

Non-equipment datasets

The non-equipment datasets are:

- **Optimisation of data collection** – software for optimisation of data collection

Re-used datasets

We have found the following reference datasets that we have considered for re-use:

- **Chemotion repository** (https://pubchem.ncbi.nlm.nih.gov/source/1195) ✔

  Owner of this dataset: Unknown.

  The dataset can be used in the provided format without any conversion needed.

  We will use version "v 1.6" of this dataset. If a new version becomes available during the project, new analyses will be done with the new version.

  We will keep a copy of the dataset and make it available with our results for the reproducibility.

  We will use the dataset as follows: to speed up collection of the data.

- **DataCite Repository** (https://datacite.org/) ✕

  We decided not use this reference dataset because of: it plays only supportive role.

We have found the following non-reference datasets that we have considered for re-use:

- **DesignSafe Data Depot Repository**
  (https://www.re3data.org/repository/r3d100012308#:~:text=Description%20The%20Desig

×

We decided not use this non-reference dataset because of: it plays only supportive role.

We will need to harmonize different sources of existing data.

We will be using data that needs to be (re-)made computer readable first. We will provide machine readable, standardised metadata to others. The data itself will be made available in computer readable form to others through a standard repository.

Data formats and types

We will be using the following data formats and types:

- **Comma-separated Values**

  It is not a standardized format. Best suits the need of our project. This is not a suitable format for long-term archiving; however, we plan to convert it to a suitable format before the end of the project. We expect to have 99 files of average size 0.1 GB (i.e. approximately 9.9 GB in total).

2. FAIR Data

2.1. Making data findable, including provisions for metadata

- **HPLC -ICP MS data** (published)
  The dataset has the following identifiers:
  - DOI: **DOI:15.8111/22222**

  We will distribute the dataset using:
  - *Domain-specific repository*: Chemotion repository. We have made other arrangements: long term mutual agreement.
    A persistent identifier will be assigned by the repository. The repository will make sure that the persistent identifier can be resolved to a digital object. The assigned persistent identifier is specified: DOI.
  - *Special-purpose repository for the project*. We will be able to support this repository for a sufficiently long time. The repository will provide a search and simple access interface.
    A persistent identifier will be assigned by an institutional data steward. The repository will make sure that the persistent identifier can be resolved to a digital object. The assigned persistent identifier is specified: PURL.

There will be different versions of this data over time; the versions will be numbered. We will be adding a reference to the published data to at least one data catalogue.

There are no 'Minimal Metadata About ...' (MIA...) standards for our experiments. However, we have a good idea of what metadata is needed to make it possible for others to read and interpret our data in the future.

We will use other solution than (electronic) lab notebooks to make sure that there is good provenance of the data analysis: we use university cloud environment.

The provenance will be captured using W3C PROV.

We made a SOP (Standard Operating Procedure) for file naming. We will be keeping the relationships between data clear in the file names. All the metadata in the file names also will be available in the proper metadata.

2.2. Making data accessible

We will be working with the philosophy *as open as possible* for our data.

The data cannot become completely open immediately because of:

- legal reasons
- non-patent business reasons: contract between authors and business partner has to be mutually signed
- we have other than paper-publishing reasons: novelty

Concerning the legal reasons, a data sharing agreement will be required. People can apply to the *CTU Data Access Committee* data access committee.

Data that is not legally restrained will be released after a fixed time period (5 years), unconditionally.

Metadata will be openly available without instructions how to get access to the data. Metadata will not be available in a form that can be harvested and indexed.

We have made the following arrangements regarding the data ownership: by service contract.

For the reference and non-reference data sets that we reuse, conditions are as follows:

- **Chemotion repository** – freely available with obligation to quote the source (e.g. CC-BY).

For our produced data, conditions are as follows:

- **HPLC -ICP MS data** (published)

The distributions will be accessible through:

- *Domain-specific repository*: Chemotion repository. We have made other arrangements: long term mutual agreement. The distribution will be available under the following license:
    - Starting 1.1.2023: Available under some restrictions, which we will follow in our project: they will be available only to the research team. Re-users will be able to get access through a specialized process: via cloud environment. The conditions will be published as part of open metadata. More infomation about the restrictions can be found here: https://en.wikipedia.org/wiki/License.
- *Special-purpose repository for the project*. It will be *Shared* with a predefined list of people. We will be able to support this repository for a sufficiently long time. The repository will provice a search and simple access interface. The distribution will be available under the following license:
    - Starting 1.1.2023: Freely available with obligation to quote the source (e.g. CC-BY).

A user of this data need specific software to be able to use it:

- CTU cloud environement (cvut.cz)

The dataset will published after initial cleanup.

2.3. Making data interoperable

We will be using the following data formats and types:

- **Comma-separated Values**

  It is not a standardized format. Best suits the need of our project.

We will be using the following standards (encodings, terminologies, vocabularies, ontologies):

- **Chemistry vocabulary** (https://www.fkit.unizg.hr/_news/32312/1%20-%20Basic%20Chemistry%20Vocabulary%20List.pdf)

2.4. Increase data re-use

The metadata for our produced data will be kept as follows:

- **HPLC -ICP MS data** (published) – This data set will be kept available for a fixed period (prepaid). – The metadata will be available even when the data no longer exists.

The following instrument datasets acquired in the project will use the following quality processes:

- **HPLC - ICP MS**
  - Calibrating measurements
  - Repeat samples / measurements
  - Data Entry validation
  - Use of controlled vocabularies
  - Other quality processes: QA + QC processes as per university recommendation

As explained in Section 2.2, our data cannot become completely open immediately.

Due to privacy reasons, the data must stay in the EU. We can use pseudonymization, anonymization, and data aggregation to make the data more openly available. For pseudonymization, we cannot make use of an existing 'trusted third party'.

There are IP reasons why our data can not be open. It is clear who owns data and documents.

Someone will be given the decision power to move documents or data to a new place after the project has finished.

We will be archiving data (using so-called *cold storage*) for long term preservation already during the project. The data are expected to be still understandable and reusable after a long time.

To validate the integrity of the results, the following will be done:

- We will run a subset of our jobs several times across the different compute infrastructures.
- We will be instrumenting the tools into pipelines and workflows using automated tools.
- We will use independently developed duplicate tools or workflows for critical steps to reduce or eliminate human errors.
- We will run part of the data set repeatedly to catch unexpected changes in results.

3. Other research outputs

We use Data Stewardship Wizard for planning our data management and creating this DMP. The management and planning of other research outputs is done separately and is included as appendix to this DMP. Still, we benefit from data stewardship guidance (e.g. FAIR principles, openness, or security) and it is reflected in our plans with respect to other research outputs.

4. Allocation of resources

FAIR is a central part of our data management; it is considered at every decision in our data management plan. We use the FAIR data process ourselves to make our use of the data as efficient as possible. Making our data FAIR is therefore not a cost that can be separated from the rest of the project.

We will be archiving data (using so-called 'cold storage') for long term preservation after the project but also already during the project. The costs for archiving data will be paid from the grant. The minimum lifetime of the archive is 12 years. The archival period can be extended – one of the principle investigators involved in the project will decide. The decision whether or not to extend the renewal be based on available budget. Data formats of data in cold storage will be upgraded if they become obsolete. Archived data will not be migrated to other storage media over time.

We will be paying for costs related to the used repositories. Further notes: it will be covered by a business partner.

We have a reserved budget for the time and effort it will take to prepare the data for publication. For making data or other research outputs FAIR, we budgeted: 250 euros.

Marek Suchánek is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Jana Freeman and Marek Suchanek are responsible for maintaining the finished resource.

To execute the DMP, additional specialist expertise is required. We will be hiring new people with additional expertise. The required expertise from new people is: data stewardship.

We require the following hardware or software in addition to what is usually available in the institute: Dedicated processors (hardware) to capture a high amount of data in a short period of time.

5. Data security

Project members will not store data or software on computers in the lab or external hard drives connected to those computers. They can carry data with them on password-protected laptops. All data centers where project data is stored carry sufficient certifications. All project web services are addressed via secure HTTP (https://...). Project members have been instructed about both generic and specific risks to the project.

The risk of information loss in the project or organization is acceptably low. The risk of information leak in the project or organization is acceptably low. The risk of information vandalism in the project or organization is acceptably low.

All personal information will be processed in pseudonymized form only. We have a specific way of pseudonymization: masking technique will be used.

The archive will be stored in a remote location to protect the data against disasters. The archive need to be protected against loss or theft. It is clear who has physical access to the archives.

6. Ethics

For the data we produce, the ethical aspects are as follows:

- **HPLC -ICP MS data**
    - It does not contain personal data.
    - It does not contain sensitive data.

Data we collect

We will collect data connected to a person, i.e. "personal data". We ask the data subjects for their consent. We ask for consent for anonymization; we will anonymise first and all further processing is on the anonymous data. The consent form will be available for re-users.

7. Other issues

We use the Data Stewardship Wizard with its *Common DSW Knowledge Model* (ID: dsw:root:2.4.2) knowledge model to make our DMP. More specifically, we use the https://internal.ds-wizard.org DSW instance where the project has direct URL: https://internal.ds-wizard.org/projects/99719f7e-23e6-47d6-a005-9f0a6599255d.

We will be using the following policies and procedures for data management:

- **Recommendations by CTU International Advisory Board**
  https://www.cvut.cz/en/international-advisory-board
  To be aligned with the International Advisory Board that provides the opinion on principal directions of scientific and educational programmes and activities, research programmes and their evaluation at CTU.