J. Heyrovský Institute of Physical Chemistry

Data Management Plan Instructions for the J. Heyrovsky Institute of Physical Chemistry of the CAS

July 2022

This document provides instructions on how to create a data management plan (DMP) for submission to and consideration by the Data Management Team at the J. Heyrovsky Institute of Physical Chemistry, of the Czech Academy of Sciences. This document consists of 5 sections; four sections contain obligatory questions to be answered when preparing your DMP. The last section outlines several recommendations to be considered when handling data.

A. Data reuse:

Before starting a project, please consider whether there are any existing data to be reused.

1. Is there any pre-existing data for this project?
2. Are you planning to reuse this data? If yes, how?

B. New data - storage:

It is important to plan the storage and handling of data produced during a project.

3. What type, format and volume of data will you generate and collect during the project?

   Example: I. spectra, CSV format, 20 GB/year; II. images (electron microscopy), TIFF format, 1 TB/year; III. Molecular dynamics trajectories XTC format, 200 GB/year.

4. How do you plan to store your data directly after production? What back-up procedures do you have in place?

   Where are the data stored? i.e. on the hard drive of the instrument or on your own computer; NAS; institutional, national or international/domain-specific repository? If repositories, please specify details. Do you plan to store your data in multiple locations?

5. Do you envisage costs related to the storage and handling of your data? If yes, please specify.

   It is important to consider the cost of data storage for a project. Examples: costs for new NAS drives; fees for access to (inter-)national data repositories (as well as data curation, at own personal cost).

6. What is your plan for data preservation at the end of your project?

   Once the project is finished, how will the data be stored in the long-term? Do you plan to use a generic or domain specific repository (licensed) for long-term data preservation? If yes, which one? Please consider the funding costs.

C. New data – annotation/metadata:

It is important to annotate data in order to enable their future recognition and reuse.

7. How will you provide information on your data (metadata)?

Briefly describe all types of documentation (README files, metadata, etc.) which you will provide to help secondary users understand and reuse your data. The metadata file should include, at a minimum, basic details allowing other users to find the data, including provenance and whether the data is human- and machine-readable. This includes a persistent identifier (PID); the name of the person who collected and/or contributed to the data; institution; the date of collection; and the conditions of access to the data (license). Furthermore, the documentation should ideally include details on the applied methodology; data processing and implemented analytical steps; variable definitions; references to vocabularies; as well as measurement units. Wherever possible, the documentation should follow existing community standards and guidelines. Please explain how you will prepare and share this information.

8. Do you use an electronic lab notebook for data handling and annotation? If yes, which kind?

   Examples: Word document(s), Evernote, SciNote, openBIS, …

D. Licensing and data protection:

Adding an appropriate license stimulates the reuse of data. At the same time, it protects the data from inappropriate dissolution. Intellectual properties and ethical issues must be considered before sending data to any data storage.

9. Under which license will your data be shared? Please consider which limitations are applied.

   May the data be shared with others? Examples: after publishing in a paper; after patenting; after a 6-month embargo period etc. Please specify the access limitations which apply to all your data collected throughout a project. Diverse licenses may be found here*.*

10. Do you expect your data to require any special handling according to intellectual property rights?

    Data will be stored in agreement with the institutional intellectual property rules. For advice, please contact The Group for Intellectual Property.

11. Are there any ethical issues related to your data? Do you expect to collect any sensitive data?

    For advice, please contact the Committee for Ethics of Research Involving Human Subjects, at the institute.

Recommendations for data management (optional)

It is strongly recommended that you plan your data management before you start your project. Understandably, circumstances may change during a project. The DMP is therefore not a finite, but a live and ongoing document which may be edited. It is advisable to open your DMP document every 6-12 months in order to re-assess and reconsider your plans.
In addition to the issues covered by the questions above, we would like to add a few points which may further guide you in your data management during a project.

1. It is important to keep all your data organized according to pre-defined rules. The rules are defined by the software/protocols handling the storage location (e.g., NAS or repository storage space). Unfortunately, this is not always handled by the application offering the user interface for data management. Often, electronic lab notebooks (ELN) implement appropriate data organization for selected data storage systems.
2. In the event that data needs to be deleted, please be aware that any metadata MUST stay unaltered (with a note that the data were deleted; and why). New data can use the same identifier (PID) but *versioning* should be applied; e.g., spectra_chromophore_220426a (deleted) and spectra_chromophore_220426a.1 (recorded).
3. Consider defining a *data curator* for your data. This does not have to be an IT expert, although a person with insight into your research domain and specific data is favourable. A person can be trained as a data curator in a number of possible upcoming workshops – please contact the Heyrovsky Open Science Team (data stewards) for more details surrounding the training options. The *Data curator* controls how the data are stored and provides feedback to the users (with how the data have been stored) with the aim to store well-documented data in a safe place. Having a dedicated data curator requires the allocation of some funds for this position; the benefit of acquiring a trained member of staff to be responsible for data curation would greatly out-weigh any employment costs for this position and would save each individual colleague from having to learn the details in curating their own data.
4. We recommend that you carefully set up the *access rights* in your repository. It is important to protect your RAW data by providing READ access, but also by blocking the WRITE access to anyone else in the group. The data curator has FULL rights despite whether the data creator leaves the team or not.
5. The use of selected Electronic Lab Notebooks is strongly recommended. It is important to annotate (describe) methodological details for the experiment at the time the experiment is conducted. Subsequent revisions are acceptable (even after several months) but this may take a lot of time and effort when needing to provide a full report on older data. An ELN keeps all your information in a machine-readable format. Many ELNs also offer automatized data management tools (e.g., openBIS).
6. If working with *personal data* (especially of patients), please consider contacting an expert from the Working Group of the Czech National Repository Platform on Sensitive Data (website in Czech only, but contacts are fluent in English).

Created and agreed by members of the **Heyrovsky Open Science Team** (HOST; Eva Pluhařová, Stefan Swift and Marek Cebecauer), April-July 2022.

Data Management Plan for the Molecular Simulations of Catalysts for CO2 reductions.

Data Reuse and Storage

Is there any pre-existing data for this project?

Yes, structures of similar molecules in the .xyz format published in papers (https://onlinelibrary.wiley.com/doi/full/10.1002/anie.201814339) we cite in scientific part of the proposal. Empirical parametrization (force field) for several components of the systems is available as a part of Gromacs software package (https://manual.gromacs.org/current/user-guide/force-fields.html).

Are you planning to reuse this data? If yes, how?

Yes, the structures will be used for benchmark quantum chemical calculations. After conversion to the .gro format, they will serve as initial conditions for molecular dynamics simulations. The force field parameters for selected molecules will be employed as is.

Data Storage

What type, format and volume of data will you generate and collect during the project.

The data related to the benchmark quantum chemical calculations will be in the form of .inp and .log Gaussian files which are human readable .txt files. The classical molecular dynamics trajectories will be in the Gromacs .xtc format. Other Gromacs input files (.mdp, itp) and ouput files (.log) are human readable .txt files. The trajectory analysis will be done using Gromacs tools (https://manual.gromacs.org/current/user-guide/cmdline.html#commands-by-name) and .sh scripts. Metadata will be in the form of .txt files. The estimated amount of data produced during the project is 400 GB.

How do you plan to store your data directly after production? What back-up procedures do you have in place?

The data will be produced on the computer clusters of the institute (https://www.jh-inst.cas.cz/cs/about-departments/pristrojove-vybaveni-oddeleni-vypocetni-chemie) and in CESNET (https://metavo.metacentrum.cz/cs/state/index.html) with back-up procedures implemented. Selected data will be backed up to the National Data Storage system provided by CESNET (https://du.cesnet.cz/en/start).

Do you envisage costs related to the storage and handling of your data? If yes, please specify.

The part of the contribution to the administration of the institutional computer cluster related to the data handling is 15.000 Kc/year. The CESNET services are currently for free, but it may change during the last year of the project.

What is your plan for data preservation at the end of your project?

The analyzed data are expected to be published in journal articles. The new .xyz structures and force field parameters will be part of the Supporting Information. These are inherently

preservative. The input files, output files, scripts used for analysis and trajectories will be stored in the National Data Storage system provided by CESNET. The input files and scripts will also be stored at external hard drives.

New data – annotation/metadata:

How will you provide information on the data (metadata)?

The input files are human readable files which contain all information necessary to reproduce the calculations. The output files are human readable too and they always contain name of the file owner, date, version of the software and the path on the computer cluster where the simulation was executed. If pre-processing is needed it will be described in a README .txt file. The purpose and details of the analysis will be provided in the README too. Each subproject will be stored in one folder which will contain a README file describing its contents.

Do you use electronic lab notebooks for data handling and annotation? If yes, which kind?

Not yet. All information and parameters necessary to reproduce the calculations is present in the input files. Notes, description of the procedures and analysis is provided in the README .txt files.

Licensing and data protection:

Under which license will your data be shared? Please consider which limitations are applied.

The data stored in the National Data Storage system provided by CESNET will be given a READ permission to any authenticated repository user after publishing the journal article.

Do you expect your data to require any special handling according to intellectual property rights?

No additional intellectual property steps are going to be required, in addition to standard procedures described in the institutional ’s guidelines.

Are there any ethical issues related to your data? Do you expect to collect any sensitive data?

No

J. Heyrovský Institute of Physical Chemistry

EXAMPLE 2

Data Management Plan for the Characterisation and Quantification of Species emitted from cell cultures stressed inflammatory factors.

Data Reuse and Storage

Is there any pre-existing data for this project?

Yes, the data measured in our lab by Violetta Shestivska in 2014-2016, Project: Combination of SIFT-MS with electrochemical methods for real-time quantification of volatiles released by damaged bacterial and cell cultures (GACR).

Are you planning to reuse this data? If yes, how?

Yes, the data will be used as positive control – cells under oxidative stress by hydrogen peroxide. No PID provided for the data. Data are available on our cloud storage system: OneDrive.

Data Storage

What type, format and volume of data will you generate and collect during the project.

Raw Data will be in the form of .csv files, directly given by the Syft Voice200 Instrument. Raw data from the profile 3 instrument are acquired in the form of .mse files and will be converted to .csv for long-term storage and sharing. Manipulated and data processing (the calculations of exact quantities and comparisons) will be performed in MS Excel and the output stored in the form of .csv files. Metadata will be in the form of .txt files. Python Scripts (.py format) will also be attached to the data for the automation and processing purposes. Photos of experimental set-ups will be taken and stored as either .jpg or .png files. The estimated volume of data to be produced for this project will be 40 MB.
Images of cells stressed by inflammatory factors will be stored in OME-TIFF format. Approximately, 100 GB of image data are expected throughout the project. Functional assays will be evaluated using Fiji/ImageJ plugins. The results will be stored as SPSS portable files. Tables and graphs will not exceed 1 GB. The RAW video data will be stored in .MP4 format. 1TB of video files is expected.

How do you plan to store your data directly after production? What back-up procedures do you have in place?

Data will be back-up automatically onto the shared OneDrive folder associated with the research group. This OneDrive folder may be accessed from 3 separate and secure computers. The data will also be on the hard drive of one separate computer.
Microscopy data will be stored in local NAS file system storage (www.jh-inst.cas.cz/Synology16) with RAID 5 (for safety). READ only access will be provided. WRITE access to RAW data is only provided for the data curator, for all other data WRITE access is provided to the OWNER and data curator. Selected (key) data will be backed up to the National Data Storage system provided by CESNET (https://du.cesnet.cz/en/start).

Do you envisage costs related to the storage and handling of your data? If yes, please specify.

No costs are currently expected for the storage of these data.

New HDD for the NAS system (4 TB) will be bought at 5000 Kc. The salary of data curator (0.2 FTE; 600.000 Kc/3 years). ELN (openBIS) license and services (60.000 Kc).

What is your plan for data preservation at the end of your project?

The final data is expected to be published within a journal article which is inherently preservative. The intermittent data and final data which was not chosen for the purpose of publication will be deployed to the National Repository Platform (NRP; installed in 2024) for LTP (long-term preservation) with a yearly cost of 500 Kc (5 years, 2500 Kc). The data will also be stored on the hard drives of the computers or the NAS file system used by the principle worker, at least until the principle worker who worked on the data leaves the institute.

New data – annotation/metadata:

How will you provide information on the data (metadata)?

A README Notepad (.txt) file will accompany the quantitative MS data once uploaded onto the institute and national server services. The description within the README file will indicate where and when the data were acquired; the name of the person who acquired the data and who manipulated the data; as well as the conditions of access to the data (licence). The RAW datafiles given by the instrument are machine readable. The PID will be provided by the NRP for the deployed data.
These specific identifier pieces of information will be described as a list at the start of the README file. This will be followed by the experimental procedures which will be written out in the README file as the data are collected. The steps taken for the quantitation and the manipulation step of data will also be described, along with the journal references, glossaries and names of algorithms used, in order to achieve the datasets. The measurement units of c/s (counts per second) as well as ppbv (parts per billion by volume) will be mentioned in the README file (.txt).
Microscopy experiments will be recorded using ELN (openBIS). The system involves the RDM module to associate all necessary information/metadata (providence, experimental setting and procedures, data processing, project structure) with data submitted to the repositories. Any persons involved in these measurements will be asked to use the ELN. Technical information about the microscope setup is an integral part of the OME-TIFF (images) and MP4 (video) files.
Once shared publicly in the national repository, the data will acquire a PID (DOI format).

Do you use electronic lab notebooks for data handling and annotation? If yes, which kind?

The essential notes from each MS experiment are recorded in the comments sections of the .csv files which originate from the Profile 3 instrument. For the Voice200 instrument, notes are recorded in a paper notebook and will be transferred to a Notepad file (.txt).
OpenBIS ELN will be used for microscopy experiments. The system includes a RDM module.

Licensing and data protection:

Under which license will your data be shared? Please consider which limitations are applied.

MS: There will be no embargo applied. The data is also envisaged to be available through access to publication records from the chosen journal article of choice. Data will also be open access from the Institute and national repositories. No License is needed.

Microscopy: The NAS file system is not connected to the internet. Data access will be provided by: i) existing connections to the NAS system – within the Department of Biophysical Chemistry; ii) extended READ access to the collaborators from the Department of Chemistry of Ions in Gaseous Phase; and iii) the National Repository Platform under CC-BY-NC 4.0 license (upon upload to the repository; https://creativecommons.org/choose/) to any authenticated repository user. No embargo period will be applied.

Do you expect your data to require any special handling according to intellectual property rights?

No additional intellectual property steps are going to be required, in addition to standard procedures described in the institutional 's guidelines.

Are there any ethical issues related to your data? Do you expect to collect any sensitive data?

Gas phase MS data of species from breath analysis will be acquired as part of the project. These data will come from patients in a hospital and from healthy volunteers. Any personal and sensitive data acquired from hospital patient volunteers needs to be handled correctly with a security key applied to the data, at the point of storage. The data will be fully anonymised before being uploaded to any unsecured storage systems (e.g., NAS file system).