J. Heyrovský Institute of Physical Chemistry

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

Version 2.0
September 2025

This document explains how to create a Data Management Plan (DMP) for submission to and consideration by the Heyrovsky Open Science Team (HOST) at the J. Heyrovsky Institute of Physical Chemistry of the Czech Academy of Sciences. The document consists of six sections, five of which contain obligatory questions to be answered when preparing your DMP. The final section provides recommendations for handling research data.

I. Project roles and responsibilities

The role of the data steward is crucial for handling data. Please ensure that someone takes on this responsibility and is mentioned in the data management plan.

1. Who are the principal investigator and data steward for this project?

   Enter their names, ORCIDs and roles. If the same person has both roles, please indicate this.

II. Data reuse

Before starting a project, please consider whether any existing data could be reused.

2. Is there any pre-existing data for this project? If so, what is the source of this data, and how do you plan to reuse it?

III. Data production and storage

It is important to plan how data produced during a project will be stored and handled.

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

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

4. How will you store your data immediately after production? What backup procedures do you have in place?

   Where will the data be stored? For example, on the hard drive of the instrument, on your own computer, on a NAS, or in an institutional, national, or domain-specific repository? If in a repository, please specify the details. Do you plan to store your data in multiple locations?

5. Do you anticipate incurring costs relating to the storage and handling of your data? If so, please specify.

   It is important to consider the cost of storing data for a project. Examples include costs for new NAS drives and fees for access to national and international data repositories, as well as personal costs for data curation.

6. What is your plan for preserving the data after the end of the project?

   Once the project has finished, how will the data be stored long-term? Do you plan to use a generic or domain-specific (licensed) repository for long-term data preservation? If so, which one?

IV. Data annotation and metadata

Annotating data is important to enable its future recognition and reuse.

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

   Briefly describe all types of documentation (e.g., README files and metadata) that you will provide to help secondary users understand and reuse your data. The metadata file should include sufficient details that will allow researchers and other stakeholders to find the data. This includes provenance and information on whether the data is human- and machine-readable. It should include a persistent identifier (PID), the name of the person who collected and/or contributed to the data, the institution, the date of collection, and the conditions of access to the data (licence). Ideally, the documentation should also include details of the methodology applied, data processing and analytical steps implemented, variable definitions, references to vocabularies, and the measurement units used. Wherever possible, the documentation should adhere to existing community standards and guidelines. Please explain how you will prepare and share this information.

8. Do you use any domain-specific metadata models or data schemas?

   A metadata model defines the structure and meaning of the metadata – that is, data about your data. It specifies how to describe a dataset's content, context, quality, and origin/provenance.
   A data schema defines how the actual data is organised – what fields or variables (classes and properties) it contains, their data types (e.g., text, number, date), relationships, and constraints.
   Common metadata models/schemas include: CIF (Crystallographic Information File), MaterialsML, JCAMP-DX, CML (Chemical Markup Language), NMReDATA, NeXus, Dublin Core (with extensions), etc.

9. Do you use electronic lab notebooks for data handling and annotation? If so, which type?

   Examples: Kadi4Mat, eLabFTW, RSpace, Evernote, openBIS, etc.

V. Licensing and data protection

Appropriate licensing stimulates the reuse of data. At the same time, it protects against the misuse of the data. Consider intellectual sensitivity and confidentiality issues before sending data to any data repositories or digital services.

10. Which licence will you use to share your data? For open access, the most common licence is CC-BY 4.0. Please consider which limitations are applied.

    Open access can be granted immediately or after an embargo period. An embargo period is used to protect data before publication or patenting and may last up to three years. Please specify the access limitations that apply to all the data collected throughout the project. Various licences can be found here*.*

11. Do you expect your data to require any special handling in relation to intellectual property rights?

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

VI. Recommendations for data management (optional)

It is strongly recommended that you plan your data management strategy before starting your project. Understandably, circumstances may change during the course of a project. The DMP is therefore not a static document, but an ongoing one that can be edited. We advise opening your DMP document every 6 – 12 months to reassess and reconsider your plans.
In addition to the issues covered by the above questions, we would like to make a few further suggestions to help guide you in managing your data during the project.

a. It is important to keep all your data organised according to predefined rules. These rules are defined by the software/protocols that handle the storage location (e.g., NAS or repository storage space). Unfortunately, this is not always handled by the application that offers the user interface for data management. Often, electronic lab notebooks (ELNs) implement appropriate data organisation for specific storage systems.
b. In the event that data needs to be deleted, please ensure that any metadata remains unaltered, with a note stating that the data has been deleted and the reason 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).
c. Consider appointing a data curator. This does not have to be an IT expert, although someone with insight into your research domain and specific data would be preferable. Data curators can be trained in a number of possible upcoming workshops – please contact the Heyrovský Open Science Team (data stewards) for details of available training options. The data curator controls how the data is stored and provides users with feedback on how the data has been stored, with the aim of storing well-documented data safely. Although allocating funds for this position is required, the benefit of acquiring a trained member of staff to be responsible for data curation would greatly outweigh any

employment costs and save each colleague the time and effort of learning how to curate their own data.

d. We recommend that you carefully configure access rights within your repository. It is important to protect your RAW data by providing READ access but blocking WRITE access for anyone else in the group. The data curator should have FULL rights, regardless of whether the data creator leaves the team.
e. The use of selected Electronic Lab Notebooks is strongly recommended. It is important to annotate (describe) methodological details for the experiment at the time it is conducted. While subsequent revisions are acceptable, even after several months, providing a full report on older data may be time-consuming and labour-intensive. An ELN stores all your information in a machine-readable format. Many ELNs also offer automated data management tools (e.g., openBIS and Kadi4Mat).
f. If you are working with personal data, particularly that relating to patients, please consider contacting an expert from the Working Group of the Czech National Repository Platform on Sensitive Data. The website is also available in English.

Members of the **Heyrovský Open Science Team** (Eva Pluhařová, Alfredo González, Graciela Eguía and Marek Cebecauer) have updated and agreed to this, September 2025.

J. Heyrovský Institute of Physical Chemistry

EXAMPLE 1

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

I. Project roles and responsibilities

1. Who are the principal investigator and data steward for this project?

Principal investigator: Petr Novotný https://orcid.org/0000-0001-2345-6789
Data steward: Alena Nováková https://orcid.org/0000-0002-1825-0097

II. Data reuse

2. Is there any pre-existing data for this project? If so, what is the source of this data, and how do you plan to reuse it?

Yes, the structures of similar molecules in the .xyz format are published in the papers cited in the scientific part of the proposal (https://onlinelibrary.wiley.com/doi/full/10.1002/anie.201814339). The empirical parametrisation (force field) for several components of the systems is available as part of the Gromacs software package (https://manual.gromacs.org/current/user-guide/force-fields.html).
These structures will be used for benchmark quantum chemical calculations. After conversion to the .gro format, they will serve as the initial conditions for the molecular dynamics simulations. The force field parameters for the selected molecules will be used as they are.

III. Data production and storage

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

The data relating to the benchmark quantum chemical calculations will be in the form of human-readable .txt files, i.e. .inp and .log Gaussian files. The classical molecular dynamics trajectories will be in the Gromacs .xtc format. Other Gromacs input files (.mdp and .itp) and output files (.log) are also human-readable text files. Trajectory analysis will be performed 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. It is estimated that 400 GB of data will be produced during the project.

4. How will you store your data immediately after production? What backup 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 CESNET (https://metavo.metacentrum.cz/cs/state/index.html), with backup procedures implemented. Selected data will be backed up to the Czech National Repository (https://data.narodni-repozitar.cz/).

5. Do you anticipate incurring costs relating to the storage and handling of your data? If so, please specify.

The contribution to the administration of the institutional computer cluster relating to data handling is 15,000 CZK per year. CESNET services are currently free of charge, but this may change in the final year of the project.

6. What is your plan for preserving the data after the end of the project?

The analysed data are expected to be published in journal articles. The new .xyz structures and force field parameters will form part of the Supporting Information. These are inherently preservable. The input and output files, the scripts used for the analysis, and the trajectories will be stored in the Czech National Repository. The input files and scripts will also be archived on external hard drives.

IV. Data annotation and metadata

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

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

8. Do you use any domain-specific metadata models or data schemas?

Computational materials data will be structured according to MaterialsML and, where applicable, aligned with the OPTIMADE API specification.

9. Do you use electronic lab notebooks for data handling and annotation? If so, which type?

Not yet. All the necessary information and parameters to reproduce the calculations are present in the input files. Notes, procedure descriptions and analyses are provided in the README.txt files.

V. Licensing and data protection

10. Which licence will you use to share your data? For open access, the most common licence is CC-BY 4.0. Please consider which limitations are applied.

After the journal article has been published, any authenticated repository user will be given READ permission to the data stored in the Czech National Repository.

11. Do you expect your data to require any special handling in relation to intellectual property rights*?*

No additional intellectual property steps will be required beyond the standard procedures described in the institution’s guidelines.

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.

I. Project roles and responsibilities

1. Who are the principal investigator and data steward for this project?

Martina Dvořáková / Principal investigator and Data Steward / https://orcid.org/0000-0002-1694-233X

II. Data reuse

2. Is there any pre-existing data for this project? If so, what is the source of this data, and how do you plan to reuse it?

Yes, these are the data that were measured in our laboratory by Violetta Shestivska between 2014 and 2016 as part of the project: Combination of SIFT-MS with electrochemical methods for real-time quantification of volatiles released by damaged bacterial and cell cultures (GACR).
The data will be used as a positive control for cells under oxidative stress caused by hydrogen peroxide. No PID has been provided for the data. The data are available on our cloud storage system: OneDrive.

III. Data production and storage

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

The raw data will be in the form of .csv files, provided directly by the Syft Voice200 instrument. Raw data from the profile 3 instrument are acquired in the .mse format and converted to .csv for long-term storage and sharing. Data manipulation and processing (i.e. the calculation of exact quantities and comparisons) will be performed in MS Excel, with the output stored as .csv files. Metadata will be in the form of .txt files. Python scripts in .py format will also be attached to the data for automation and processing purposes. Photos of experimental set-ups will be taken and stored as either .jpg or .png files. It is estimated that this project will produce 40 MB of data.
Images of cells stressed by inflammatory factors will be stored in OME-TIFF format. Approximately 100 GB of image data is 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. RAW video data will be stored in MP4 format. One TB of video files is expected.

4. How will you store your data immediately after production? What backup procedures do you have in place?

Data will automatically be backed up to the shared OneDrive folder associated with the research group. This folder can be accessed from three separate, secure computers. The data will also be stored on the hard drive of one computer.
Microscopy data will be stored in the local NAS file system (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 to the data curator; for all other data, WRITE access is provided to the OWNER and data curator. Selected data will be backed up to the Institutional Repository (https://data.narodni-repozitar.cz/heyrovsky/datasets/all/).

5. Do you anticipate incurring costs relating to the storage and handling of your data? If so, please specify.

Currently, no costs are expected for storing these data.
A new 4 TB HDD for the NAS system will be purchased for 5000 Kc. The salary of the data curator is 600.000 Kc for 0.2 FTE other three years. The ELN licence and services will cost 60.000 Kc.

6. What is your plan for preserving the data after the end of the project?

The final data is expected to be published in a journal article, which is inherently preservative. Intermittent and final data not chosen for publication will be deployed to the National Repository Platform (NRP; installed in 2024) for long-term preservation (LTP) at an annual cost of 500 Kc (5 years, 2500 Kc). The data will also be stored on hard drives or the NAS file system.

IV. Data annotation and metadata

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

Once uploaded onto the institute and national server services, a README Notepad (.txt) file will accompany the quantitative MS data. The description in the file will indicate when and where the data were acquired, who acquired and manipulated the data, and the conditions of access to the data (licence). The raw data files provided by the instrument are machine-readable. The PID will be provided by the NRP for the deployed data.
These pieces of identifying information will be listed at the start of the README file. This will be followed by the experimental procedures, which will be documented in the README file as the data are collected. The quantitation and data manipulation steps will also be described, along with the journal references, glossaries, and names of algorithms used to produce the datasets. The measurement units of c/s (counts per second) and ppbv (parts per billion by volume) will be mentioned in the .txt README file.

Microscopy experiments will be recorded using an ELN. This system uses the RDM module to associate all the necessary information and metadata (e.g. providence, experimental settings and procedures, data processing and project structure) with the data submitted to the repositories. All 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 (image) and MP4 (video) files.
Once the data has been shared publicly in the institutional repository, it will acquire a PID in DOI format.

8. Do you use any domain-specific metadata models or data schemas?

No.

9. Do you use electronic lab notebooks for data handling and annotation? If so, which type?

Essential notes from each MS experiment are recorded in the comments section of the .csv files originating from the Profile 3 instrument. For the Voice200 instrument, notes are initially recorded in a paper notebook and then transferred to a Notepad (.txt) file.
Kadi4Mat ELN will be used for microscopy experiments.

V. Licensing and data protection

10. Which licence will you use to share your data? For open access, the most common licence is CC-BY 4.0. Please consider which limitations are applied.

MS: No embargo will be applied. The data will also be available via access to the publication records of the chosen journal article. The data will also be openly accessible from the institute and national repositories. A CC-BY 4.0 licence will be used.

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 for collaborators in the Department of Chemistry of Ions in Gaseous Phase; iii) the National Repository Platform under a CC-BY-NC 4.0 licence (upon upload to the repository: https://creativecommons.org/choose/). There will be no embargo period.

11. Do you expect your data to require any special handling in relation to intellectual property rights?

Standard procedures described in the institutional guidelines are all the intellectual property steps that are required.