SNSF Data Management Plan

EPFL Library and ETH Zurich Library template

Version 1.0

SNSF Data Management Plan

The recommendations provided in this document are intended for EPFL researchers.

Institution	
EPFL	
Responsibilities	
Principal Investigator: (Specify name and email)	Xile Hu
Data management plan contact person: (Specify name and email)	Nanjun Chen

Further information: **https://www.epfl.ch/campus/library/services/services-researchers/rdm-guides-templates/**

1. Data collection and documentation

1.1 What data will you collect, observe, generate or re-use?

Questions you might want to consider:

- What type, format and volume of data will you collect, observe, generate or reuse?
- Which existing data (yours or third-party) will you reuse?

Briefly describe the data you will collect, observe or generate. Also mention any existing data that will be (re)used. The descriptions should include the type, format and content of each dataset. Furthermore, provide an estimation of the volume of the generated datasets.

(This relates to the FAIR Data Principles F2, I3, R1 & R1.2)

➢ [see *List of recommended file formats*]

The data produced from this research project will fall into three categories:

1. The various reaction parameters for the synthesis. The summary of the experiments.
2. The spectroscopic and general characterization data. They include data from instruments including electron microscopy (TEM and SEM), X-ray photoelectron spectroscopy (XPS), nuclear magnetic resonance (NMR), X-ray powder diffraction (XRD),

infra-red (IR), UV-Visible absorption spectroscopy, X-ray absorption spectroscopy (XAS), ultraviolet photoelectron spectroscopy (UPS), Raman, elemental analysis, gas absorption, Mass Spectrometry etc.

3. Data from catalytic tests and mechanistic studies, such as electrochemical data, data from High Pressure Liquid Chromatography (HPLC), Gas Chromatography (GC), Gas Chromatography Mass Spectrometry (GC-MS), etc.

Data in category 1 will be documented in .docx, pptx., and .pdf.

Spectroscopic data in category 2 will be produced in their native formats, and converted to .csv and .txt for datasets, and .tif and .pdf for images.

The experimental instruments and simulations used within NCCR Catalysis will produce data with a wide range of formats, including but certainly not limited to:

Nuclear magnetic resonance (NMR) .fid, .mnova
Infrared spectroscopy .csv
UV-vis spectroscopy .csv
Raman spectroscopy .csv
X-ray diffraction .cif
Transmission electron microscopy (TEM) .tiff
Gas chromatography flame ionization detector (GC-FID) .gcd
Gas chromatography-mass spectroscopy (GC-MS) .xms
Liquid chromatography-mass spectroscopy (LC-MS) .raw
Electron paramagnetic resonance (EPR) .dsc, .dta

Data in category 3 will be in their native formats, and converted to .csv and .txt for datasets, and .pdf for graphics.

We anticipate that the data in category 1 will amount to approximately 1 GB, in category 2 will be 1 TB (mostly due to microscopy image), and in category 3 will amount to approximately 100 GB.

Contact for assistance:

EPFL: Research Data Library team (researchdata@epfl.ch)

1.2 How will the data be collected, observed or generated?

Questions you might want to consider:

- What standards, methodologies or quality assurance processes will you use?

- How will you organize your files and handle versioning?

Explain how the data will be collected, observed or generated. Describe how you plan to control and document the consistency and quality of the collected data: calibration processes, repeated measurements, data recording standards, usage of controlled vocabularies, data entry validation, data peer review, etc.

Discuss how the data management will be handled during the project, mentioning for example naming conventions, version control and folder structures. (This relates to the FAIR Data Principle R1)

The experimental records and observations are recorded in an internal electronic notebook (An ELN system developed by EPFL/ISIC) or sometimes by hand-written notes followed by digitization (scanning). The analytical data are collected by the instruments that generated them; they are processed by the native programs associated with the instruments. A periodic quality control process will be applied to remove errors and redundancies. Errors include for example incorrect handling and machine malfunction. The quality control process will be documented. For analytical data, the data collection is done through instrument standardised data acquisition programs. For data related to catalytic test and mechanistic studies (e.g., electrochemical data), lab-standardized protocols will be used.
The quality of experimental records and observations will be controlled by repeating experiments.

The various experimental procedures and associated compound characterization will be written up using the American Chemical Society standard formatting in a Word document (.docx), which will then be converted to PDF.

Files will be named according to a pre-agreed convention, e.g., **ProjectW-ReactionX-GenerationY-ScientistZ-YYYYMMDD-HHmm.csv.** The dataset will be accompanied by a **README** file which will describe the directory hierarchy, as well as the experimental protocol. This should allow the data to be understood by other members of our research group and add contextual value to the dataset should it be reused in the future.

1.3 What documentation and metadata will you provide with the data?

Questions you might want to consider:

- What information is required for users (computer or human) to read and interpret the data in the future?

- How will you generate this documentation?

- What community standards (if any) will be used to annotate the (meta)data?

Describe all types of documentation (README files, metadata, etc.) you will provide to help secondary users to understand and reuse your data. Metadata should at least include basic details allowing other users (computer or human) to find the data. This includes at least a name and a persistent identifier for each file, the name of the person who collected or contributed to the data, the date of collection and the conditions to access the data.

Furthermore, the documentation may include details on the methodology used, information about the performed processing and analytical steps, variable definitions, references to vocabularies used, as well as units of measurement.

Wherever possible, the documentation should follow existing community standards and guidelines. Explain how you will prepare and share this information. (This relates to the FAIR Data Principles I1, I2, I3, R1, R1.2 & R1.3)

The data will be accompanied by the following contextual documentation, according to standard practice for chemistry projects:

1. Spreadsheet documents which detail the reaction conditions.
2. Text files which detail the experimental procedures and compound characterization.

Files and folders will be named according to a pre-agreed convention YXZ, which includes for each dataset, identifications to the researcher, the date, the study and the type of data (see section 1.2).

The final dataset as deposited in the chosen data repository will also be accompanied by a README file listing the contents of the other files and outlining the file-naming convention used.

2. Ethics, legal and security issues

2.1 How will ethical issues be addressed and handled?

Questions you might want to consider:

- What is the relevant protection standard for your data? Are you bound by a confidentiality agreement?
- Do you have the necessary permission to obtain, process, preserve and share the data? Have the people whose data you are using been informed or did they give their consent?
- What methods will you use to ensure the protection of personal or other sensitive data? Ethical issues in research projects demand for an adaptation of research data management practices, e.g. how data is stored, who can access/reuse the data and how long the data is stored. Methods to manage ethical concerns may include: anonymization of data; gain approval by ethics committees; formal consent agreements. You should outline that all ethical issues in your project have been identified, including the corresponding measures in data management. (This relates to the FAIR Data Principle A1)

Environmental protection and safety:

The PI assures that appropriate health and safety procedures conforming to relevant local/national guidelines/legislation are followed for staff involved in this project. The health and safety of all participants in the research (investigators, subjects involved or third parties) must be a priority in all research projects (see http://securite.epfl.ch/page-34437-en.html). The project will be conducted under EPFL’s Lex 1.5.1 – Directive concerning occupational health and safety (DSST). The present directive determines the assignment of functions relating to health and safety in the workplace. It specifies the responsibilities of all the actors who must work as part of a network at EPFL. It also forms an integral part of risk management, at both CEPF and EPFL levels.

2.2 How will data access and security be managed?

Questions you might want to consider:

- What are the main concerns regarding data security, what are the levels of risk and what measures are in place to handle security risks?
- How will you regulate data access rights/permissions to ensure the security of the data?
- How will personal or other sensitive data be handled to ensure safe data storage and transfer?

If you work with personal or other sensitive data you should outline the security measures in order to protect the data. Please list formal standards which will be adopted in your study. An example is ISO 27001-Information security management. Furthermore, describe the main processes or facilities for storage and processing of personal or other sensitive data. (This relates to the FAIR Data Principle A1)

Data will be stored on the centralized file storage system (group folder) managed by our IT department. The access to the data is managed through the EPFL identity management system, which is a secured system following the best practices in terms of identity management. Our central storage facility has redundancy, mirroring and is monitored.

2.3 How will you handle copyright and Intellectual Property Rights issues?

Questions you might want to consider:

- Who will be the owner of the data?
- Which licenses will be applied to the data?
- What restrictions apply to the reuse of third-party data?

Outline the owners of the copyright and Intellectual Property Right (IPR) of all data that will be collected and generated including the licence(s). For consortia, an IPR ownership agreement might be necessary. You should comply with relevant funder, institutional, departmental or group policies on copyright or IPR. Furthermore, clarify what permissions are required should third-party data be re-used. (This relates to the FAIR Data Principles I3 & R1.1)

The work may generate intellectual properties. The intellectual property rights are set out in EPFL policies. The IP protection will be assisted by the EPFL Technology Transfer Office. Data relevant to IP are confidential until patent applications are filed. The aim is to then publish the work in a research journal and to publish the supporting data under an open Creative Commons Attribution (CC BY) license.

Data not suitable for IP protection can be published and shared without the above restriction.

3. Data storage and preservation

3.1 How will your data be stored and backed-up during the research?

Questions you might want to consider:

- What is your storage capacity and where will the data be stored?
- What are the back-up procedures?

Please mention what the needs are in terms of data storage and where the data will be stored.

Please consider that data storage on laptops or hard drives, for example, is risky. Storage through IT teams is safer. If external services are asked for, it is important that this does not conflict with the policy of each entity involved in the project, especially concerning the issue of sensitive data.

Please specify your back-up procedure (frequency of updates, responsibilities, automatic/manual process, security measures, etc.)

Storage and back up will be in three places:

- On personal computer of each researcher
- On a portable storage device (hard drive) of the group
- On institutional collaborative storage

Each researcher will be responsible for the storage and back up of data. **The backup will be done weekly** on institutional server, and monthly on portable hard drives Backups on the institutional infrastructure are automated using the RSYNC tool. The lab has access to up to 1 terabyte of information storage, which can be expanded if needed. The institution's Collaborative Storage is backed-up on a regular basis.

3.2 What is your data preservation plan?

Questions you might want to consider:

- What procedures would be used to select data to be preserved?
- What file formats will be used for preservation?

Please specify which data will be retained, shared and archived after the completion of the project and the corresponding data selection procedure (e.g. long-term value, potential value for re-use, obligations to destroy some data, etc.). Please outline a long-term preservation plan for the datasets beyond the lifetime of the project.

In particular, comment on the choice of file formats and the use of community standards.

Data will be stored for a minimum of three years beyond award period, per funder's guidelines. If inventions or new technologies are made in connection data, access to data will be restricted until invention disclosures and/or provisional patent filings are made with the institutional Technology Transfer Office (TTO).

4. Data sharing and reuse

4.1 How and where will the data be shared?

Questions you might want to consider

- On which repository do you plan to share your data?
- How will potential users find out about your data?

Consider how and on which repository the data will be made available. The methods applied to data sharing will depend on several factors such as the type, size, complexity and sensitivity of data.

Please also consider how the reuse of your data will be valued and acknowledged by other researchers.

(This relates to the FAIR Data Principles F1, F3, F4, A1, A1.1, A1.2 & A2)

It is recommended to **publish data in well established** (or even certified) domain specific **repositories**, if available :

- re3data is a repository directory allowing to select repositories by subject and level of trust (e.g. certifications)
- EPFL Library provides support to guide researchers in the choice of an appropriate (disciplinary) repository.
  A project for an EPFL institutional repository is currently ongoing (expected for 2018).
  However, EPFL strongly encourages its researchers to use disciplinary repositories when they exist.

In domains for which no suitable subject repositories are available, generalist repositories are available.
Among the most common used:

- Zenodo (free, maximum 50GB/dataset, hosted by CERN)
- Dryad (120$ for the first 20GB and 50$ for additional GB, Non-profit organization)
- Figshare (free upload, maximum 5GB / dataset, commercial company)

The studies will be published as research articles in scientific journals. Data underpinning the studies are included in the articles. Moreover, a large amount of supporting data will be published as online supplementary materials. All major publishers in our field support the deposit of such data in a digital format, freely accessible to public, on their websites. At the same time, these data will be deposited at EPFL's InfoScience platform, freely accessible to public. The datasets for each publication will be uploaded to Zenodo server, with the corresponding doi number cited in each paper.

Each dataset publication should contain the following metadata:

● Title

● Abstract describing the dataset

● Authors including affiliations

● Description of the folder structure and the contents

● A license file; NCCR Catalysis recommends using the Creative Commons Attribution 4.0

International license (CC-BY-4.0).

If Zenodo is used as the data repository, these requirements are automatically fulfilled as this data is requested during the submission process. If another qualified data repository is chosen that does not automatically include this metadata, it can be written in a README file that is included in the dataset.

In addition to these requirements, researchers are encouraged to ensure that the deposited datasets respect the following guidelines. These guidelines can be quite technical and the Data

Officer can support researchers in understanding and implementing them where necessary.

● Text files should be encoded using UTF-8 whenever possible.

● Standard and non-proprietary file formats should be used as much as possible.

● When providing code- or instrument-specific data, or proprietary formats, a description

of the format specifications should be included in the metadata.

● The folder structure should be logically organized and well described in the metadata.
● Use descriptive folder and file names, avoiding spaces and special characters.

The data published are searchable through google scholar and other specialized search engines. Links to InfoScience are available from our website. Crystallographic data will be additionally deposited in Cambridge Structural Database (https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/), which is freely available to the public.

4.2 Are there any necessary limitations to protect sensitive data?

Questions you might want to consider:

- Under which conditions will the data be made available (timing of data release, reason for delay if applicable)?

Data have to be shared as soon as possible, but at the latest at the time of publication of the respective scientific output.

Restrictions may be only due to legal, ethical, copyright, confidentiality or other clauses. Consider whether a non-disclosure agreement would give sufficient protection for confidential data.

(This relates to the FAIR Data Principles A1 & R1.1)

Data which underpins any publication will be made available at the time of publication. Data related to intellectual property rights are confidential until patent applications are filed.

4.3 All digital repositories I will choose are conform to the FAIR Data Principles

[CHECK BOX] Yes

4.4 I will choose digital repositories maintained by a non-profit organisation

[RADIO BUTTON yes/no] Yes