Data Management Plan

# 1. Data collection and documentation

## 1.1 What data will you collect, observe, generate or reuse?

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 data sets. (This relates to the [FAIR Data Principles](http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf) F2, I3, R1 & R1.2.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Type | Content | Format | Volume per dataset | Overall data volume |
| Two-Photon Microscopy | Brain barriers and fluid transport | TIFF (.tif)  Imaris format (.im) and video (.mp4) (visualizations) | 30 GiB (raw data)  30 GiB (image correction)  30 MiB (visualizations) | 12 TiB |
| Near Infrared Imaging (NIR) | Anatomy and liquid flow in cerebrospinal fluid spaces | Zen Format (.czi )  Exported as TIFF (.tif) or video (.avi) | 450 MiB | 60 GiB |
| Fluorescence and Confocal Imaging | 2D images of tissue sections | Nikon format (.nd2)  Zen format (.czi)  Export as TIFF (.tif) | 8.2 MiB (Nikon multichannel image)  9 MiB (TIFF conversion) | 500 MiB |
| Magnetic Resonance Imaging (MRI) | CSF instant  velocity and  immune cell  distribution | Raw data in ‘fid’  file with text files  (.txt)  Reconstructed  data in ‘2dseq’  file exported as  TIFF (.tif) | 20 MiB | 2 GiB |
| In vivo Synchrotron Radiation X-ray computed microtomography (SRµCT), ESRF | 3D images of mouse skulls with contrast agent injected into cerebrospinal fluid spaces | Raw binary (.edf)  Reconstructions exported as TIFF (.tif) | 1.9 TiB (projections)  2.3 TiB  (reconstructions) | 168 TiB |
| Ex vivo Synchrotron Radiation X-ray computed microtomography (SRµCT), Soleil | 3D images of mouse brain tissues with contrast agent injected into the vasculature | TIFF (.tif) | 4.5 TiB (projections)  12 TiB  (reconstructions) | 198 TiB |
| Image acquisition and processing software | In-house image acquisition and processing software  Scripting of repeated tasks | Source code (.c, .R, .m, .py, .sh, .ijm, .bat) | < 1 MiB | < 100 MiB |
| Processed and annotated data for reanalysis by other researchers or educational purposes | Cerebrospinal fluid space atlas | OME-TIFF (.tif) | 12 TiB | 12 TiB |
| Computational fluid dynamic simulations | 3D fluid dynamics simulations of cerebrospinal fluid flow and transport using commercial software | Simulation files (.sim, .simh, .cas, .dat)  Exported data and images (.csv,.tif) | 10 GiB (initial simulation file, including mesh and set-up)  10 GiB (per saved intermediate time step; typically 50-100 saved time-steps per simulation)  10 GiB (final simulation results) | 20 TiB |
| CSF transport simulations | In-house code for CSF production and transport.  Simulation results (including a copy of the code version and set-up used) | Source code (.c, .m, .py, .sh)  Results as text files (.csv) | < 2 MiB | < 200 MiB |

## 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](http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf) R1.)

Data collection:

Image data will be collected using instruments tabulated below. Instruments are calibrated and set up according to manufacturer’s guidelines.

For use of the external synchrotron facilities, calibration and measurement protocols of the custom-built hardware and software are handled by the facility staff.

Computational models are developed according to community standards and validated against experimental data.

|  |  |  |
| --- | --- | --- |
| Type | Instrument | Software |
| Two-Photon Microscopy | LaVision Biotec TrimScope II | Imaging with the Trimscope software  Image correction with In Vivo Follow 2  Volume rendered 4D images – Imaris |
| Near Infrared Imaging (NIR) | Zeiss StereoLumar.V12 | Image sequences in Zen software |
| Fluorescence and Confocal Imaging | Nikon Eclipse E600 and Zeiss LSM800 | Nikon NIS Elements software and Zen software |
| Magnetic Resonance Imaging (MRI) | Bruker BioSpec 14.1T  MRI | Imaging with Paravision  360  Reconstruction and  analysis with Matlab |
| In vivo Synchrotron Radiation X-ray computed microtomography (SRµCT), ESRF | Facility custom setup | Imaging with facility in-house software  Reconstruction with facility in-house software (PyHST2) |
| Ex vivo Synchrotron Radiation X-ray computed microtomography (SRµCT), Soleil | Facility custom setup | Imaging with facility in-house software  Reconstruction with in-house software |
| Image acquisition and processing software | Computers | Text editors and integrated development environments (IDE) |
| Processed and annotated data for reanalysis by other researchers or educational purposes | Computers | In-house image processing software  Fiji/ImageJ |
| Computational fluid dynamic simulations | Computers | StarCCM+, Ansys Fluent |
| CSF transport simulations | Computers | Text editors and integrated development environments (IDE) |

For data documentation, it will be responsibility of:

* The principal Investigator to introduce the group leaders to the documentation procedure
* The group leaders to introduce new students and group members to the documentation procedure
* The group leaders to send experimental protocols to members of the other groups for review of completeness and understandability, and then to upload to ZORA.
* Each researcher to annotate data with metadata
* The Principal Investigator to check the repositories biannually with all participants to assure data and protocols are being properly processed, documented, and stored.

File organization and versioning:

Image data are raw data that may not be altered or updated, and therefore do not require versioning. Data are organized first by type (see section 1.1.) then date of acquisition. Processed data are kept separately.

Simulation data are similarly treated as raw data that may not be altered of updated. Simulations are first organized by type and then by simulation data. For simulations conducted with in-house software, the code version used is automatically saved together with the simulation data and similarly treated as raw data. Post processing data are kept separately.

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

Questions you might want to consider:

* What information is required for computers or humans 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](http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf) I1, I2, I3, R1, R1.2 & R1.3.)

Metadata will be provided as data entry into the databases of the respective repositories as much as possible and follow the community guidelines of the respective repository. Metadata that cannot be provided in the respective repository databases will be provided as downloadable text files or on the central repository on ZORA. Templates for these metadata will ensure consistency across the repository.

Experimental protocols

* Experimental Materials & Methods
* Experimental Standard operating procedures
* Experimental Checklists

Experimental protocols will be compiled after major experimental sessions and provided on ZORA.

Image data

* Image acquisition settings
* Content description
* File sizes and formats (incl. image dimensions and bit depth for raw binary formats)

Image acquisition settings are generated automatically during scans and stored in various formats depending on the image modality (header in image files, XML-files, text files). Content descriptions will be written after every major experimental session and stored both in a central database and along with the data.

Fluid dynamic simulations

* Simulation date
* Software version
* Simulation settings
* Initial and boundary conditions

Metadata of the computational simulations will be generated automatically during the simulation. These will be saved in standard .csv file format. When relevant, these text files will be amended with data linking the simulation to the underlying experimental conditions.

# 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](http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf) A1.)

The proposed project involves animal experiments. All interventions, no matter how minor, are considered as harmful, as they are not executed to the benefit of the individual animal involved. As such, the harm done to the animal has to be weighed against the societal benefits of the research. This weighing of interest is performed independently of the researchers by governmental ethics committees, which have to review and approve every experiment individually. No experiment will take place without such approval. To minimize harm, we have designed the project with consideration of the 3R principles of animal research (Replace, Reduce, Refine).

There is currently no cell culture model that adequately mimics the anatomical structure of the brain cerebrospinal fluid spaces. Computational models are not available, as our project involves the very generation of such models. They may serve to replace certain animal studies in the future, but require the investment of animal experiments to provide necessary underlying data and validation through our project in the present.

Acknowledging that mice are indispensable for our scientific goals, and that interest in this research field has been strongly rising, we have made transparency of our methods and reusability of our data a priority. Our efforts in providing SOPs openly and independently of journal publications, we enable other researchers to access our most up-to-date protocols within a short timeframe, allowing them to refine their own animal experiments and reduce unnecessary repeat optimization trials. By providing all our data at repositories, the research groups in computational modeling can opt to reuse our data rather than perform their own experiments. We expect that this will – looking at the field as a whole – reduce the number of animals used.

## 2.2 How will data access and security be managed?

Software:

* Access to the NAS is password protected and only accessible by research group members and the institute’s IT administrators.
* Backup and snapshots can only be accessed by the institute’s IT administrators.

Hardware:

* NAS and backup NAS are located in locked rooms that are only accessible to the institute’s IT administrators.
* Backup tapes are stored in a safe at the institute that is only accessible to the principle investigator and two senior research group members. The second copy is stored in a library archive, only accessible by the university library staff.

Data in repositories are fully public, with no access restrictions.

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

Data will be published as Creative Commons Attribution 4.0 International (CC BY 4.0).

Ownership is governed by the respective institutions' staff regulations.

# 3. Data storage and preservation

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

Processed and frequently used raw data are stored on a Synology NAS, arranged in RAID 6 configuration (can tolerate failure of two hard disks). Data are backed up with snapshots to a second NAS every night. Snapshots for the last year are kept. Offline backups are written to tape once a year.

For SRµCT data, synchrotrons additionally maintain tape backups of all raw data collected at their facilities as a general policy, independent of users. We additionally archive all raw data on LTO-7 tapes immediately after returning from the synchrotron, and only retain frequently used raw data on the NAS due to the data volume.

## 3.2 What is your data preservation plan?

Storage:

We aim to upload all data to repositories, which have policies in place for long-term archival. With the exception of SRµCT data, all data is less than 50 GB per dataset and can be uploaded to Zenodo. SRµCT data collected at ESRF are offered for access in the ESRF Data Repository after an embargo period as a general policy by ESRF, independent of users. SRµCT collected at Soleil will be provided via the ETH Research Collection.

NAS backups are archived on two offline copies on LTO-7 Tapes and stored at two different sites. File hashes for all files are calculated to allow identification of potential data corruption and serve as table of contents. As a general policy, we retain all raw data collected, regardless of their usefulness for our immediate scientific needs, as they may prove useful for future studies or computational models.

For SRµCT data, synchrotrons additionally maintain tape backups of all raw data collected at their facilities as a general policy, independent of users.

File formats:

All image data collected in proprietary formats (.czi and .nd2) will be exported and stored as TIFF (.tif). Image data uploaded to Image Data Resource (IDR) will be converted to OME-TIFF as per the repository’s requirements.

Videos in our projects are generated for visualizations only. As they do not constitute raw data, we prioritize interoperability over lossless compression. They are exported by the image visualization software as MPEG-4 Part 2 ASP in an AVI container or MPEG-4 Part 10 AVC in an MP4 container.

Reconstructed 3D volumes of SRµCT data collected at ESRF will be exported as TIFF, while the raw projections will be retained in the ESRF data format (.edf). As this is simple binary raw data with a 1024 byte header, it is an open, interoperable and well-documented file format. While still less interoperable than TIFF, it allows for direct reconstruction of 3D volumes using the ESRF’s PyHST2 reconstruction software, which would be used by all researchers not using their own, custom reconstruction software. As researchers with such specialized know-how should have no issue opening and converting raw binary data, and researchers without this know-how would use PyHST2, we consider the ESRF data format (.edf) more operable than TIFF for raw projections.

Source code is stored as ASCII text files.

Simulation project files from commercial software packages are inherently tied to the simulation software used in their generation and no open standard exists. They are thus retained in their proprietary formats. Simulation set-up will be exported as ASCII text files. Results of simulations will be exported as CSV files.

Simulation results from in-house software packages will be exported as CSV files. Source code will be stored as ASCII text files along with them.

# 4. Data sharing and reuse

## 4.1 How and where will the data be shared?

|  |  |
| --- | --- |
| * Central repository linking to other repositories * Experimental protocols * Hardware schematics * Documents * Metadata not in other repositories | [Zurich Open Repository and Archive](https://www.zora.uzh.ch/) (ZORA) |
| * Microscopy image data * MRI image data * Simulation set-up and final result files   Second copies of:   * In-house software source code * Platform-independent in-house software application (wrapped in Virtual Machine) * Experimental protocols * Hardware schematics * Documents * Metadata not in other repositories | [Zenodo](https://zenodo.org) |
| * SRµCT image data collected at ESRF | [ESRF Data Repository](https://data.esrf.fr/) |
| * SRµCT image data collected at Soleil | [ETH Research Collection](https://www.research-collection.ethz.ch/) |
| * Processed and annotated data for reanalysis by other researchers or educational purposes | [Image Data Resource](https://idr.openmicroscopy.org/) (IDR) |
| * In-house software source code | [Github](https://github.com) (All software will be submitted to Zenodo as well) |
| * Platform-independent in-house software application (wrapped in Virtual Machine) | [SourceForge](https://sourceforge.net/) (All software will be submitted to Zenodo as well) |

## 4.2 Are there any necessary limitations to protect sensitive data?

No

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

Yes

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

Yes