# Data Management Plan- BiUM Model

# content of the mySNF form

UNIL | Université de Lausanne
Faculté de biologie et de médecine



Dr. Cécile Lebrand
FBM Publication & Data Management Unit- 2017
Bibliothèque Universitaire de Médecine CHUV

# Data Management Plan – content of the mySNF form

| Question | Help text |
|---|---|
| **1 Data collection and documentation** | |
| ***1.1 What data will you collect, observe, generate or re-use?***<br>Questions you might want to consider:<br>- What type, format and volume of data will you collect, observe, generate or reuse?<br>- 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* F2, I3, R1 & R1.2) |
| ***1.2 How will the data be collected, observed or generated?***<br>Questions you might want to consider:<br>- What standards, methodologies or quality assurance processes will you use?<br>- 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) |
| ***1.3 What documentation and metadata will you provide with the data?***<br>Questions you might want to consider:<br>- What information is required for users (computer or human) to read and interpret the data in the future?<br>- How will you generate this documentation?<br>- 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.<br>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) |

# 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?

## *Additional information 1.1*

***Dataset (data set)**: In a database, a data set is a set of data relating to a specific type of information.*
***Types:** raw data versus analysed data, observational data, experimental data, statistical data, survey results, etc.*
***Format:** Choosing the appropriate file format is crucial to access and potentially reuse your data.*
*Open file formats can be used by anyone because the file specifications are publicly available. On the contrary, proprietary file formats work only with software provided by the vendor, and when the software is no longer supported, files in this format are usually unreadable.*
*Therefore, as far as possible, **favor universal or open file formats to share or archive data, such as those recommended below**:*
*https://www.bium.ch/en/publication-open-access/data-management/#5*
*http://datadryad.org/pages/policies#formats*

## *Response suggestion 1.1*

This project will generate research **data on …… from …….** We list the specific types of data, equipment, software, data storage formats, and data sharing formats for all research data in this project in Tables 1 and 2 below.

***Table1: Types of data, formats and scales for new data***

| ***Types*** | ***Equipment*** | ***Software*** | ***data storage format*** | ***data archiving/sharing format*** | ***Volume*** |
|---|---|---|---|---|---|
| ***Microscopy images*** | | | | | |
| Raw data: microscopy cell images | Zeiss LSM 710 Quasar | ZEN lite software | .liff | .tiff uncompressed, JPEG2000 | 500 GB |
| Secondary data: 3D Z-stack reconstructions and processed images | | Imaris 7.2.1 software; Fiji/ImageJ; Adobe Photoshop CS5 | .ims, .tif series, .PSD | .tiff uncompressed, JPEG2000 | 1 TB |
| Analysed data: cell quantifications | | Imaris 7.2.1 software, Excel | .ims, .xlsx | .xlsx; .csv | 3 GB |
| Raw data :time lapse video microscopy | Leica SP5 | LAS AF Lite 4.0.11706 | .czi files;.Avi,.Mov | MPEG-4; Motion JPEG 2000 | 500 GB |
| Analysed data: tracking function | | Metamorph software 6.0 | .xlsx | .xlsx; .csv | 2 GB |
| ***Western Blots*** | | | | | |
| Raw data: cell images | | | | | 1 GB |
| Analysed data: quantification | | | | | 500 MB |
| | | | | | ***TOTAL =*** |

***Table2: Types of data and formats for reused data***

| ***Types*** | ***data storage format*** |
|---|---|
| ***Genetics databases*** | |
| RNA sequencing | |

***If animal research data*** **(mice, non-primate studies) is included, see additional DMP suggestions:**

***Types: Animal management data***

- Administrative data: animal experimentation Form A (authorization demand) and B (delivered animal authorization number XXXX).
- Animal husbandry data (temperature, humidity, epizootie, vaccination, feeding, water delivery, cycle day/night, husbandry statistics, animals provenance, mutants, etc).
- Animal physiology and behavioral follow-up data.
- Animal experimental design data (blinding, attrition, randomization, allocation concealment, power analysis, termination criteria., genotyping after animal delivery, genotyping before and after experiment, etc).

## *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?

## *Additional information 1.2*

- ***Recommended conventions for File Naming from UBC library for*** *http://researchdata.library.ubc.ca/plan/organize-your-data/*

*Short, descriptive filenames and a simple hierarchy of files make navigation and location easier.*
*Set up conventions for your project, document them for all other team members and be consistent.*

- ***Use a short unique identifier (e.g. Project Name or Grant #)***
  - *DO: CHHM*
  - *DON'T: Centre for Hip Health and Mobility*
  - *BECAUSE: Short filenames prevent the need for side scrolling and column adjustment.*
- ***Include a summary of content (e.g. Questionnaire or GrantProposal) as part of the file name***
  - *DO: FileNm_**Guidelines**_20140409_v01.docx*
  - *DON'T: FileNm_20140409.docx*
  - *BECAUSE: Files will be easier to find.*
- ***Use _ as delimiters. Avoid these special characters: & , * % # * ( ) ! @$ ^ ~ ' { } [ ] ? < > –***
  - *DO: FileNm_Guidelines_20140409_v01.docx*
  - *DON'T: FileNm Guidelines 2014 04 09 v01.docx*
  - *BECAUSE: Different computer systems handle special characters differently – filing order, etc.*
- ***Keep track of document versions either sequentially (e.g. v01, v02,) or with a unique date and time ( e.g. 20140403_1800)***
  - *DO: FileNm_Guidelines_20140409_**v01**.docx*
  - *DON'T: FileNm_Guidelines_20140409_**Review**.docx AND FileNm_Guidelines_20140409_**Investigation**.docx*
  - *BECAUSE: Two years from now, you won't remember what you meant.*
- ***Denote dates in YYYYMMDD format***
  - *DO: Use 20140403*
  - *DON'T: Use 04032013*
  - *BECAUSE: Computers sort YYYYMMDD in chronological order.*
- ***Make folder hierarchies as simple as possible***

- *DO: F:/* ***Env/LIBR/****DataMgmt_FileFormats_20140409_v01.docx*
- *DON'T: F:/****Environment/Library/Woodward/Data/Education/Materials/Draft/2014/04/-****DataMgmt_FileFormats_20140409_v01.docx*
- *BECAUSE: Complex folder hierarchies are harder to navigate and offer more opportunities for filing errors. System back-ups may take longer.*

- ***Tutoriel Mantra en ligne:*** *"Organizing data"*
- ***Tools for research data management and versioning***
  - *Consider using an electronic note book: http://www.labarchives.com/ and LIMS: https://www.genohm.com/slims/*
  - *Consider using version control software for data Data Versioning: https://osf.io/ or manuscript Data driven, executable articles in Authorea: https://www.authorea.com/product*

## *Response suggestion 1.2*

**Methodologies for data collection / generation**
***Data generation and analyses from research data summarized in Table 1 "Types of data, formats and scales".***

- **Procedure (and equipment) used to produce all datasets (linked to the grant application paragraphs?). Supplement the information**
- Data file naming and labeling (short unique identifier and summary of content) & organization into folders and sub-folders.
- Versioning of data subsets distinguished by a subscript (v01, v02, etc) with date attached to the file name.
- Filename will appear as **DatasetName** (short unique identifier_summary of content)_**VersionNo_Date** (YYYYMMDD).

***If animal research data*** is included add animal facility management software (PyRAT).
***If human research data*** is included add patient data management software (**supplement the information** e.g. soariane/data wharehouse, secutrial ….).

**Data quality and standards**
***Routine procedures for conducting high quality research, for publication in peer reviewed journals:***

- Instrument /machine calibration and control processes.
- Standard protocols in the field **(refs?)** for data collection ensuring reliability and consistency.
- Appropriate experimental design, data recording and data validation (controls, randomization/blinding, sampling/replicates, experimental versus hypothesis driven-protocol) ensuring internal validity.
- Training (techniques & data management) for lab staff to ensure high quality data (PhD students and continuing education via regular workshops in *Open Science* at the FBM doctoral school and FBM library).
- Peer review of data: regular supervision and lab meetings to ensure that procedures have been carried out correctly and that all data are properly recorded.

## *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?

## *Additional information 1.3*

- ***« Documentation, metadata, citation" tutoriel en ligne, Mantra.***

***Documentation:***

- ***Research data need to be documented at various levels:***
  - ***Project level****: what the study set out to do, how it contributes new knowledge to the field, what the research questions/hypotheses were, what methodologies were used, what sampling frames were used, what instruments and measures were used, etc. A complete academic thesis normally contains this information in detail, but a published article may not. If a dataset is shared, a detailed technical report will need to be included for the user to understand how the data were collected and processed. You should also provide a sample bibliographic citation to indicate how you would like secondary users of your data to cite it in any publications, etc.*
  - ***File or database level****: how all the files (or tables in a database) that make up the dataset relate to each other; what format they are in; whether they supercede or are superceded by previous files. A readme.txt file is the classic way of accounting for all the files and folders in a project.*
  - ***Variable or item level****: the key to understanding research results is knowing exactly how an object of analysis came about. Not just, for example, a variable name at the top of a spreadsheet file, but the full label explaining the meaning of that variable in terms of how it was operationalised.*
- ***Some examples of data documentation are:***
  - *laboratory notebooks & experimental protocols*
  - *questionnaires, codebooks, data dictionaries*
  - *software syntax and output files*
  - *information about equipment settings & instrument calibration*
  - *database schema*
  - *methodology reports*
  - *provenance information about sources of derived or digitised data*

***Metadata:***

- *The term metadata is commonly defined as "data about data", **information that describes or contextualises the data.***
- *The difference between documentation and metadata is that the first is **meant to be read by humans** and the **second implies computer-processing** (though metadata may also be human-readable).*
- *Documentation is sometimes considered a form of metadata, because it is information about data, and when it is very structured it can be. The importance of metadata lies in the potential for machine-to-machine interoperability, providing the user with added functionality, or 'actionable' information.*

*Metadata should be as complete as possible, using the standards and conventions of a discipline, and should be machine readable. Metadata should always accompany a dataset, no matter where it is stored.*

***Many academic disciplines have formalized specific metadata standards. You can consult them on:***

- *Digital Curation Center.*
- *https://fairsharing.org/standards/*
- *Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences.*

➢ ***Readme file for data sharing "What is a README file, and how do I make mine as useful as possible?"***

***Dryad guidance*** *https://datadryad.org/pages/readme*
*A README file is intended to help ensure that your data can be correctly interpreted and reanalyzed by others.*
*There are two ways to include a README with your Dryad data submission:*

- *Provide a separate README for each individual data file (view an example).*
- *Submit one README for the data package as a whole (view an example).*

***Dryad recommend that a README be a plain text file containing the following:***

- *for each filename, a short description of what data it includes, optionally describing the relationship to the tables, figures, or sections within the accompanying publication*
- *for tabular data: definitions of column headings and row labels; data codes (including missing data); and measurement units*
- *any data processing steps, especially if not described in the publication, that may affect interpretation of results*
- *a description of what associated datasets are stored elsewhere, if applicable*
- *whom to contact with questions*
- *If text formatting is important for your README, PDF format is also acceptable.*

## *Response suggestion 1.3*

**Data documentation**
To document our research, we will use laboratory notebooks, methodology reports, noting, experimental protocols, information about equipment settings & instrument calibration, database schema, DMP **(refer to the list above),** all essential components of data management.
***If animal research data*** *is included add electronical spreadsheets and animal facility management software (PyRAT).*
***If human research data*** *is included add patient clinical records ( e.g.* soariane/data wharehouse, secutrial ….).

**Metadata**

**Readme XML file** *https://www.bium.ch/en/publication-open-access/data-management/#4*
All our generated datasets will be accompanied by a **DataCite standard metadata in a Readme XML file.** The DataCite Metadata Schema for Publication and Citation of Research Data allow data to be understood and reused by other members of the research group and add contextual value to the datasets for future publishing and data sharing. We will generate the Readme XML file automatically using the DataCite Metadata Generator after filing the form requesting intrinsic metadata. The Readme XML file ensures compatibility with international standards and is human as well as machine-readable.

- **Mandatory elements** will include the file name for the results (field Title)/creators name (field Creator)/affiliation (field creator affiliation)/type of data (field Resource Type).
- **Recommended elements** will include key words (field Subject)/date of data creation (field Date)/link to electronic notebook (field Related Identifier)/details on the methodology used, analytical and procedural information, definitions of variables, vocabularies and units of measurement (field Description).
- **Optional elements** will include information on the size / format / version / access /funding.

**Files with internal metadata**: **supplement the information** e.g. microscope images contain a range of metadata (magnification, lens, zoom, gain, etc…).

**Metadata for publishing datasets on nonprofit unstructured data repositories such as Zenodo or Dryad (see section 4):**

- We will use similar standard XML metadata (DataCite Metadata Schema or Dublin Core) to publish and share unstructured datasets. The XML metadata that ensures machine readability / interoperability is generated after filing the repository submission form requesting intrinsic metadata. Metadata comprise, in addition, a persistent identifier, a publication date and conditions of access to (type of license) the dataset.
- We will upload, alongside the dataset, Readme files in text and/or XML (see previous paragraph) formats with more detailed information.

**Use of discipline metadata standards: Supplement the information (refer to the list above).**

| **2 Ethics, legal and security issues** | |
|---|---|
| ***2.1 How will ethical issues be addressed and handled?***<br>Questions you might want to consider:<br>- What is the relevant protection standard for your data? Are you bound by a confidentiality agreement?<br>- 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?<br>- 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) |
| ***2.2 How will data access and security be managed?***<br>Questions you might want to consider:<br>- What are the main concerns regarding data security, what are the levels of risk and what measures are in place to handle security risks?<br>- How will you regulate data access rights/permissions to ensure the security of the data?<br>- 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) |
| ***2.3 How will you handle copyright and Intellectual Property Rights issues?***<br>Questions you might want to consider:<br>- Who will be the owner of the data?<br>- Which licenses will be applied to the data?<br>- 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 reused. (This relates to the *FAIR Data Principles* I3 & R1.1) |

# 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?

## *Additional information 2.1*

## *Research on human subjects.*

- ***Research data confidentiality***

- *FBM/CHUV researchers conducting research on human subjects should consult the Commission cantonale d'éthique de la recherche sur l'être humain before planning research, data use and data sharing.*
- *Swiss researchers must comply with the general privacy protection law according to the Federal Act on Data Protection.*
- *They must also adhere to the Federal Act on Research involving Human Beings, created to protect the dignity, privacy and health of human beings involved in research. Data concerning humans made publicly available must remain totally confidential and be anonymized. Researchers should include a provision for data sharing in the informed consent.*
- *The following article may help you to prepare your data for self-archiving: Hrynaszkiewicz I et al. (2010) Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Trials 2010 11:9. http://www.trialsjournal.com/content/11/1/9*

- ***Services at FBM UNIL- CHUV***

- ***Centre de Recherche Clinique de Lausanne (CRC).*** *FBM-UNIL/CHUV researchers conducting clinical study research should consult the CRC as early as possible if planning a prospective clinical study; either interventional (trial) or observational. The CRC provides services spanning from concept/design to publication, including solutions for electronic data capture, data management and statistical analysis. For more information, please feel free to contact the CRC service.*

- ***Unité de valorisation des données et des échantillons biologiques (VDE).*** *FBM-UNIL/CHUV researchers conducting research on human subjects and using samples from the Biobanque Institutionnelle de Lausanne must consult the VDE service before planning research and data use, for proper data codification and de-identification. For more information, please feel free to contact the VDE service*

- ***Vital-IT*** *is a Competency Centre in Bioinformatics and Computational Biology that provides infrastructure, support and technological R&D for life science and clinical research in Switzerland and internationally. It is a platform helping Swiss and international researchers to manage codification /de-identification /anonymization, storage, analyses and publication of genomic, proteomic and metabolomic big datasets. For more information, please feel free to contact the service at the following address: http://www.vital-it.ch/about#contact*

- ***What do you need to consider when making your data openly available?***

***Dryad's guidance:*** *"When making data publicly available, any human subjects data must be properly anonymized and prepared under applicable legal and ethical guidelines. When de-identifying your data, both direct and indirect identifiers need to be considered.* ***Dryad does not allow direct identifiers, but a dataset may contain up to 3 indirect identifiers.*** *Direct identifiers include variables such as the participant's name, initials, email address and postal code; indirect identifiers are data that if combined might lead to identification (see additional guidance on human subjects data).*

- *https://www.bium.ch/en/publication-open-access/data-management/#10*

## *Response suggestion 2.1*

- **If animal research data (mice, non-primate studies,..)**

**Permission and data management of animal data throughout the project:**

- We will manage animal data in accordance with the « Ordonnance sur la protection des animaux, section 6. Documentation et statistique » (Art 143 Registre des animaux ; Art 144 Procès-verbaux de l'expérience)

- Form A and Form B on eTV (Swiss Federal electronical TierVersuch software)

**For research data sharing:** This project will not produce sensitive human personal data and will not necessitate specific limitations on data sharing.

- **If human sensitive data**

**Encoding/de-idenfication / anonymization of data; gain approval by ethics committees; formal consent agreements:** Manage the above three title points according to your situation. If you plan to consult the CRC, VDE or VitalIT for correct encoding/de-identification /anonymization of your data you should mention it in your DMP and ask the SNSF for cost coverage (refer to the information above).

**For research data sharing:** If some data cannot be shared publicly because they are bound by legal, ethical or confidentiality criteria, you should explain their specific constraints (refer to the information above). Data privacy for sensitive information related to personal and private information will be handled carefully by our service. Indeed, our service will divulgate and provide open-access to sensitive human data only after receiving **explicit consent of the individuals** as well as assuring **privacy protection through proper data anonymization.** We will ensure that shared data do not contain information that identifies, or that used in conjunction with other publicly available information to personally identify an individual, following *DRYAD's recommendations*.

## *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?

## *Response suggestion 2.2*

**Main concerns regarding data security, levels of risk and measures in place to handle security risks**

- Loss of data, Hard Disk Failure Rate / SSD Failure Rate (See 3.1).
- Global Malware Threats.
- Data Theft

**To overcome these issues we use an antivirus & adapted behavior:**

- Installation and regular updates of our antivirus application smb://nas.unil.ch/soft/Antivirus/ Symantec/@SEP
- Applying security patches (OS and applications)
- Activation of a Firewall
- No response to mail phishing
- Not opening attachments contained in unsolicited emails
- Inactivating Flash Player & Popups
- Backups

**Data access rights/permission to ensure the security of the data: supplement the information**

**Handling of personal or other sensitive data to ensure safe data storage and transfer**

- **If animal research data (mice, non-primate,...)**
  Animal administrative, husbandry, follow-up data and animal experimental design are sensitive data and are stored according to security procedure of institution.

- **If human personal data**

**At FBM-UNIL**

- **Data storage and regular back-up solutions at FBM-UNIL**
  See 3.1below

- **Data protection**

- **The Ci service at UNIL** provides appropriate protection of hosted data on its technical infrastructure, including the confidentiality and integrity of the data. The Ci will not modify or transmit the data except under application of Swiss law and in cases of support and prevention of technical problems. For information security, the **UNIL Institution mainly refers to the following international standards: ISO 27001** - *International information security standard*.
- **Extremely sensitive Data will be encrypted** by our lab members for storage on NAS-UNIL (for example with the "Cryptomator"), on USB/external disks or on laptops (Mac : FileWault, Windows : Bitlocker). This guarantees that external individuals do not have access to our documents.

- **Data sharing via SwitchDRIVE and data transfer via SWITCHfilesender.**
  *https://wwwfbm.unil.ch/wiki/si/start?id=en:public:documentation:services_switch*

**In the CHUV**
*http://tribu.intranet.chuv/content-19.06.2017_17_40.pdf*
*http://tribu.intranet.chuv/content-17.02.2016_11_26.pdf*

- **Data storage and regular back-up solutions at FBM-UNIL**
  See 3.1below

- **Data protection** *http://tribu.intranet.chuv/content-17.02.2016_11_26.pdf* and see : politique de sauvetage dans l'annexe B 4.6.
  The DSI-CHUV service provides appropriate protection of hosted data on its technical infrastructure, including confidentiality and integrity of the data. The DSI-CHUV will not modify or transmit data except under application of Swiss law and in cases of support and prevention of technical problems. For information security, the Institution mainly refers to the following international standards:
  - **ISO 27001** – *International information security standard*
  - **ISO 27002** - *Code of practice for information security controls*
  - **ISO/IEC 27002** - *Information security management in health*
- Sensitive data will be stored without encoding on the **NAS-CHUV (share disk for CHUV clinical data) with very high protection.** For additional analysis, sensitive data will be **encoded/ de-identified** by our lab members **(or VDE/CRC)** for storage on the NAS-CHUV, on USB/external disks or on laptops (Mac : FileWault, Windows : Bitlocker). This guarantees that external individuals do not have access to our documents.
- **Transfer** of encrypted data via *Filecare*.

## *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?

## *Additional information 2.3*

- *https://www.bium.ch/en/publication-open-access/data-management/#10*
- ***PACTT (Powering Academia-industry Collaborations and Technology Transfer).*** PACTT is the joint technology transfer office of the University of Lausanne (UNIL) and the University Hospital of Lausanne (CHUV). Contact us for commercialization of research results, protection and management of intellectual property, negotiation and management of

collaboration contracts with industry and other institutions, or if you need advice on the creation of a start-up company. For more information, please feel free to contact the service at the following address: *pactt.info@chuv.ch*

***Response suggestion 2.3***

## Intellectual property for datasets

Concerning content, individual content items are not copyrightable, while in most jurisdictions; such as the European Union and USA, data collection involving creativity is copyrightable. The structural elements of a database involving originality is not explicitly covered by copyright in Switzerland (see code des obligations https://www.admin.ch/opc/fr/classified-compilation/19110009/201401010000/220.pdf ).

**According to the UNIL-CHUV contract, any data created or modified in the course of our professional activity as a UNIL collaborator belongs to the UNIL (except for copyrights belonging to the creator).** This means that for any use other than scientific and academic purposes, UNIL approval is necessary (eg for patenting and commercialization) (see *Directive du Conseil de Direction UNIL-CHUV du 02.12.2009 relative aux contrats et à la valorisation de la recherche*).

## Open licenses for data

We will promote sharing and unlimited use of the data that we produced using **explicit licences**. For sharing our data, we will use a **creative common CC0 license as recommended by the UNIL University** or a **CC By license that is suitable for data sharing**. The CC0 license is a 'public domain dedication', i.e. a waiver of **all** our rights including those of attribution. The CC By license allows others to distribute, remix, tweak and build upon our work, even commercially, as long as they credit us for the original creation.

## Restrictions applying to the reuse of third-party data?

If you cannot share some data because of concern due to commercial and patenting issues, **you should contact the PACTT and explain the specific constraints (see above).**

| **3 Data storage and preservation** | |
|---|---|
| ***3.1 How will your data be stored and backed-up during the research?***<br>Questions you might want to consider:<br>- What are your storage capacity and where will the data be stored?<br>- 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.) |
| ***3.2 What is your data preservation plan?***<br>Questions you might want to consider:<br>- What procedures would be used to select data to be preserved?<br>- 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 reuse, 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. (This relates to the *FAIR Data Principles* F2 & R1.3) |

# 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 are your storage capacity and where will the data be stored?

What are the back-up procedures?

## *Response suggestion 3.1*

## DATA storage and regular back-up at FBM-UNIL

https://wwwfbm.unil.ch/wiki/si/fr:administratif:organisation:general:personnes

- The person in charge of the service's IT : Mathieu Noverraz
- **Appropriate data storage is the responsibility of the Principal investigator Prof. X or the Lab manager X**

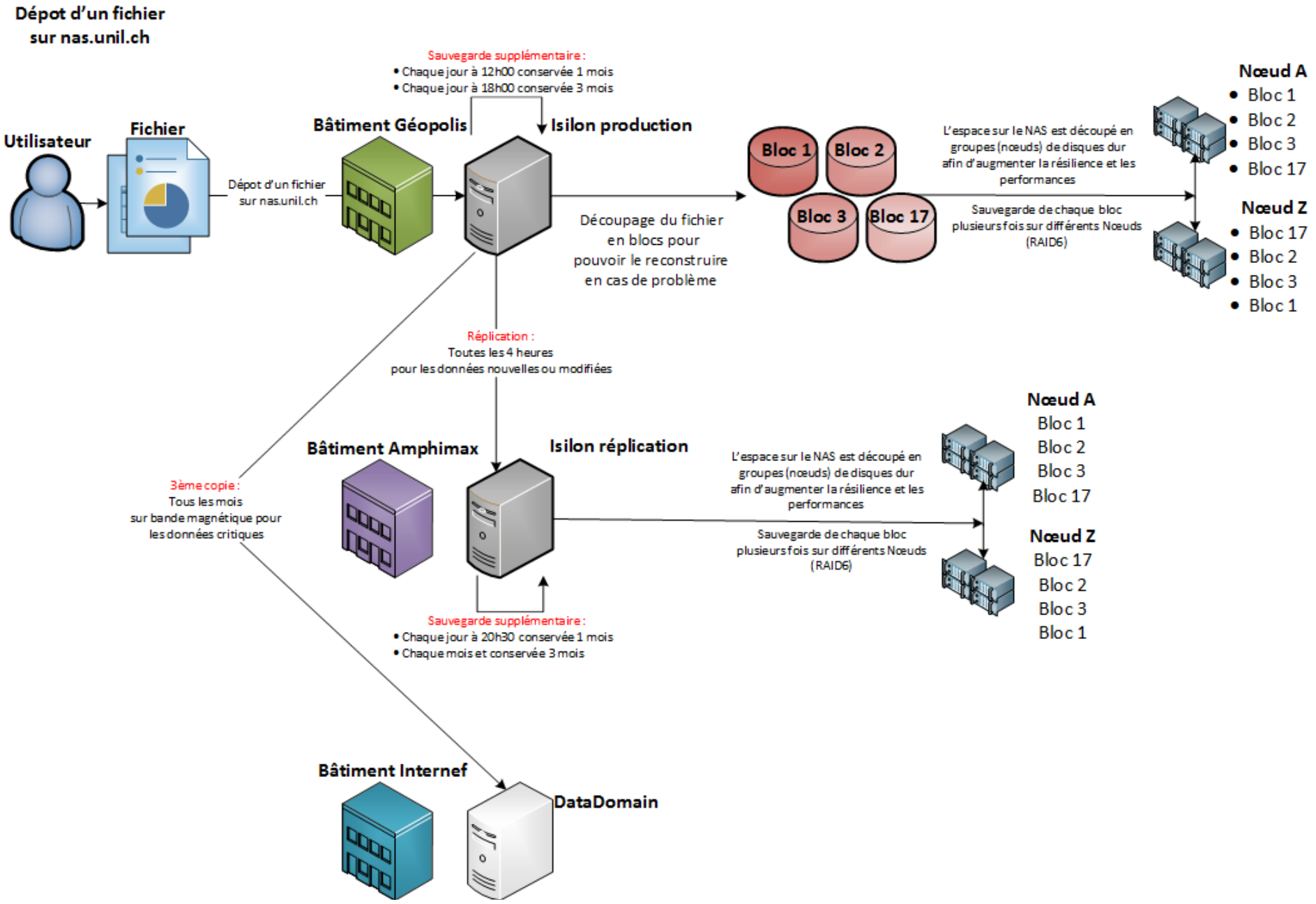
➢ **Desktop Backups via online Backups**

- CrashPlan PROe http://www.code42.com
- 100 GB per user
- Data are stored on the Ci local datacenter
- More info on:
  - services-au-personnel/sauvegardes/postes- personnels.html
  - https://wp.unil.ch/cinn/2015/04/le-quota- de-crashplan-voit-double/

➢ **Central Storage Tools in the IT service at FBM-UNIL**

**Isilon Network Attached Storage NAS (NAS EMC Isilon):**

- NSF, CIFS (Win/OSX/Linux);
- User-limited access rights management
- Access to files from anywhere in the world using a vpn;
- Upload: ~114 MB/s and Download: ~114 MB/s.
- Very high resistance to hardware and software failures;
- Security measures (protection against viruses, loss of data, etc.).
- **Data back-up and safeguarding**: **3 copies of the data**
  - ➢ Incremental backup every hour (snapshot);
  - ➢ Full daily backup of the data (2 times per day) on the NAS;
  - ➢ Replication of data;
  - ➢ Third copy for long-term archiving;
- **Costs (1 TB: 2 copies ~ 500.- CHF/year) (ask SNSF for cost coverage)**


Dépot d'un fichier
sur nas.unil.ch
Sauvegarde supplémentaire :
• Chaque jour à 12h00 conservée 1 mois
• Chaque jour à 18h00 conservée 3 mois
Utilisateur
Fichier
Dépot d'un fichier
sur nas.unil.ch
Bâtiment Géopolis
Isilon production
Découpage du fichier
en blocs pour
pouvoir le reconstruire
en cas de problème
Bloc 1
Bloc 2
Bloc 3
Bloc 17
L'espace sur le NAS est découpé en
groupes (nœuds) de disques dur
afin d'augmenter la résilience et les
performances
Sauvegarde de chaque bloc
plusieurs fois sur différents Nœuds
(RAID6)
Nœud A
• Bloc 1
• Bloc 2
• Bloc 3
• Bloc 17
Nœud Z
• Bloc 17
• Bloc 2
• Bloc 3
• Bloc 1
Réplication :
Toutes les 4 heures
pour les données nouvelles ou modifiées
Bâtiment Amphimax
Isilon réplication
3ème copie :
Tous les mois
sur bande magnétique pour
les données critiques
L'espace sur le NAS est découpé en
groupes (nœuds) de disques dur
afin d'augmenter la résilience et les
performances
Sauvegarde de chaque bloc
plusieurs fois sur différents Nœuds
(RAID6)
Nœud A
Bloc 1
Bloc 2
Bloc 3
Bloc 17
Nœud Z
Bloc 17
Bloc 2
Bloc 3
Bloc 1
Sauvegarde supplémentaire :
• Chaque jour à 20h30 conservée 1 mois
• Chaque mois et conservée 3 mois
Bâtiment Internef
DataDomain

## DATA storage and regular back-up at the CHUV

***Central Storage Tools at the DSI service CHUV***

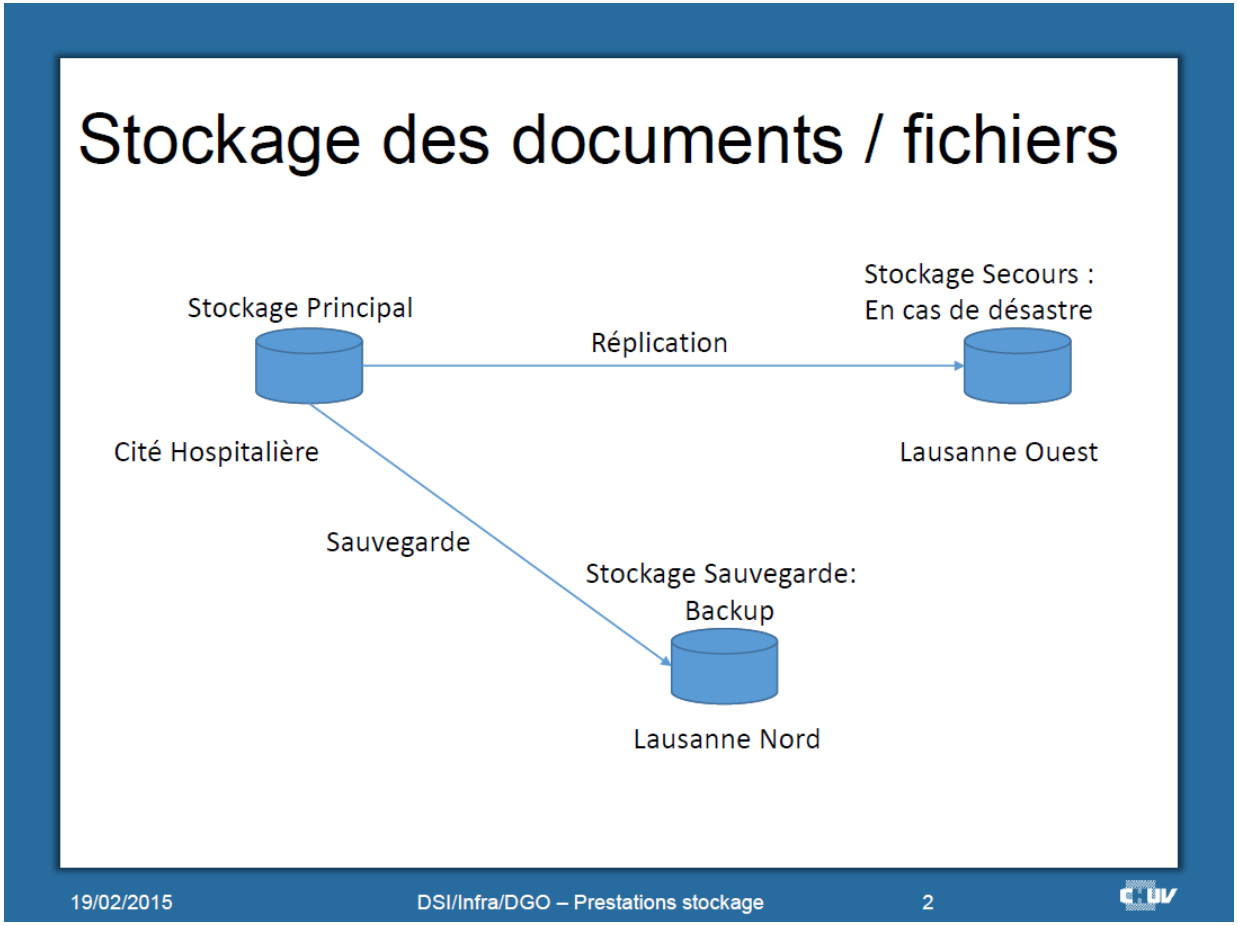
- The person in charge of service's IT at DSI: Nicolas Rosat
- **Appropriate data storage is the responsibility of the Principal investigator Prof. X or the Lab manager X**

➢ **Space for personal data (Accessible via unit H)**

- Also known as RPS (Répertoire Personnel Sécurisé)
- User-limited access rights management
- Immediate and regular access
- Backup to the backup system by the DSI
- Time snapshots
- Available soon after disaster
- Storage of personal information

➢ **NetApp Network Attached Storage NAS:**

- NSF, FTP, Win.
- User-limited access rights management
- Access to files from anywhere in the world using a vpn.
- Very high resistance to hardware and software failures.
- Security measures (protection against viruses, loss of data, etc.).
- Adequate protection of hosted data including confidentiality and integrity (see politique de sauvetage dans l'annexe B 4.6).
- **Data back-up and safeguarding: 2 copies of the data**
    - Incremental snapshot backup.
    - Full daily backup of the data on the Share Disk (1 x daily, usually overnight; retention 8 versions for daily backups, 2 versions for weekly backups).
    - Replication of the data on the backup disk (1 x daily, usually overnight; retention 30 versions for daily backups, 40 versions for weekly backups).
- **Costs (1 TB: 2 copies ~ 800.- CHF/year) (ask SNSF for cost coverage)**




**OR**

**DATA archiving, Data base development and data curation at Vital IT**
***https://www.vital-it.ch/services/infrastructure/storage***

**Vital IT supports different kinds of storage and data management systems:**

- Parallel and distributed **file systems** with more than 2.3 PB (Peta Bytes) of fast disk space
- **Hierarchical Storage Management (HSM)** systems with more than 3.6 PB (Peta Bytes) of space on magnetic tape (on-demand accessible from the cluster file system)
- Large-scale **relational database servers** (SQL-based)

**Costs for data archiving (1 TB: 2 copies 430 CHF/year) (ask SNSF for cost coverage)**

| Storage | 215 CHF per TB/year |
|---|---|

| Unit operational costs | Price per TB/year (CHF) |
|---|---|
| **U.1: Direct costs**<br>Depreciation (for compute nodes with a life span ≤ 4 years) & hardware of non-enduring value (≤ 20K CHF) | 51.00 |
| **U.2: Other direct costs**<br>Maintenance personnel and contracts | 139.00 |
| **U.3: Indirect costs**<br>Administrative personnel & central services | 25.00 |
| **U.T: Total costs** | **215.00** |

**File system types**

| | |
|---|---|
| **Archive** | Each file is archived in at least two copies.<br>**The storage amount is multiplied by 2.** |
| **Scratch** | A single copy of the file is stored in a RAID file system. |

***If animal research data*** **(mice, non-primate studies) is included, see additional DMP suggestions:**

- Animal administrative data Form A and B are stored for **6 years** by animal husbandry unit exceeding legal period of 3 years required by animal protection Art 144, and as electronical form in eTV for unlimited time period.
- Animal husbandry data PyRAT are stored **for unlimited time period** by animal husbandry unit exceeding legal period of 3 years required by animal protection Art 143.
- Animal experimental design and follow-up data are stored in NAS for unlimited time period and processed as dataset publication on long time guarantee preservation dataset publication platform Zenodo in case of a confirmatory paper to be published.

## *3.2 What is your data preservation plan?*

Questions you might want to consider:

What procedures would be used to select data to be pre- served?

What file formats will be used for preservation?

## *Additional information 3.2*

*https://uniris.unil.ch/researchdata/*

## *Response suggestion 3.2*

**Long-term data archiving: we will archive collected data on the UNIL-NAS research storage space (see Directive 4.2, art. 2.4).**

**Appropriate data archiving is the responsibility of the Principal investigator Prof. X.**

- **Published data:**
  Deposition and open data via Zenodo or Dryad repositories will ensure longevity of the data in the long-term. In addition, the UNIL-NAS or CHUV-NAS retains primary and secondary research data supporting published articles for **at least 10 years after publication.**

- **Unpublished data:**
  We will save unpublished high-quality final data generated during this project for re-use in our future projects. We will make some data available for use by researchers in future collaborations if our lab no longer uses it. Bad quality data will be permanently discarded at the end of the project.

| **4. Data sharing and reuse** | | |
|---|---|---|
| | ***4.1 How and where will the data be shared?***<br>Questions you might want to consider<br>- On which repository do you plan to share your data?<br>- 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) |
| | **4.2 Are there any necessary limitations to protect sensitive data?**<br>Questions you might want to consider:<br>- 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) |
| | **4.3 I will choose digital repositories that are conform to the FAIR Data Principles.** [CHECK BOX] | The SNSF requires that repositories are conform to the FAIR Data Principles (Section 5 of the guidelines for researchers, SNSF's explanation of the FAIR Data Principles).<br>If there are no repositories complying with these requirements in your research field, please deposit a copy of your data on a generic platform (see examples).<br>If no data can be shared, this is a statement of principles. |
| | **4.4 I will choose digital repositories maintained by a non-profit organisation.** [RADIO BUTTON yes/no]<br>→ If the answer is no: "Explain why you cannot share your data on a non-commercial digital repository." | The SNSF supports the use of non-commercial repositories for data sharing. Costs related to data upload are only covered for non-commercial repositories. |

# 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?

## *Additional information 4.1*

*http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx*
*https://www.bium.ch/en/publication-open-access/data-management/#2*
*https://www.bium.ch/en/publication-open-access/data-management/#5*
*https://www.bium.ch/en/publication-open-access/data-management/#7*

## *Response suggestion 4.1*

### Data sharing at the latest at the time of publication, via non-profit digital repositories

### Repositories making data available.

We will make supplementary files and key **datasets accompanying a publication to demonstrate reproducibility openly available in appropriate digital data repositories** that conform to the **Fair Data principles** and maintained by a **non-profit organisation**.

- We will share specific datasets via **domain-specific public repositories** (**see list recommended by PLoS**; **FAIRsharing.org**; **re3data.org**) **(state which datasets and data repositories)**.
- **Unstructured data** will be shared via the H2020 data repository **Zenodo-FBM/CHUV community** or **Dryad** **(choose one after checking the comparative table, or consider both and make a final selection depending on the journal's requirements)**. The following two data repositories fulfill biomedical journals' and **SNSF's requirements** (allowing publishing FAIR data, non-commercial).
- **Ask the SNSF for cost coverage**

### How we will make data available.

**We will use specific formats for data sharing** to ensure a file's preservation and re-usability (see table DMP1.1, BiUM recommended formats and Dryad recommendations). In some cases, we will upload the "Original" file along with the converted file.
**For field specific standards consult** https://fairsharing.org/standards/?q=
Datasets will be given a **Digital Object Identifier (DOI)** and **associated metadata**. The DOI corresponding to the datasets in the repository will be included in the article's reference list, allowing identification and access of any dataset in a publication.

### Visibility and valorization of datasets

We will also link DOIs to appropriate records in the University's publication repository **Serval**, to enhance the dataset's visibility. Metadata about datasets will be publicly searchable and discoverable and will indicate how and on what terms the datasets can be accessed. We will display information about datasets on the lab's webpage, on researcher profile pages on **unisciences**, as well as on researchers ORCID iDs, which will increase the visibility of the datasets.

### How will reuse of our data be valued?

**We will share data using CC0 or CC BY licenses** that will become citable products of research (see **Joint Declaration of Data Citation Principles**).

## 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)?

### *Additional information 4.2*

*More information for Open Data licences*
*https://www.bium.ch/en/publication-open-access/data-management/#10*
*https://www.bium.ch/en/publication-open-access/data-management/#11*

### *Response suggestion 4.2*

- **No specific limitations on data sharing**

We do not anticipate that this study will generate patentable data or proprietary data that would need protection.
**Prof. X** will decide when to publish and make Research Data accompanying the article publicly available including whether to supply research data to a new user.
**Use of the data**

- We will share data between lab members without limitation.
- We will make datasets openly available on appropriate digital data repositories (see 4.1) at the latest at the time of publication.
- We may make data available before publication upon demand by potential new collaborators.

**Restrictions or delays to sharing**

This project **will not produce sensitive human personal data** and **will not necessitate specific limitations on data sharing**. We will restrict data sharing only to ensure novelty of publication or **in cases of concern related to commercial and patenting issues.** Except for this, we will share data as widely as possible using CC0 or CC BY licenses.

➢ **Limitations on data sharing for human personal data.**

As mentioned in 2.1 and 2.2 above, we will share sensitive human data very carefully specifically due to legal, ethical and confidentiality issues.

**Prof. X** will decide when to publish and if it is possible to make Research Data accompanying the article publicly available including whether to supply research data to a new user.

**Use of the data**

- We will restrict encoded and de-identified Data access to specific lab members in charge of the project.
- We will make anonymized datasets openly available on appropriate digital data repositories (see 4.1) or give restrictive access after deposition on the NAS at our institution at the time of publication at the latest.
- We may make encoded data available before publication upon demand by potential new collaborators.

**Restrictions or delays to data sharing**

This project **will produce sensitive human personal data** and **will necessitate specific precautions and limitations for data sharing**.

- Our service will divulgate and provide open-access to human sensitive data only after the explicit consent of the individuals and only with privacy protection through proper data anonymization. We will ensure that shared data do not contain information that identifies, or that could be used in conjunction with other publicly available information to identify personally an individual, following DRYAD's recommendations. We will share data using CC0 or CC BY licenses.
- We will put in place restrictions on data sharing to ensure sensitive data protection if we are unable to assure perfect data anonymization. In this case the data will be stored on the NAS at our institution to insure perfect data protection.

## 4.3 I will choose digital repositories that 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** → If the answer is no: "Explain why you cannot share your data on a non-commercial digital repository."