NCCR Microbiomes

Research Data Management Strategy

NCCR Series 5

Coversheet

Title of the NCCR	Microbiomes
NCCR Director Name, first nameInstitution	Van der Meer, Jan Roelof University of Lausanne
Funding Phase	Phase I
Version (Date of submission)	Version 2 (5 September 2023)

Acknowledgements

We are greatly indebted to Valeria Di Cola (NCCR Robotics), Kaitlin McNally (NCCR Digital Fabrication), and Dominik Theler (NCCR RNA & Disease) for sharing their RDM Strategies with us. We have reused some of their language for our needs. We are very grateful to Carmen Jambé and Cécile Lebrand (UNIL-CHUV), Fabian Schmid and Julian Dederke (ETH Zurich), and Francesco Varrato and Eliane Blumer (EPFL) for helpful discussions and suggestions.

A. General Section

A1. SNSF requirements for research data

Except in the case of sensitive or protected data, the SNSF requires researchers to publish the data associated with journal articles in a **non-profit**, **FAIR-compliant public repository**, at the same time as the journal article. Published datasets are to be annotated such that external peers may verify, validate and reproduce the results.

A2. Data manager, data management organisation and budget

A2.1. Data manager

The data manager for the NCCR Microbiomes is Kendra Brown. The data manager activity is currently estimated as 0.4 FTE.

A2.2. Internal organization, roles and responsibilities

Each NCCR research group is responsible for the preparation, curation, documentation, and preservation of the group's data, and manages the submission of NCCR datasets to relevant repositories. Each research group informs the NCCR data manager about new journal articles and data publications, at a minimum once a year as part of their research progress report.

Based on the groups' information, the NCCR data manager maintains and updates an NCCR dataset index that is available on the NCCR Microbiomes website (https://nccr-microbiomes.ch/research/dataset-index/). If publications do not meet SNSF Open Science standards, the data manager advises the research group on a course of action to achieve compliance.

Additional responsibilities of the data manager are to inform the NCCR PIs about the RDM Strategy, organize trainings on data management and Open Research Data, and promote discussions among NCCR members about how to compare data across systems. The data manager enlists the help of RDM professionals in the home and partner institutions, as well as from different research consortia.

By signing this RDM Strategy, the NCCR Microbiomes PIs agree to follow its recommendations (the RDM Strategy becomes an appendix of the NCCR Internal Guidelines). If a PI does not want to comply with the RDM Strategy, they have the right to propose one that is specific for their group and to seek approval by the NCCR Microbiomes management and the SNSF. The data manager deposits the approved RDM Strategy on the NCCR Microbiomes internal cloudspace (Switch Folder), making it available to all NCCR researchers. The RDM Strategy will be made available for download on the NCCR Microbiomes website.

The internal data sharing policy appears in Annex A.

A2.3. Definition of the NCCR's Scientific Units

In Phase I, NCCR Microbiomes research is divided among six Work Packages:

- WP1: Human microbiomes
- WP2: Animal microbiomes
- WP3: Plant microbiomes
- WP4: Environmental microbiomes
- WP5: Synthetic and engineered microbiomes
- WP6: Computation

Research Data Management is conducted similarly across WPs, with the exception of WP1, which manages sensitive personal data from hospital patients and thus handles specific ethical issues. Consequently, the **Scientific Units** are defined as:

1) the WP that manages sensitive personal data (WP1)
2) the group of WPs that manage non-sensitive data (WP2–5).

A2.4. Data management budget

The salary of the data manager is part of the general management. The activity of the data manager is currently estimated as 0.4 FTE.

The charges for data preparation lie with each NCCR research group. Starting from Year 2 (2021-2022), we will offer NCCR researchers the option to pay for data deposition in non-profit, FAIR-compliant repositories if required. In addition, costs for workshops and trainings will be covered.

Budget Phase I (July 2020 – June 2024) with foreseen costs per year (in CHF).

Description	Year 1	Year 2	Year 3	Year 4	Years1-4
Data manager salary					
Data preparation	0	0	0	0	0
Data deposition in a repository	0	0	500	500	1,000
Workshop/training	0	0	1,000	1,000	2,000

A3. Intellectual property rights and copyright

A3.1. IP Rights

The intellectual property right procedures within the NCCR Microbiomes follow the NCCR Contract (§26), which states that in principle IP rights conform to the directives of the home institutions (UNIL , ETH Zurich) or the institutions of the PIs (EPFL, CHUV, University of Zurich, University of Bern). The SNSF, the NCCR Microbiomes and the individual host institutions maintain the right to publish the results of all associated research and data.

PIs specify IP rights obtained without the NCCR ('Existing background'), as part of the annexes of the NCCR Internal Guidelines.

A3.2. Copyright

Copyright and the choice of licensing remain with the individual creators of the work and data. With the exceptions of personal data, preexisting licenses or claims by any third party, the NCCR Microbiomes recommends to its members to grant permission to use their work and data by applying a CC0 or CC-BY 4.0 license. These licenses allow anyone to use and modify (CC-BY: so long as attribution is given to the creator). The licenses allow for commercial use.

B. General Policies for all scientific units (WP 1–6)

B1. Person responsible

The NCCR data manager (Kendra Brown).

B2. Data collection and documentation

B2.1. Description of the data collected, observed, generated or re-used

Researchers collect, observe, generate or reuse numeric (e.g., measurements, simulations, recorded

data) and textual (e.g., subjective findings, observations, reports) data, as well as code (e.g., software, programs), and various documentation data (e.g., images, video, mixed media).

Files include (but are not limited to): Office files (.docx, .xlsx, ...), OpenOffice files (.odt, .ods., ...), text files (.txt, .rtf, .pdf, ...), Image files (.tif, .jpg, ...), Tabular data files (.csv., ...), Code files (Matlab, R, Python, Perl, ...), Sequence files (FASTA, FASTQ, BAM, ...), Flow Cytometry files (.fcs), Mass Spectrometry files (.raw, .cdf, ...).

The size of data produced by all WPs collectively in Phase I is estimated to be a maximum of 200 Tb.

NCCR data may include protected data, such as data shared by a third party under confidentiality or under a specific agreement, data to be subject to a patent, or data to be subject to a license to a third party for commercial purposes (e.g. start-up).

Information about management of sensitive data, which concerns WP1 (Scientific Unit 1) can be found in Section B2.

B2.2. Documentation and metadata provided

To accompany each dataset published in a repository, researchers are to prepare a README.txt file of descriptive metadata. These metadata are required by the SNSF, in accordance with the FAIR Principles: https://media.snf.ch/qpI103GWawFodFF/FAIR_principles_translation_SNSF_logo.pdf. Researchers are encouraged to follow domain-specific metadata standards if possible.

1) General Information
    a. Title (a name given to the dataset or the research project that produced it)
    b. Creator (the name and affiliation and if available ORCID of the person who collected or contributed to the data)
    c. Date or period of collection
2) Data and file overview
    a. Abstract/description of the dataset (What does it contain? If applicable, a short description of each file)
    b. What metadata standard was used (if any)
3) Sharing and access information
    a. License (CC0, CC-BY, ...)
    b. Persistent identifier of the dataset (DOI, ARK, ...)
4) Methodological information
    a. Description of methods for data collection or generation (if available include links or references to publications or other documentation containing experimental design or protocols used)
    b. Description of methods used for data processing
    c. Additional definitions (e.g., of variable names)

If unique tools or proprietary software are needed to reuse the data, this should also be mentioned (in the section Methodological information) with, if possible, the tools or links to the provider made publicly available. For scripts or code, researchers will add the programming languages used, the version used (libraries, compiler, packages, etc.), and the license to reuse.

B3. Ethics and security issues

B3.1. Handling of ethical issues

No sensitive personal data will be generated in Scientific Unit 2. Anonymized personal data generated in Scientific Unit 1 may be reused in Scientific Unit 2. In these cases, PIs of Scientific Unit 1 are responsible for obtaining the necessary consent from the owners of the personal data, (see section C3.1).

PIs using experimentation in mice (Hapfelmeier, Hardt, Slack) are responsible for obtaining the necessary permissions from their cantonal authorities. They will forward a copy of authorizations to the NCCR data manager.

B3.2. Management of data access and security

If protected datasets are generated within Scientific Unit 2 (e.g. datasets relating to potential for patenting or copyright), they will be stored on group-specific servers with user-restricted access and backed up daily by institutional IT services.

B4. Data storage and preservation

B4.1. Data storage and back-up during the active research phase

Every research group defines its own data storage strategy backs up data regularly to avoid loss. Data are primarily stored on institution servers and institution computers with routine backup plans (in principle daily); the research groups rely on their institution's IT resources and services to ensure data storage and back-up according to their needs.

B4.2. Data preservation plan

Data related to publications will be preserved in a long-term data repository (see section B5).

The NCCR Microbiomes encourages the use of open formats, if possible, to facilitate the reuse of data. These formats include .txt, .pdf (PDF/A), .odt for text files; .csv for datasets; .tif, .jpeg, .png, for images. Additional information and recommendations about file formats can be accessed here: https://documentation.library.ethz.ch/display/RC/File+formats+for+archiving.

https://www.unil.ch/openscience/home/menuinst/open-research-data/gerer-les-donnees-de-recherche-research-data-management/archivage--partage.html.

For long-term preservation (≥10 years) of unpublished data of high-quality and value for reuse, NCCR groups can use the archiving services proposed by their respective institutions (e.g. UNIL DCSR).

B5. Public data sharing

B5.1. Public data sharing locations

The SNSF requires researchers to publish the data associated with journal articles in a non-profit,

FAIR-compliant public repository, at the same time as the journal article.

NCCR researchers enter their journal articles in their respective institutional repository, i.e., UNIL/CHUV-Serval, ETH Research Collection, EPFL-Infoscience, or Unibe BORIS.

For data publication, the NCCR Microbiomes recommends the use of Zenodo (www.zenodo.org) as a general repository for data and code related to publications. Zenodo is maintained by CERN, a non-profit organisation, and respects the FAIR Data Principles. It is available to researchers at any institution. Every publication made on Zenodo is assigned a DOI for unique identification and citation. Data deposited on Zenodo will be stored for the lifetime of the repository (≥20 years). The NCCR has created the NCCR Microbiomes community on Zenodo, to which NCCR researchers can link their submissions.

Alternative general repositories include the ETH Research Collection (https://www.research-collection.ethz.ch/), available to ETHZ researchers or the EPFL Academic Output Archive (ACOUA) (www.epfl.ch/campus/library/acoua-support), available to EPFL researchers. NCCR members may also choose to publish their data in other repositories, as long as they meet the SNSF requirements.

When compliant field-specific repositories are available, researchers are encouraged to publish their data there, in order to enhance their visibility and potential re-use. For example, sequencing data are best deposited in the European Nucleotide Archive (www.ebi.ac.uk/ena/browser/home) or the US National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/). Proteomics and metabolomics data may be deposited on the ProteomeXchange consortium (www.proteomexchange.org/) and MetaboLights (www.ebi.ac.uk/metabolights/). Image data may be deposited on BioImage Archive (https://www.ebi.ac.uk/bioimage-archive/). Software codes are often placed on Github (www.github.com) and Gitlab (www.gitlab.com) for version control during development, but to ensure long-term storage, the copy of the code associated with a publication is to be deposited on Zenodo (www.zenodo.org).

The data manager checks that all data related to an NCCR journal article are published in accordance with SNSF requirements. If the data are not properly published, the data manager will contact the author with advice for doing so. In the case of sensitive or protected data, the data manager will check that structured metadata are made available, with clearly defined conditions of data re-use.

C. Specific policies for Scientific Unit 1 (WP1)

C1. PIs concerned

CHUV: Greub, Guery, Resch

UNIL: Vonaesch

C2. Data collection and documentation

C2.1. Description of the data collected, observed, generated or re-used

Data generated or (re)used in Scientific Unit 1 include personal and personal sensitive data. Personal data is information related to a natural person (who is identified or can be identified directly or indirectly with another dataset). Personal data are considered sensitive when they are related to, e.g.,

the natural person's racial/ethnic origins, physical or mental health, sexual life, political opinions, or religious beliefs.

Data in Scientific Unit 1 can also include protected data, such as data shared by a third party under confidentiality or under a specific agreement, data to be subject to a patent, or data to be subject to a license to a third party for commercial purposes (e.g. start-up).

The size of produced data by Scientific Unit 1 in Phase I is estimated to be a maximum of 2 Tb.

C2.2. Documentation and metadata provided

For sensitive data, researchers are to publish descriptive metadata rather than the dataset itself. The README.txt file of datasets related to personal data, including sensitive data, will also include the following: the categories of personal data; the categories of recipients of the personal data; the used method of anonymization (if any); the access condition of the data (if the data are not released); the planned storage duration.

C3. Ethics and security issues

C3.1. Handling of ethical issues

The collection and processing of personal or sensitive data is subject to authorization by the ethical commission on human research of Canton Vaud (Commission cantonale d'éthique de la recherche sur l'être humain - CER-VD). Necessary permissions for experimentation in mice are obtained by the PI (Vonaesch) from the cantonal authorities. PIs in Scientific Unit 1 will forward a copy of authorizations to the NCCR data manager.

PIs of Scientific Unit 1 are responsible for obtaining the consent of the owners of the personal or sensitive data to the online and public publication of their data in encoded and/or anonymized form, as well as to the re-use of the published data by third parties.

C3.2. Management of data access and security

The personal data must be collected and processed in conformity with the cantonal law and the Swiss Federal law on data protection (and other laws if/when applicable, such as the EU-GDPR). Personal sensitive data will be stored on encrypted servers at UNIL (Vonaesch) and CHUV (Greub, Guery, Resch) with user-restricted access.

If protected datasets are generated within Scientific Unit 1 (e.g., datasets relating to potential for patenting or copyright), they will be stored on group-specific servers with user-restricted access and backed up daily by institutional IT services.

C4. Public data sharing

C4.1. Public data sharing locations

NCCR researchers in Scientific Unit 1 enter their published journal articles in UNIL/CHUV-Serval.

In the case of sensitive or protected data that cannot be published in public repositories, NCCR researchers are to publish structured metadata in an appropriate repository (see Section B5.1) with clearly defined conditions of data re-use.

C4.2. Data sharing constraints

In general, for sensitive datasets, metadata are to be publicly accessible, with clearly defined conditions of data re-use. However, if complete metadata cannot be published without compromising the anonymity of personal data, then only truncated metadata can be published. If the consent from participants is not obtained, no sample and metadata can be conserved, used or published unless authorized by the Ethical Committee under Art. 34 of the Federal Act on Research involving Human Beings.

Annex A. Internal data sharing policy

Each research group stores and maintains the data produced in NCCR projects according to their own internal procedures. In general, data produced within the NCCR that are not yet publicly available are stored on group-specific servers with automated backup plans (in principle daily); the research groups rely on their institution's IT resources and services to ensure data storage and back-up. Sensitive or confidential data (patients' information; current or future copyrights, licenses or patents) are protected by user-restricted access.

Collaborating laboratories develop their own operating procedure and policy regarding data sharing between the involved researchers, depending on how many parties are involved and of the nature of the collaborations (e.g., within or across institutions). The research groups, with the help of their institutional IT support, provide necessary accreditation to their NCCR collaborators in other groups and institutions.

The NCCR Switch drive Project Folder is available to NCCR researchers as a collaborative workspace and means to share data. NCCR members may contact the data manager to set up project space. Furthermore, to encourage data re-use within the NCCR, the data manager curates an 'internal dataset index' listing the datasets associated with the NCCR projects.

As agreed in the NCCR Consortium Agreement, data and knowledge produced and used within the NCCR are generally exempt of confidentiality among the NCCR members. Confidentiality is nevertheless maintained for defined existing background[1] (as specified in the NCCR Consortium Agreement) and for data which cannot be shared for ethical reasons (e.g., sensitive information related to hospital patients). With regard to the scientific community and the public outside of the NCCR, NCCR members are to respect the confidentiality of unpublished research data, especially data for projects with commercial potential.

[1] Existing background means information, materials and knowledge held by the NCCR member prior to the beginning of the project, as well as any intellectual property rights which are needed for carrying out the project.