Research Data Management Policy
Stratingh Institute of Chemistry

2025

Document Version History

Version	Date	Author	Description
0.1	27. February 2024	Jelte Nimoth	First Draft
0.2	9. January 2025	Jelte Nimoth	Version discussed with board
1.0	25. February 2025	Jelte Nimoth	Published version

Table of Contents

Introduction

The Stratingh Institute for Chemistry is a research institute embedded within the Faculty of Science and Engineering (FSE) of the University of Groningen (UG). It consists of several research groups performing research in molecular and supramolecular chemistry that spans across synthesis, catalysis, functional materials, bio-organic chemistry and chemical biology to systems chemistry and complex molecular systems.

The present institute policy is based on the FSE Research Data Management Policy (2022) as well as on the following regulations that corroborate the faculty policy:

- UG Research Data Policy (2021)
- Netherlands Code of Conduct of Research Integrity (2018)
- European Code of Conduct for Research Integrity (2017)
- European General Data Protection Regulations (GDPR)
- Faculty's policy and procedures on Research Ethics (under development)

The institute policy further specifies the faculty data management guidelines and outlines how scientists working within the Stratingh Institute are expected to deal with their research data during the time their research is being undertaken at the institute as well as how the research data is stored in the long-term (archival).

The Stratingh Institute recognizes the importance of FAIR (see Definitions) data management to increase the reusability of research data. The institute encourages researchers to adhere to the FAIR principles. This document functions as a guideline to help researchers to manage their research data in a FAIR manner throughout the whole research life cycle. One aim of this document is to guarantee the verifiability of the data generated from the research taking place at the Stratingh Institute.

The Stratingh Institute also endorses the principles of Open Science and encourages its researchers to make their data openly available (see Data collection, Storage, Archival and Sharing).

In line with good scientific practice, all those involved in data collection and management must meet the requirements and the procedures described in this document, keeping in mind that a common denominator in most cases of alleged scientific misconduct has been the absence of a complete set of verifiable data. The retention of accurately recorded and retrievable results is of utmost importance for the progress of scientific inquiry. Scientists must have access to their original results in order to respond to questions including, but not limited to, those that may arise from suggestions of impropriety. Moreover, errors may be mistaken for misconduct when the primary experimental results are unavailable.

Definitions

This section defines the terms used throughout the document.

Research Life Cycle

In the context of this document, the *research life cycle* refers to the whole research process. It includes all relevant components, from the initial planning phase to the final publication of the research results as well as its long-term preservation and reuse.

Data

In the context of this document, *data* is defined as any form of digital data that is generated during research. This includes raw (primary) data, processed (secondary) data, but also metadata, digital lab journal entries.
Code/software that is generated during research is also part of this definition of data and therefore affected by this guideline (see Software/Code as Data).
While non-digital data, e.g. physical/chemical/biological samples, is also part of this definition of data, it is treated separately as certain special considerations apply for these types of analog data (see Analog Data).
Moreover, special considerations for all types of textual data should be taken into account (see Special Considerations)

Metadata

Metadata can be summarized as *data about data*. Examples for metadata could be additional information about certain data (sets) that described its origin, underlying experimental parameters, information about processing that was performed, etc.
Most generated data already contain certain metadata information which is added by the system that generated the raw data, but metadata does not have to be stored directly within the data files itself. For example, providing a README file with certain information about a data set also constitutes a form of metadata.
The usage of metadata is an important pillar for the proper documentation of research data which helps to add an additional layer of structure to the data as well as increases its findability, and finally, reusability potential.

Publication

In the context of this document, *publication* is any document that can be accessed, electronically or in printed form, by a given set of users. A non-exhaustive list of publication types includes: BSc, MSc and PhD theses, internship reports, workshop/conference/symposium proceedings and abstracts, journal/magazine articles, books, book chapters, patents, technical reports, etc.

Data Management

As the name already suggested, the term data management summarizes all kinds of different tasks related to the handling of data. Within research, this includes the organization, storage, preservation and sharing of research data. In this case, the term *research data management*, short RDM, is usually used.

Data Storage

Data storage describes the way the research data is stored during the whole research life cycle. Long-term storage after a research project is finished, is also known as archiving/archival.

Data Repository

In the context of this document, a (*data) repository*, is a digital environment in which research data can be stored and shared. In contrast to normal data storage, data in a repository gets a digital identifier that makes it citable. Moreover, data in repositories is enriched with metadata and clear re-use requirements (licenses, etc.) are assigned.

Research Data Management Plan (RDMP)

A research data management plan, short RDMP, is a document that is created for a new research project which describes what data will be generated in the project and how it is handled during the research life cycle.

FAIR Data

FAIR data is data that fulfills the FAIR principles:

- Findability
- Accessibility
- Interoperability
- Re-Usability

FAIR data has the goal of maximal re-usability in mind, and it is important to note that FAIR data documentation should be in formats that are machine-readable.

Open Science

Open Science (OS) is a multifaceted approach that aims at making scientific research openly available for interested third parties, professional as well as amateur. OS encompasses multiple areas from open-access publishing to active engagement of society in the research project. In the context of this document, OS advocates to make research data openly accessible.

Roles and Responsibilities

Institute

- Ensures that the institute's RDM policy is available and communicated to institute members.
- Ensure that the institute's RDM policy is up-to-date and aligned with the respective policies on faculty and university level.
- Ensure that an RDMP template is available for the institute.

Principle Investigators/Group Leaders

- Ensures that an RDMP is created for all research projects (including BSC/MSc, PhD, postdoc, and research projects of visiting researchers).
- Ensures that the RDMPs are continually kept up to date and are adhered to by all project members.
- Ensures that students and staff who performed research projects in the respective groups properly store and archive their research data.
- Ensure that the RDM policy is followed and the data is properly stored after the end of a research project.
- Ensures that the RDMPs align with requirements of funding agencies, research consortia, private or any other third parties.
- Ensures that a data and material transfer agreement (D/MTA) is in place before data is transferred outside of the UG (see Special Considerations).
- Ensures that data is either destroyed or kept after the 10 year minimum archival period (see Retention Periods).

All Persons involved in the Research

- Ensure that research data needed to reproduce research findings are properly stored for at least 10 years starting from the end of the research project and made available upon request.
- Ensure that the stored data adheres to the FAIR principles.
- Ensure that a RDMP is in place which covers details on the type of raw/processed data (including storage and access regulations) that is generated in their research projects.
- Ensures that the data is properly documented and relevant metadata is stored in a suitable data repository.

Data Management Guidelines

The below guidelines further elaborate on the specific data management-related objectives that apply **for all research conducted at the Stratingh Institute**.
In general, all scientific data that underlies a publication has to be properly documented and stored for a minimum of 10 years.

Verifiability of Research Results

Results published in scientific papers must:

- refer to the data from which conclusions are drawn;
- show how these conclusions have been derived;
- allow verification of the original research data/dataset.

In order to guarantee verifiability of the data, all researchers within Stratingh need to decide, based on the nature of their research, what data needs to be collected and stored to guarantee the verifiability of their results, and what file format will be the most appropriate for storage (see the respective section about data formats).

Research Data Management Plan (RDMP)

Every research project at Stratingh is accompanied by an RDMP. The aim of the RDMP is that it functions as a reference to see which research is on-going at Stratingh, but also helps the research participants to estimate in advance what resources are needed for their research and which special considerations must be taken into account. It also provides a tool to guide the main researcher to think about a proper structure of the data. On the one hand, this increases the re-usability of the data. On the other hand, the requirement to think about the data and its structure also gives additional guidance to particular researchers . Finally, the RDMP is the go-to document for future researchers (e.g. other group members) that want to re-use (part of) the data. The information provided in the RDMP should therefore be written in a clear and understandable way to also benefit 3rd parties.

The information collection in the RDMP should therefore clearly state:

- Participants and affiliations.
- Topic of the research.
- Estimated type and amount of data generated.
- Used tools.
- How the data is documented.
- How the data is stored during research.
- How the data is archived after the research.
- Access to the data during/after the research.
- If software/code was produced as part of the project, how this is maintained and shared (see also Software/Code as Data).
- If there are any additional (ethical, legal) considerations for the re-use of the data.

For RDMPs, the following regulations apply:

- The Stratingh Institute provides a template that aligns with the present policy, and which should be used for projects at the Stratingh Institute.
- The UG RDMP web portal is the place to fill in and store all RDMPs. RDMPs can be printed/saved as PDF for additional purposes.
- The first version of the RDMP should be filled in **at the start** of the research project or no longer **than 3 months** after the project was started.

- The RDMP is **a living document** which should reflect the reality of the research project. If changes occur, it is the responsibility of both the project participants and the group leader that **a new version** of the RDMP is created.
- Upon request, RDMPs are made available to third parties.

Data Collection, Storage, Archival and Sharing

All digital forms of data and metadata, including code produced during research as well as digital lab journals, are collected, stored and archived in UG storage facilities that guarantee access and retention for the duration of a project, but also after researchers leave the university. For suitable storage solutions during and after the active research phase, the UG Digital Competence Centre (DCC) has an overview of available IT solutions that can be used.

All analytical, raw data should be saved as “read-only” and not altered. Instead, additional versions should be created during processing of these data during research. In case further adjustments of the data take place, it is recommended to store these modifications as additional versions, without deleting previous versions. In cases where this is not possible, for example due to huge file sizes of the different versions, researchers should judge for themselves if and how many different versions of (processed) data are kept. Moreover, it should be noted that for long-term archival of finished research projects, it is not needed to store all versions of processed data that were created during research. For the archival stage, it is important that at least the raw data, its documentation, as well as the processing that was applied to the raw data to derive the respective results are stored. Moreover, the main processed data should be also archived as long as this does not result in unnecessary high amounts of needed storage capacity.

For every publication, suitable information should be added describing which data was used in that particular publication and how it was processed (incl. details on the methods, tools, software, etc.) to draw the conclusions. This may already occur in the form of supplementary information according to the requirements of the scientific journal.

In line with the FAIR principles, data should be treated as a potential source of further reuse. Therefore, it should be clearly and properly documented at all stages of the research process, from data collection to final long-term storage (archival). Especially at the archival stage, it is recommended to **add a separate document to the data set** that summarizes important information of its structure, origin and processing.

Note:
Raw primary data and metadata that is collected and stored at an external institute falls under the responsibility of this external institute. Therefore, such data does not need to be deposited in the UG provided storage facilities.

Retention Periods

All forms of data underlying a publication have to be saved and archived for a period of 10 (ten) years, starting from the time the research project ends, but at least for 10 years after the last publication resulting from that project.
This applies for raw and processed data, code produced, but also all related digital and paper lab journals.
After the required retention period, it is the responsibility of the PI to decide if the data will be destroyed or kept beyond this period.
In special cases, the scientific director of the institute has the authority to demand that data is kept beyond the necessary 10-year period.

Naming of Data

While the institute does not enforce certain naming schemes, this document will outline some considerations that should be taken into account:

- If possible, try to avoid special characters (e.g. é!?*&à) in the name of the data objects.
- If possible, also try to avoid whitespaces in names of data objects.
- If numbering data objects, it is recommended to use multiple instead of single digits (e.g. 001 instead of 1) as this helps to prevent sorting problems.
- See if there are standards in your group or field that should be applied.
- Most important: Find a clear naming convention, document it and then stick to it.

Data Formats to use

While the institute does not enforce the usage of specific data formats, this document will outline some considerations that should be taken into account:
For the optimal (long-term) re-usability and interoperability of the data, it is recommended to use open (non-proprietary) and established file formats whenever possible.
DANS provides a list of preferred file formats which can be used when deciding for/against a format.
It should be furthermore considered that while open formats are the preferred selection, the standards that are used in a scientific field are often of a proprietary nature. In those cases, it is recommended to apply the standard formats that are used in the field. If the conversion is easily achievable, it is in this case recommended to add the data objects in its original, proprietary format as well as in the converted, open format to the saved/published data sets.

Data Access and Sharing

In general, data generated within the Stratingh Institute that is of interest to other scientific communities and/or external parties can be made available upon request, unless there are restrictions limiting availability of the data, such as publishing agreements, compromising intellectual property interests, or other third-party agreements as outlined in an NDA. If such restrictions exist, they should also be mentioned in the RDMP that accompanies the research project.

Data that relates to a publication (see definition above), should be made openly available with the following additional information:

- Metadata to increase its findability, but also clarify the content.
- Clear access and re-use information.

Researchers can openly publish data in one of the available scientific data repositories. The Stratingh Institute does not enforce one specific data repository, but the UG DCC has information about the archiving and publishing in research data repositories on its web page.
The institute furthermore does not enforce a specific license when data is openly published, but recommends the usage of one of the following Creative Commons (CC) licenses (also see Special Considerations):

- CC0 (public domain)
- CC BY
- CC BY-SA

Information on where the data is published (which data repository and how to access it) must be also recorded in the RDMP.

Recommended IT Solutions

The UG as well as the national Dutch organization for IT support in education and research, SURF, provide a selection of different tools that can and should be used throughout the different phases of a research project.
The UG DCC maintains a web page that lists the different solutions that are provided by the university and SURF.

Software/Code as Data

For some research, software or code, for example for processing of the data, are also generated. The current document acknowledges and emphasizes the treatment of this software/code as a product of research which is therefore also affected by the mentioned RDM regulations. Software/code of this type should be kept and documented following the best practices of reproducible workflow and adhere to the best practices defined by UG DCC.
If the software will be made open, then the repository should contain the appropriate license and a citation file.
While the Stratingh institute does not enforce a specific software license, it recommends one of the following open-source software licenses (also see Special Considerations):

- Copyleft: GNU GPLv3
- Permissive: MIT, Apache 2.0

And also refers to this helpful software license overview, and to the local data steward as well as the UG DCC for further points to get assistance while choosing a suitable software license (see Getting Help).

To facilitate the handling of software/code in a FAIR manner, the Stratingh Institute will provide a research software management plan (RSMP) as a part of the institute's RDMP which will ask about:

- Information on where and how the software code is stored and how access can be obtained.
- Information on the license that is chosen when publishing the software/code.
- Measures taken during the project to ensure long-term sustainability of the software developed in the project (maintenance).
- Measures that will be taken to support the software after the completion of the project.
- Resources that are needed to ensure the long-term usability and availability of the software, and how these resources are funded or obtained.

Analog Data

While analog data, for Stratingh especially chemical samples, are also affected by this data management policy, some special considerations apply for this type of data:

Chemical and Biological Samples
While chemical, and also biological, samples are one of the main analog outputs of a lot of research projects at the Stratingh Institute for Chemistry, the requirements that are listed in this document for digital data, do not apply for this kind of analog data at the Stratingh Institute for Chemistry.

This is due to:
- Chemical samples are often unstable in the long-term.
- There are limited physical storage capacities available.
- Keeping track of physical objects in the long-term can be very difficult.
- In most cases, there is not much benefit from also storing the samples along with the documentation of the experiments that they resulted from.

This means that chemical samples **do not** have to be:
- Stored for 10 years,
- made openly available.

Nevertheless while samples do not need to be stored for the long-term, it should be clearly documented for chemical samples from which experiments they resulted. Therefore, sample naming has to align with the naming scheme specified in the respective RDMP and lab journal entries as well as the sample descriptions should relate to each other.

It should be noted that the statement that chemical samples do not have to be stored long-term, does not mean that they cannot be stored long-term. Due to the wide variance of analog samples produced at the Stratingh Institute, it is up to the individual groups to decide how long they want to store their chemical samples.

Analog lab journals
Analog lab journals provide an essential part of the experimental documentation, and as textual data also provide a potentially valuable input for future usage (see Special Considerations). They should therefore be regularly, at the latest at the end of the research project, digitized and stored with the other data. If there are reasons not to add the journals to the stored data, the

stored data should contain a note that explains where the original experimental observations (lab journal) are stored and how they can be accessed.

Special Considerations

Textual Data

In the context of an ever increasing usage and impact of large language models and other machine learning techniques, all kinds of textual data that are created during research should be treated as a potential source of re-usable information, and handled accordingly.

Licensing

Upon publication of data (sets) and software/code, it is important that a license is specified to clarify the conditions for reuse. In case of data, Creative Commons (CC) licenses are the preferred way of licensing. CC licenses exist in different levels of reuse restrictions, starting from CC0 (donation to public domain) to the restrictive CC BY-NC-ND (attribution, non-commercial, no derivatives allowed). In the case of the publications of scientific data (sets), CC0 or one of the two CC BY type licenses (CC BY and CC BY-SA) are commonly selected with the latter two requiring attribution when the data is re-used. When choosing a license for data, it should be furthermore considered that CC0 has the maximal reuse potential and that even while it does not require attribution of the original authors, this is still a common good practice in the scientific community. When choosing a license, potential funder requirements should be taken into account. Also certain data repositories might require a certain license.
For software, it is equally important to specify an open-source software license, and it is here important that the selected license has a high compatibility with other open-source licenses to increase the reuse of the code/software, and therefore also its impact. The licenses recommended by Stratingh are some of the most commonly used open-source software licenses with a high compatibility with other common licenses. It should be noted that while CC licenses can be also used for software, it is discouraged to do so as this might result in compatibility problems and it is also not needed as specific open-source licenses exist for code/software. Furthermore, researchers should be aware of the distinction between permissive and copyleft software licenses. While the former basically allows all reuse of the code/software with no or very limited restions, copyleft licenses require the code/software which contains copyleft licensed parts to be also licensed under the same conditions.

Research using sensitive/personal Data or ethical Assessment

Research at Stratingh does not use sensitive/personal data and ethical assessments are therefore not needed for the research projects at the institute. In cases that this is needed, the institute policy refers back to the data management policy of the FSE which elaborates on these special considerations.

Data and Material Transfer Agreement (D/MTA)

A (D/MTA) is recommended every time data/materials are transferred to a receiving party outside of the UG, or when they are received from an external party by UG researchers. While the usage of a D/MTA is not mandatory, it is still recommended to have these agreements in place as they clarify responsibilities, questions about the ownership, and what can be done with the received materials/data.
A material transfer agreement is especially recommended if materials of high value are transferred and/or if the transfer of the material is connected to certain conditions of use.
A data transfer agreement is mainly needed if personal/sensitive data is transferred to another partner. In these cases, the Privacy & Security coordinators of the faculty are the first point of contact which can also provide templates.
In both cases, the agreements should be signed by all involved parties.

Getting Help

For help concerning data management, the local data support (data steward) of the institute is the primary point of contact. The UG DCC is the second line of support for data management related questions.