Institute for Molecules and Materials
Radboud University

RESEARCH DATA MANAGEMENT POLICY OF THE INSTITUTE FOR MOLECULES AND MATERIALS

1. Preamble

This document addresses the research data management policy of the Institute for Molecules and Materials (IMM). The IMM produces a diverse set of data, ranging from the synthesis of materials and molecules, mechanical drawings, toward software and characterization data of various materials and molecules. This document summarizes the types of data acquired and lists the policy adopted in the institute with respect to data organization, data archiving, data security, data access and privacy, open access data, and training of researchers in data management.

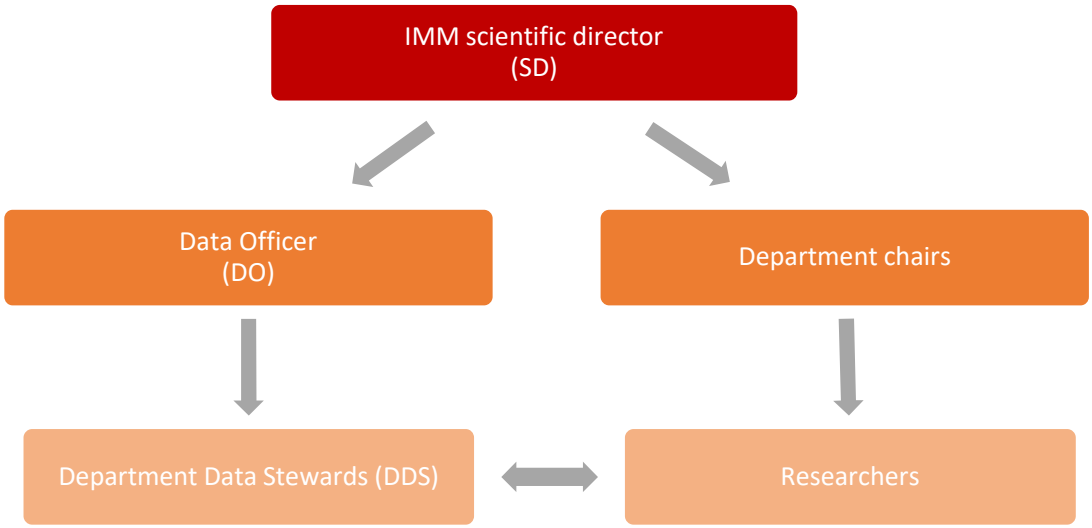
These policies have been developed and approved by the IMM board, and will be carried out and enforced by an institute's organizational structure, in which institute level data management is carried out by a data officer, in combination with appointed department data stewards which supervise day-to-day department specific data management. This policy addresses the growing movement toward organizing Research Data Management (RDM) both by the university[1], and the national governing bodies and funding agencies[2] to create a RDM policy concerning proper research data collection and production.

This policy is directed at addressing the needs in terms of data management of the IMM as a research institute. Research at IMM is situated at the interface of theoretical and experimental physics, material science, chemistry and biophysics/biochemistry. This broad and highly interdisciplinary domain is typically located in the so-called "long tail of data management".[3] The research is an array of many highly non-standardized research projects, which are only moderately data-intensive at the individual level, but the sum of which constitutes a large data volume. The other end of the spectrum contains 'big science' projects, in which researchers share a single large model or data set or facilities that produces vast, highly standardized data sets. In the "long tail" the total volume of data is large, but it is distributed over many individual researchers. This makes data stewardship and standardization particularly difficult to address. The scientific process is typically driven not by re-use of data itself, but by sharing of protocols, and (in some cases) samples.

[1] https://www.radboudnet.nl/onderzoek/onderzoek-visie-beleid-kwaliteit/onderzoeksbeleid/research-data-management/centraal-beleid/
[2] https://www.nwo.nl/research-datamanagement
[3] https://amolf.nl/wp-content/uploads/2017/09/DMPolicy_AMOLF_ARCNL.pdf

2. Organizational structure

Due to the diversity of the data produced resulting from the interdisciplinary nature of the institute, we have developed the following organizational structure:



- IMM Scientific director (SD): responsible for implementation of the RDM at the institute level

- Institute data officer (DO): responsible for implementation of data management policy throughout the institute, and manages the department data stewards. The data officer is also the link from the institute to the university.

- Department data stewards (DDS): the data stewards will be chosen within each department, appointed by the department head. The DDS needs to be employed at least on a temporary contract. Responsibilities include:
    - Updating the data management plan (DMP) annually (as defined below)
    - Training new researchers and students on RDM practices in the department
    - Compliance of the department with respect to the RDM policies of department and institute
    - Will initially draft and revise when necessary the data organization policy for the department, and define what types of data should be stored and archived.
    - Draft and monitor the archiving, security, and organizational rules related to data, as well as define what data will be version controlled in the group.
    - Attend meetings organized by the data officer, and occasionally attend RDM related meetings outside the institute

3. Types of data produced at the IMM

Data is defined as all recorded information, digital (and non-digital), gathered, collected, obtained or produced during or as a result of scientific research, and used for scientific development of theories, hypothesis testing, or validation of scientific findings, observations or conclusions.[4] Datasets at the IMM are typically closely linked to publications. Most publications are based on a dataset, which serves as a so-called "replication package". A replication package is the *sufficient* collection of research data (including measurements, protocols, metadata, processing methods) required to verify and replicate a scientific publication.[5] Replication packages are the basic units of information that are used to document the research process, and the basic units of storage of research data.

Since the IMM is an interdisciplinary institute comprised of disciplines ranging from physics to biological chemistry and medicinal chemistry, the datasets (replication packages) include a diverse range of data types, sizes and levels of maturity and uniformity. This data includes but is not limited to:

- Created and refined scripts and software over a diverse range of languages
- Raw images, binary files, and text files
- Processed data
- Handwritten and digital notebooks
- Documentation of experimental setups
- Mechanical drawings
- Calibration software
- Procedures on material preparation and operation of experimental setups
- Intellectual property
- Publications and grant applications
- Presentations
- Student reports and teaching materials

While this RDM policy is focused on digital data, our institute produces a diverse set of data. In the case that new types of data are produced or analysed, this list will be updated. The archiving and organization of such data, will be determined at the department level as discussed below.

4. Data organization

To enable effective and well-informed data management in the different departments within IMM, each department will have a document describing the organization of their data structure (DOS) and a policy document describing the way data gathering, analysis and storage is managed (DMP). Both documents will be updated regularly.

Data organization structure (DOS):
Each department, via the department data steward, will draft a document on the data organization structure (DOS) that illustrates how data (both digital and non-digital) is stored and archived in the

[4] Adopted from Donders RDM, https://www.ru.nl/donders/research/research-data-management/
[5] Adopted from: AMOLF RDM, https://amolf.nl/wp-content/uploads/2017/09/DMPolicy_AMOLF_ARCNL.pdf

department. This document will be updated annually, along with the DMP (as discussed below). A template of the DOS is illustrated in the appendix of this document. The following points of emphasis will be made in the department DOS:

- What data is stored, and what data is deleted. What is the policy on what raw data is archived. See the list with types of data above.
- Where data will be archived, for the defined 10 year period.
- Who has access to the data, and how data is it secured.
- How digital notebooks will be formatted, and how they relate to running projects.
- How raw data can be linked to notebook entries, and how it is connected to data which is archived on network drives.
- How manuscripts are stored, and the link to the raw data is made for each manuscript (for the IMM manuscript drive).
- Where scientific posters and presentations will be archived.

Data management plan (DMP):
Each department will draft a data management plan, which will be updated annually and managed by the DDS. Within the DMP, it will be listed what types of data are produced, by whom, and which projects this concerns for the year in question, as well as addresses the subsequent year. Moreover, the document will address any open access data, as well as what manuscripts have been produced and archived in the IMM manuscript server. This document will cover specific questions that are requested by NWO concerning data management for newly funded projects. The DMP (see attached template) will be made available to the IMM board each year, and reviewed by the data officer to confirm that proper RDM policies are in compliance.

5. Data storage and archiving

Data storage and archiving at the IMM occurs in various stages of the research process: from raw data collection, to the analysis/processing of data, to archiving of finalized datasets (replication packages) for replication and scientific integrity purposes. The storage of raw data typically is done on department network drives or local drives with regular backup to the network drives. The storage of finalized datasets (replication packages, which are typically linked to a publication) will be done on the Radboud Data Repository (RDR) or an equivalent discipline-specific repository that assigns a persistent identifier to a deposited dataset. Identifying acceptable repositories is up to the individual departments, in consultation with the data officer and scientific director.

By making use of dedicated repositories for publication data, IMM complies with the RDM policy of Radboud university[6] and most national and international funding agencies, who require that data (i.e., the replication package) are stored at the time of publication of the research at the latest, together with at least all the information necessary for potential reuse of data (metadata). The retention period for research data is a minimum of ten years. PI's at IMM can choose to make publication data available to anyone, available under a license (such as CC-BY-NC), or available only upon request and authorization. The first two options are already implemented for Data Sharing Collections in the RDR, but implementation of the latter option is still pending. Until this option has

[6] https://www.ru.nl/publish/pages/868512/rdm_policy.pdf

been implemented in the RDR, the IMM manuscript network drive (see below) will remain in place. PI's can use this CNCZ-managed network drive as a temporary alternative location for storage of publication data. When PI's decide to make their datasets available upon request and authorization, for example for reasons of Intellectual property (see "Data security and privacy" below), the departments involved must have a clear, written policy on applicable embargo periods or the conditions under which datasets will be shared. This policy will be included in the department DMPs.

Department network drives

Each department will be required to create a network drive(s), where various forms of data will be archived, being uploaded from local acquisition drives. The archiving here concerns all relevant (raw) research data, not only matured/final data which is utilized for publication. The use of the CNCZ network drive will be encouraged, as most groups already use these services. In cases where the data is not archived on the CNCZ network drive, the department head will exemplify to the data officer and IMM director, that the underlying rules are maintained.

All local drives, for example pertaining to measurement acquisition computers, containing research data will be required to sync and back up once a week to the network drives. Moreover, each network drive needs to contain some amount of version control for relevant data that requires version control (e.g., project-related notebooks, etc). The layout of the network drives will be specified by the department data steward in accordance to the data organization plan described above. New members will be instructed on the proper storage and archiving of such data.

IMM manuscript drive

The institute has a CNCZ-maintained network drive for publications and the corresponding datasets (replication packages). The philosophy will be to have a user restricted copy of every manuscript produced by every group, including a list of the raw data that was used to produce the publication. The drive will contain a folder structure for each department, and each department will be required to upload each published manuscript along with a link/description to the raw data used to generate the manuscript (except for manuscripts and datasets uploaded to the RDR). Access to the specific folders of manuscript drive, will be limited to the IMM director, and the pertinent PI's responsible for that manuscript. In cases of open data, the relevant manuscript folder can be made available with the approval of the pertinent PI.

The use of the IMM manuscript drive will continue as an alternative to the RDR, as long as the option of making Data Sharing Collections (DSCs) available on request and authorization only has not been implemented yet.

Radboud Data Repository

The standard location for deposition of finalized datasets, which typically belong to a publication, is the RDR (or an equivalent discipline-specific repository). The philosophy will be to have a registered copy of every dataset that has been used to produce a publication from every department in the RDR. The datasets in RDR will contain relevant metadata to make the dataset findable. PI's at IMM can choose to make the datasets in RDR publicly accessible or accessible only upon request and authorization. In the latter case, the department must have a clear, written policy detailing the conditions under which datasets will be shared. This policy will be included in the department DMPs. It is also possible to deviate from the general policy to deposit finalized datasets of publications in the RDR for reasons of data security/privacy or intellectual property (see "Data security and privacy"

below). These exceptions are listed and justified by departments in their DMP, according to the rules of compliance with the PI in question, the data officer, and the IMM scientific director. The metadata of these datasets and a link to the publication is always registered in RDR or on the IMM manuscript drive.

In accordance with the RDR guidelines,[7] different roles are distinguished in the data deposition process: data administrator(s), managers, contributors and viewers. Data administrators are able to create new datasets and assign one or more managers. They will be mandated by the scientific director. Dataset managers are typically PI's and data stewards (DDS), who can assign multiple contributors to a dataset. Contributors (researchers) can upload and organize data in a collection. The PI managing a dataset decides who will be granted access to the collection.

6. Data security and privacy

User access

Each department will define levels of user access connected to the network drives and repositories, used to archive various data. User access to various network drives will be controlled using the CNCZ user access control, unless other services are used and justified. Moreover, network drives will be created specifically for permanent staff, to store particular information related specifically to the department. On a case by case basis, network drives will be created for project-specific related data in each department, with case defined user access, in order to optimize security and privacy. These points will be defined in the organizational plan from each department.

Data privacy

A majority of the data produced by the institute does not concern ethical or privacy related issues, for example medical data of patients. In the unanticipated case that this is relevant, and not handled outside the institute (e.g. via the UMC) and should such issues become relevant, they will be handled case by case according to the rules of compliance with the PI in question, the data officer, and the IMM scientific director. Most data created by the institute concerns intellectual property, and access to such data will be limited and outlined in the department policies laid out by the data stewards in each department, and approved by the data officer.

In addition to user access control, all data acquisition computers related to data-taking facilities will be required to have limited external network access in order to increase security.

Intellectual property

Intellectual property is handled on a case-by-case basis by individual group leaders, as intellectual property comprises data both digital as well as non-digital data outside the scope of this RDM policy. However, all digital related data related to intellectual property will be properly archived and secured, in accordance to this policy and overseen by both the DDS and group leader, for each department.

[7] https://data.ru.nl/doc/help/faq/login-collection-roles.html

7. Open access and reuse

Publication-based datasets will in principle be made available via the RDR, but PI's can choose to make them available only upon request (see Data storage and archiving"). In those cases, the department will have a clear, written policy on how such requests are handled and what the conditions for sharing of data are. Exceptions to the deposition of datasets in RDR or the IMM manuscript drive for reasons of data security and intellectual property can apply, and should be specified by the departments in their DMP. Requests for sharing or reuse of such data are handled case by case according to the rules of compliance with the PI in question, the data officer, and the IMM scientific director.

The IMM will handle all other open data access and reuse requests (i.e., of non-published data) case by case, depending on the request. Raw data will in principle not be made available, but can be made available upon request. All raw data will remain on the network drives of the department and remain stored for a minimum of ten years.

All publications that are made open access, will be handled through the Library of Science (http://www.ru.nl/library/services/research/open-access/radboud-repository/).

8. Collaboration/data sharing outside the institute

At the moment, there is no strict policy on how to share data outside of the institute concerning collaborations. As data production and sharing at our institute does not concern ethical data, e.g. patient data, there are no strong security issues related to using commercial services. The use of Surfdrive and Radboud Data Repository (i.e., in the form of Data Acquisition Collections) will be recommended. Alternative solutions are being explored in collaboration with CNCZ to establish some secure collaboration service (equivalent to the currently used CNCZ servers), which meet the privacy and security requirements but also enable user access control to shared data (e.g. European/international collaborators).

9. Training

Training of proper RDM policies will be carried out within each department, by the DDSs. They will be responsible to train every new member on the following elements:

- Data organization trees
- Proper backup/archiving policies
- Data security and user access
- Use of version control
- Formatting of manuscripts

In addition to the DDSs, the data officer will be available for any general questions about implementing RDM.