Research Data Management Condensed Matter Theory Groups in the Institute for Theoretical Physics at the University of Göttingen

This Research Data Management concept is based on the Research Data Guidelines of the University of Göttingen, the DFG's Codex on Good Scientific practice (version 2022), and the RDM concepts of DFG Research Unit FOR 2414 and CRC 1073. Moreover, the standards of the scientific disciplines represented in the CMT groups are taken into account.

Milestones and What's New

Milestones

- 2019: Introduction of standard access to data, software etc. within scientific collaborations.
- 2020: Per default, processed data are made public.
- 2021: Introduction of a written CMT RDM concept.
- 2022: Evaluation of structure of Git repositories, documentation, etc.
- Mid 2022: Self-written software should per default be made public.
- Mid 2023: Evaluation of software publication.

What's New

- June 2022: New Section H on Documentation.
- Oct. 2022: Revision of Section C (mainly concerning publication of self-written software).
- Oct. 2022: Revisions in the general section Goals and Standards, e.g., concerning data reuse.

Goals and Standards in the Scientific Disciplines

The CMT research groups (Blöchl (TU Clausthal, Adjunct Member ITP), Heidrich-Meisner, Manmana, Kehrein) use a broad spectrum of numerical methods from the theory of strongly correlated electron systems (SCES) and the *ab-initio* description of solid state systems, including:

- Exact Diagonalization (ED).
- Lanczos/Krylov Methods.
- Density matrix renormalization group (DMRG) methods and more generally, matrix-product state approaches.
- Density functional theory (DFT).
- Semiclassical methods, as, for instance, used for the description of nuclear degrees of freedom or the truncated Wigner approximation.

In the scientific subfields represented in our CMT groups, data are produced to advance the scientific understanding of a problem via interpreting the data in the context of a scientific project/problem/question. Research data are generically produced to be part or the basis of a publication and usually have no standalone value in our communities. A reuse of research data is, in our scientific communities, not currently an established standard of how our science is advanced. Reuse mostly occurs for the purpose of independent verification of external scientist's own codes/algorithms or for benchmarks with other methods' results. An exchange of research data usually occurs only concerning processed data (see below).

Research data are usually produced by self-written programs and may be postprocessed by standard software. The data usually result from many independent simulations (both hardware

and software). Therefore, besides the long-term storage of data for the reproducibility of research results in general, we put a particular emphasis on preserving and documenting the underlying program codes. As a default, we aim at making complex self-written software that is essential to comprehend research results publically available. In order to account for competitive interests of the code authors (e.g., doctoral students), on a case-by-case basis involving all parties, decisions are made as to whether program code is made public or not. Such exceptions can arise to protect the competitive and career interests of early-career scientists and whenever the research is not published yet (as is often the case at the completion of a master thesis). Moreover, license restrictions for larger program packages (as often used for DFT simulations) need to be considered as well.

We established standards for the access to data (within collaborations and for external scientists), programs and documentation within scientific collaborations, both within the Göttingen CMT groups but also whereever external collaboration partners are involved. The RDM concept presented here also applies to Bachelor- and Master theses.

(National) Standards for RDM and Metadata in the form of a written RDM policy or guidelines do currently not exist in our scientific disciplines. This concept is adapted to the standards of the RDM policies of DFG Research Unit 2414 and of CRC 1073 and takes into account the recently published DFG codex on Good Scientifc Practice. A unique standard for data formats does also not exist in our communities, due to the highly specialized physical problems and topics.

We emphasize the importance of arXiv.org as the central international knowledge platform in our commnuities. The arXiv is the place where scientists initially search for our research results and data. The systematic publication of all manuscripts that are submitted for publication and partially also of research data (processed data) in common data formats are central aspects of our concept.

RDM Concept

A. Definitions: Data types

   We distinguish 4 types of data:

   a. Raw data: Data that are produced by a self-written program, a measurement, or a software.

   b. Processed data: Data that result from processing raw data. These are typically shown in publications or in a thesis.

   c. Codes: These are typically self-written programs with implementations of algorithms that are tailored for a specific problem. This category further includes larger programming packages and routines using commerical or open-source software packages.

   d. Meta data: These are, e.g., data format, location of data, access rights, reference to a publication.

B. Data storage

   a. During the duration of a project, data are stored on the already available media (GoeGRID Cluster/rocks, home directory, etc.), with automated back-up at the GWDG. Users of ITP issued laptops are responsible to ensure a regular back-up onto University or GWDG resources.

   b. For each project, a git-repository (or similar) hosted at GWDG (e.g., gitlab.gwdg.de) is used. All scientists that are involved into the project obtain access to the repository, including external collaboration partners. In the repository, you are recommended to store (specifics can vary in agreement with your supervisor and/or collaborators): (i) processed data, including everything that is shown or directly underlies a scientific publication/thesis, (ii) codes, (iii) documentation of the project's evolution, (iv) manuscripts

and figures. The repositories are maintained and remain accessible after the termination of a project.

c. For codes, versioning tools are used that are suitable for the specific project (e.g., the functionalities of a git repository).

d. After termination of a project (e.g., after a publication or submission of a thesis), the research data are moved into the ten-year-archive in the home-directory of the responsiple author (author of a thesis ot the person who produced and/or analyzed the data) at ITP or on the rocks cluster. These directories are automatically stored at GWDG for at least 10 years after the termination of a person's affiliation with ITP. Large amounts of data are directly transferred to GWDG by the responsible person (see instructions on ITP's compute wiki).

C. Public access to Research Data

a. Processed data are published after the conclusion of a project (usually after publication or submission of a thesis) on arXiv.org (along with a publication, as ancillary files), on Gro.Data, Zenodo, or via establishing public access to a git repository. If there are reasons why data are not made public (e.g., for a master thesis for which a later publication is intended), then this will be documented.

b. Raw data (e.g., DMRG wave functions, matrices in ED, eigenvectors in ED) are usually not made publically available, since there is no standard, need or benefit for their reuse or exchange in our scientific communities. Data will be made available upon request.

c. (Complex) Self-written codes are usually made public, e.g., in GRO.data, Zenodo, or a public git repository. Licensing restrictions and reusage rights for larger programming packages need to be observed, competitive interests of early-career reseaerchers need to be taken into account. Exceptions and their reasons will be documented.

D. Data Formats

a. For the publication of processed data, these formats should be used: ASCII, .csv, .XML, .json. In essence, non-proprietary formats should be used.

E. Data Management Plans (DMPs)

A uniform standard for DMPs is currently being developed for the CMT groups and will be introduced in 2023/2024. These DMPs will be stored in a central cloud (hosted at GWDG) that is accessible to all CMT groups.

F. Meta data

a. The git repositories and the location of published data should contain the following information: Reference to a publication (if applicable), name of the corresponding author, name of the institution where the data was produced, location of long-term storage (including instructions on the directory structure and how data can be retrieved), ORCID IDs of the authors (if applicable), reference to the arXiv identifier, name of contact person for data requests. Data are further complemented by metadata that are needed to understand scientific context and how the codes are started (e.g., input parameters), and how convergence/data quality is controlled.

b. Meta data for publications within DFG FOR 2414 are provided on the FOR 2414 webpage hosted in Frankfurt in a uniform format.

G. Training on RDM

   a. Annual joint seminar on RDM and on the specific implementation in the CMT groups for all group members, Bachelor and master thesis students.

   b. Access to RDM training offered by GAUSS graduate school at Georg-August-Universität Göttingen and by eResearch.

   c. Access to RDM training offered by FOR 2414 and CRC 1073 for funded persons and those involved in scientific projects within FOR 2414 and CRC 1073.

H. Documentation

   a. During a project: For every project (publication, thesis), the evolution of the project is documented in an electronic labbook, including, e.g., milestones, topics and results of discussions, decisions leading to changes in the methods, models, or parameter ranges. This documentation is stored in the git-repo that is accessible by all people involved in the project.

   b. Post-project documentation: The git-repo should contain: processed data shown in publication/thesis, software/codes, additional relevant processed data (e.g., convergence plots, other quantities, results for other parameters), the electronic lab-book, metadata. Information on: which data are made public (if not, explain why), and which data are stored long-time (if not, why not – discuss long-term value), whether software is publically available, where, and under which license (if not, explain why).

This RDM concept will continuously be further developed and adapted to new infrastructure and developments in our scientific communities.

October 2022