

Data Management Plan: Generic Template with Extension for High Performance Measurement and Computing

Template

Dr. Manuel Hora, ORCID 0000-0002-5156-0687

Technical University of Munich (TUM), University Library

Benjamin Farnbacher, ORCID 0000-0002-1489-6501

Technical University of Munich (TUM), TUM School of Engineering
and Design, Chair of Aerodynamics and Fluid Mechanics

Last update on December 22, 2021

Content

Preface	3
Template	4
DMP submitted by (name, address, contact).....	4
Submitted on (date and time)	4
Project name.....	4
Formal context of the project	4
Mandated policies (TUM only)	4
Principal investigators.....	4
Further researchers	4
Begin and end of project.....	4
Project description	4
Relevant research data.....	5
Reuse of existing data	5
Data format.....	5
Software for data processing	5
Size of data.....	6
Storage and backup.....	6
Documentation.....	6
Metadata.....	6
Reproducibility	6
Quality assessment	6
File naming and organization of data.....	7
Versioning.....	7
Data sharing within the project	7
Protection of sensitive data.....	7
Archiving.....	7
Publication	8
Responsibility for data management	8
Costs for data management	8
Last update of this DMP	8
Additional remarks	9
Additional HPMC-Information	9

Preface

This data management plan (DMP) template has been created within the context of the consortium *NFDI4Ing - the National Research Data Infrastructure for Engineering Sciences*. Target group are engineers conducting and post-processing high-resolution and high-performance measurements and simulations on high performance computing systems (HPC), represented by the NFDI4Ing archetype DORIS.

This template is based on standards and best practices at Technical University of Munich (TUM). Most of the queries are generic and can be transferred to different institutions and disciplines, even though most of the linked guidelines are institutional. Browsing appropriate institutional policies should accompany the draft of a new data management plan.

This data management plan template is generic for all kind of research domains. The special high performance measurement and computing (HPMC) queries are outlined on the last page.

Further information

- NFDI4Ing – the National Research Data Infrastructure for Engineering Sciences: <https://nfdi4ing.de/>
- NFDI4Ing-archetype Doris: <https://nfdi4ing.de/archetypes/doris/>
- Research Data Management at TUM: <https://www.ub.tum.de/en/research-data-management>

Template

DMP submitted by (name, address, contact)

Submitted on (date and time)

Project name

...

Formal context of the project

Please describe the context of the project briefly, e.g. is this project part of a larger research collaboration? Are there external partners? In case of a funded project, name the funding agency and the project ID.

Mandated policies (TUM only)

For TUM projects, the following policies apply:

<https://www.it.tum.de/en/projects/research-data-management/>

<https://www.ub.tum.de/en/open-access-policy>

https://www.forte.tum.de/fileadmin/w00bgt/www/Patent- und Lizenzbuero/tum_patentpolitik_ip_policy_en.pdf

Funding organizations might issue further policies. Please be aware that some funding organizations require a special format for DMPs, e.g. European Union.

Principal investigators

For each person, please state the name and institution.

Further researchers

For each person, please state the name and institution.

Begin and end of project

Please estimate the end date and update these dates, if necessary.

Project description

Please provide information on the scientific background and impact. Then state the project goals and outline the approach for this project.

Relevant research data

Please list all data, which will be collected, generated or processed in this project. Examples for data are:

- Surveys among patients
- Photographs of churches
- NMR spectra of proteins
- Source code for simulations
- Technical drawings
- Statistics on global CO2 emissions
- Annotated text corpora

Please consider all of these data in the remaining questions of this DMP, if applicable.

Reuse of existing data

- Please explain which research data is already available and will be reused for this project.
- Sources of data could include own previous research, work from collaborators, or published data.
- Please provide permalinks to the data (if published) or provide contact information of the data producer.
- Please check, whether you are allowed to use these data with regards to intellectual property rights.

Data format

If the data is available in digital form, please name the file formats. If the format is not common in your field of study, provide links to documentation of the format. For archiving, open and simple file formats are preferred. For more information, see:

<https://www.ub.tum.de/en/publishing-research-data>

If the data are not available in digital form, please name their form and explain, where and for how long they will be stored.

Software for data processing

Please name the software used for data processing, including the versions.

If you develop software, please consider the software as research data. Consider sharing scripts for data processing together with the data, if the license permits.

Size of data

If your estimated amounts of data exceed 50 GB and 10 000 files, please provide orders of magnitude for file numbers and size of the data sets.

E.g.: Ca. 100 000 .tif images, in total approx. 10 TB

Storage and backup

Please explain where the data is stored and which backup strategies are employed.

E.g.: Data are stored on a local server, with backups at LRZ using IBM spectrum protect.

Documentation

Please explain how the data will be documented.

E.g.:

- (Electronic) lab note book
- Log files
- Readme files
- Annotation within the data files

Metadata

If applicable, please provide links to used metadata standards. For lists of metadata standards, refer to:

- <https://nfdi.de>
- <https://rd-alliance.github.io/metadata-directory/standards/>
- <https://www.dcc.ac.uk/resources/metadata-standards>

If no suited standards are available yet for your discipline, please describe, which information will be provided to describe the data and facilitate further use.

Reproducibility

Please state, whether reproduction of the data is possible and outline the efforts and requirements for reproduction.

Quality assessment

Please describe the methods employed for quality control.

E.g.:

- Measurement of triplicates
- Cross-checking with other data
- Automatic plausibility tests (please provide a link)

- Peer review of data publication

File naming and organization of data

Please describe the conventions for file naming and explain, how your data will be organized.

E.g.:

Files are organized in directories according to specific research questions. Within these directories, files are organized with respect to method and accompanied by a readme file. File names follow the structure: YYMMDD_topic_method_version.

Versioning

Please describe how you handle different versions of files. Do you e.g. create full copies of old versions or use version control software like git? How do you document version differences?

Data sharing within the project

Please describe which data should be available when for whom.

Explain, how the data will be shared, e.g.:

- <https://workbench.ub.tum.de>
- <https://syncandshare.lrz.de/>

How do you prevent version conflicts in collaborative work?

Protection of sensitive data

Please explain, whether data are sensitive due to protection of personal data, confidential information from industry, concerns for public safety etc. If so, describe your means to protect the data.

E.g.:

- Password protection
- Encryption
- Anonymization

Archiving

Please explain, which data will be archived and for how long. State, which (intermediate) data will be deleted and how long after generation.

Consider, how time-consuming or expensive reproduction of data would be, which resources are available for long-term archiving and which requirements apply, e.g. rules for good scientific practice or laws covering protection of personal data.

Publication

Please state, which data will be published. Pay attention to requirements by the funding organization. Consider especially, which data are interesting for other research projects or difficult / expensive to reproduce.

For planned data publications, provide a link to the intended data repository. If data will not be published open access, explain the requirements for access. Further information can be found here:

<https://www.ub.tum.de/en/publishing-research-data>

Explain, under which license data will be published.

Will there be an embargo period?

Responsibility for data management

According to the TUM guidelines for research data management, the project head is responsible for data management and can delegate responsibilities for distinct aspects. If specific aspects will be delegated, describe the responsibilities and name the persons in charge.

Please name the persons responsible for:

- Data collection
- Documentation
- Quality assessment
- Archiving
- Publication
- Technical infrastructure
- Data protection

Costs for data management

Please state, whether you need to cover special costs or require additional personal resources for data management.

E.g.:

- Position for a data steward
- Costs for data storage
- Costs for data publication

In many cases, you can apply for funding for data management.

Last update of this DMP

It is recommended to check and update the DMP at least once per year.

Additional remarks

...

Additional HPMC-Information

Should already be included within the proposal / proposal can be attached to the DMP.

- Describe why a HPC-storage unit is necessary, and what should be accomplished using the resources
- Amount of storage space requested
- Number of files to be stored
- Time period for the storage
- Regular bandwidth requirements (MB/s), expected average data transfer in 24 hours (TByte/day)
- Special requirements regarding bandwidth for data access (e.g. high fluctuation in data transfer amount; overrun of provided average bandwidth; relatively high bandwidth in relation to storage space)

Acknowledgement

Manuel Hora and Benjamin Farnbacher would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.