Contract between doctoral student and supervisor

W2W

(2019-2023)

§1 Purpose of the contract

The contract between PhD student and supervisor(s) enables their relationship to be transparent. The planning and implementation of the PhD research project should be designed by the supervisor(s) and the PhD student together, in order for the project to be completed within a reasonable timeframe and with high quality. The individual situation of the PhD student should be considered. The successful completion of the PhD thesis is not guaranteed by the signature of this contract.

§2 Persons involved

A contract is concluded between the following persons:

..... (PhD student)
..... (main supervisor)
..... (other supervisor, if applicable)
..... (other supervisor, if applicable)

§3 PhD project

The PhD student intends to work on a dissertation, which title is
.....
.....
at the …..... [Faculty]
in the ..... [University].

The dissertation is planned to start on
..... (start: month/year)
and to be finished on
..... (expected end: month/year).

§4 Duties and responsabilities of the supervisor

(1) The supervisor makes sure that the PhD student is familiar with the current rules applying to PhD programs at the host University.

(2) The supervisor strives to provide the appropriate working conditions to the PhD student.

(3) The supervisor commits to regularly and professionaly avise the PhD student. The supervisor also commits to attend meetings regulary (at least every three months) about the work in progress of the PhD student, taking into consideration the work plan and the work schedule.

(4) The supervisor encourages the PhD student to work independently. The supervisor also supports the PhD student by providing access to national and international scientific environments, by introducing her/him to working groups and scientific networks, by encouraging her/him to take part in workshops, conferences, summer schools, by helping her/him to prepare presentations, by providing her/him with information on possibilities to publish articles and by helping her/him in the writing process.

(5) The supervisor supports the PhD student regarding her/his career plan and mentions possibilities for further disciplinary and interdisciplinary qualification.

(6) The supervisor agrees with the PhD student on the disputation procedure.

(7) The supervisor assesses the work submitted by the PhD student promptly and in a neutral way.

(8) The supervisor strives to find reviewers outside of the host institution for the PhD student. These reviewers should, if possible, be principal investigators of the Transregional Collaborative Research Center „Waves to Weather“ (W2W; SFB / TRR165).

(9) The supervisor will report on the progress of the PhD student, as part of the annual project report.

§5 Duties and responsabilities of the PhD student

(1) In agreement with the main supervisor, the PhD student produces a detailed and structured workplan and work schedule. The PhD student must inform the supervisor in case major changes have been made to the work plan and work schedule.

(2) With the prior agreement of the main supervisor, the PhD student attends specific courses and multidisciplinary courses.

(3) The PhD student regularly reports (at least twice a year) on the work in progress to the supervisor(s). The report (approximately 1-page long) contains a description of the achievements since the last report or since the start of the PhD, the overall progress on the thesis, the past and planned participation in lectures, conferences, guest lectures, doctoral days and specific workshops, and the interactions between the PhD student and the second supervisor. In addition, the PhD student submits parts of the results (e.g. chapter of the thesis, draft of article) to the supervisor(s) following the work plan and the work schedule.

(4) The PhD student strives to present her/his scientific results to the international community by publishing articles in peer-reviewed journals and by presenting these results at confe-erences.

(5) It is recommended that the PhD student use the W2W templates for oral and poster presentations. The templates are available here:
https://gitlab.physik.uni-muenchen.de/w2w/presentation_templates.
In case university templates must be used, the W2W logo should appear on the slides and on the poster.

(6) The financial support of W2W must be acknowledged in publications. The following phrasing is required: *"The research leading to these results has been done within the subproject "<name of the subproject>" of the Transregional Collaborative Research Center SFB/TRR 165 "Waves to Weather" (www.wavestoweather.de) funded by the German Research Foundation (DFG)."*

§6 Involvment in the Collaborative Research Center „Waves to Weather" (SFB/TRR165)

(1) The PhD thesis will contribute to the project ….....….....
….....
within the Transregional Collaborative Research Center „Waves to Weather" (W2W).

(2) In addition to the regular meetings with her/his supervisor, the PhD student presents her/his work in progress to the project group at least once a year and receives feedback from the project members in order to improve the PhD project if possible.

(3) Measures and regulations to help combining family and scientific activities are implemented within W2W, within the host institution and within the partner institutions. They apply, and are available, to the PhD student.

(4) All project scientists must comply with the W2W Data Management Plan.

(5) The agreements made within the doctoral program remain unchanged.

§7 Conflict situations

In case of conflicts, the PhD student and the supervisor(s) can ask for advice and support from other persons involved in W2W, e.g., Early Career Scientists committee members, Equal Opportunity committee members, co-speakers, and Early Career Scientists contact person in the Steering Group.

§8 Compliance with the good scientific practice principles and ethical guidelines

The persons signing the present contract agree to comply with the principles of good scientific practice and ethical guidelines.

Location, date and signature

..... (location, date, PhD student)

..... (location, date, main supervisor)

..... (location, date, other supervisor)

..... (location, date, other supervisor)

Attachement:
- Work plan
- Work schedule
- W2W Data Management Plan

Last updated: October 2019

Data Management Plan for Waves to Weather

1. Administrative Data

Funding Agency	DFG
Grant Reference Number	SFB/TRR 165
Project Name	Waves to Weather
Project Description	See https://wavestoweather.de
PIs	See https://wavestoweather.de/people/index.html
Project Data Contact	Robert Redl <robert.redl@lmu.de>
Date of First Version	18.09.2019
Date of Last Update	23.10.2019
Related Policies	• DFG Leitlinie zum Umgang mit Forschungsdaten[1] • DFG Leitlinien zur Sicherung guter wissenschaftlicher Praxis[2]

2. Data Collection

2.1. Sources

W2W is a project in theoretical meteorology. Data that is collected or produced consist of input or output data of meteorological models or meteorological observations collected by third-parties, which are used for model evaluation. In addition, program code is used or created, which is also to be regarded as a research result. Within this document, the following types of data are distinguished:

Type	**Description**
External Data	Data created by third-parties (e.g., weather services). These data are collected to ensure easy access for all project members and to ensure preservation of data used in publications, if this is not guarantied by the source[3].
Internal Data	Data created by numerical experiments of individual sub-projects of W2W. Data in this category are usually a derived product of code and external data and thus reproducible if both are preserved.
External Code	Program source code provided by third-parties. This includes weather models like COSMO or ICON.
Internal Code	Program source code produced by project members. These can be actual software projects (e.g., program libraries) as well as scripts used to create publications.

2.2. Formats

Data (internal and external) are stored wherever possible in one of the two self-describing formats netCDF or GRIB. These formats are well-established within the meteorological community and supported by numerous software tools, which ensures long-term accessibility. Open-source software libraries for reading and writing both formats exist. Beyond that, GRIB is a standard data exchange format of the World Meteorological Organization. In cases where usage of these formats is not

1 https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf
2 https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf
3 Due to large data volumes, weather services usually have short retention times for model output.

possible, e.g., when data are provided in different formats and conversion is not feasible due to the large amount of data, additional metadata along with code to read the data are added.

Code (internal and external) is stored in GIT-Repositories. Text files should be stored in UTF8 encoding to ensure portability between computer platforms.

2.3. Structure

In contrast to the climate community, where a high level of prescribed structure is common, we here only recommend a basic directory structure and leave room for adaption for the large variety of different projects planned within W2W. Our basic structure follows the standard of the Climate and Environment Retrieval and Archive, version 2 (CERA2) of the German Climate Computing Center (Deutsches Klimarechenzentrum, DKRZ) Long-Term Archive[4]. The following components are recommended for internal data:

```
/Project/Experiment/[Dataset-Group]/Dataset/[Additional-Info]
```

Component	Description
Project	Name of the sub-project responsible for the data (e.g., “A1”). For collaborations this may include multiple projects (e.g., “A1+A7”). This top-level folder is created be Z2 and permissions are granted for project members.
Experiment	Name of a numerical experiment within a project (e.g., “nature-run”). An experiment may also be a work-package of a sub-project.
Dataset-Group	An optional level of organization. Datasets that belong together within one experiment may be grouped in a folder. This might be used to group all “input” and all “output” Datasets together.
Dataset	A folder containing the actual data files. A folder with source code is also considered to be a dataset.
Additional-Info	An optional folder with documents, scripts, or plots which add to the understanding of the data.

External data are kept within the same folder structure, but with a modified meaning of the first components:

Component	Description
Project	Type of the collected data: “forecast”, “observation”, …
Experiment	Name of the Dataset, e.g., “ERA5”, “IFS-oper”, …

2.4. Data Live-Cycle

A central element of the practical data management is the software iRODS[5]. It provides the possibility to attach metadata to datasets, makes datasets searchable by metadata (e.g., key-words), provides a simplified interface to local tape archives, and enables data-exchange between the participating sites. To work with data, project scientists are provided with a per-project working directory, which can be organized as outlined above. Data within this working directory are not considered final and is used for ongoing analysis and modifications. Once a dataset or experiment is finalized (i.e., it is ready for publication or internal sharing), it is archived along with metadata. A typical workflow for a model experiment may contain the following steps:

4 https://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf
5 https://irods.org

1. Collection of required input and evaluation data. The iRODS client may be used to find the data in our archive. Datasets from the archive may only be linked into the working directory and not copied (unless they are only available on tape).
2. Quality control of retrieved data.
3. Perform the actual numerical experiment. This will create one or more new datasets within the working directory (e.g., model output, plots, etc.)
4. Create appropriate metadata and documentation. A third person should be able to understand how the experiment can be repeated. The usage of combined code and documentation (e.g., in format of Jupyter Notebooks[6] or similar technologies) is encouraged but no strict requirement.
5. Archive the data using iRODS. This is usually done together with a publication of the scientific results. iRODS will transfer the data to the tape archive of the local computing center.
6. If feasible, publish the data and related code following the FAIR principles (Findable, Accessible, Interoperable, Reproducible). DFG guidelines for good scientific practice are applied (guideline 13, see related policies).
7. Delete data from the working directory when it is not needed for subsequent experiments or other projects.

2.5. Version Control

Without having multiple versions of datasets, the total amount of data is estimated to be about 3 PB. Therefore, versioning of data is considered not to be feasible. Source code (internal and external) is version controlled using GIT. Scripts and settings (e.g., namelist files), which are required to repeat experiments, are considered as source code datasets and are versioned accordingly using GIT. Repositories are managed on our Gitlab instance[7], which is hosted by the administration group of the Faculty of Physics at LMU.

3. Documentation and Metadata

3.1. Mandatory Information

The most important metadata are collected within the actual data files using a self-describing format. In addition to that, the CERA2 standard is used for the definition of mandatory information. These metadata are required when datasets are stored within the iRODS archive:

Attribute	**Description**
Entry Name	A name that ensures unique identification. The combination of Project + Experiment + Dataset-Group must be unique, individual parts may not be unique.
Summary	Short Summary, external links or references are allowed.
Keywords	A list of keywords that helps others to find the dataset. One of those should be “Waves to Weather (SFB/TRR 165)”.
Authors	The main researchers involved in producing the data, or the authors of a data

6 https://jupyter.org
7 https://gitlab.physik.uni-muenchen.de, alias https://gitlab.wavestoweather.de

Attribute	Description
	publication.
Investigator	Responsible contact person for the data in question.
Metadata	Person responsible for the metadata if different from the investigator.
Data Description	The data format as well as additional information about the data if none of the formats netCDF or GRIB are used.
License	Mandatory for external data. Includes information of whether a dataset may be (re-) shared or made available to the public.

CERA2 defines additional blocks of information, that are not directly applicable to purely theoretical experiments. Optional Information include temporal and spatial coverage as well as the quality control procedure. The attributes listed above are collected in form of iRODS attributes and stored within the iRODS catalog. In this way, data becomes easily searchable.

3.2. Documentation

Human readable files with documentation (e.g., read-me files in markdown format[8]) are recommended on the levels Experiment, Dataset-Group, and Dataset. The collected information should be sufficient to repeat an experiment and to make it comprehensible how scientific results where obtained as outlined in DFG guidelines for good scientific practice (guideline 12, see related policies).

4. Ethics and Legal Compliance

4.1. Ethics

No person-related data are collected. Thus, no ethical issues are expected.

4.2. Legal Compliance

Not all data are owned by the project or project members. To ensure intellectual property rights, the following rules are applied depending on the type of data:

Type	Regulation
External data	Data provided by third-parties remain in possession of third-parties and may only be published with their permission. Whether this permission is given is stored as a mandatory attribute of each external dataset.
Internal data	These data are freely reusable within the consortium. For publication a license has to be added. Internal data may not be shared outside of W2W if it is derived from an external dataset where publication of derived results is not permitted by the data provider.
External code	External source code may only be used in agreement with existing license agreements. Sharing or publication of external source code is only allowed in cases where the license explicitly permits it.
Internal code	Source code produced by members of W2W is intended for publication along with papers using or describing the code. See also 7.3.

8 https://www.markdownguide.org

5. Storage and Backup

5.1. Provision

The requirements for disk space within the project significantly exceed the basic capacities (“Grundausstattung”) of the participating institutes. Funding for additional storage has been applied for based on approximation of requirements, which has been carried out individually for each sub-project

The storage will be provided in collaboration with regional computing centers and administration groups at the individual sites involved in the project. Tape archives of computing centers are used for regular and automatic backups. Z2 staff at each site is responsible to ensure that at least a weekly backup of all data is created and that access to this backup is possible.

5.2. Access Control and Security

No confidential personal data are collected. Despite this, the following risks exist in the handling of data: (a) Loss of data and scientific results due to hardware failure or accidentally deletion; (b) violation of license agreements by publication of external data or source code.

Hardware failures are accounted for by usage of redundant components and regular backups; this is ensured by the local computing centers and administration groups. Accidental deletion shall be avoided by appropriate access permission. We try to be as open as possible by allowing all project members to read all data, but we only grant write permissions to the people actually responsible for the data. Technically, this is implemented by means of unix or iRODS permissions set for archived datasets. Once a dataset reached its final state (e.g., after it has been used in a publication), it becomes a read-only resource in our iRODS system to prevent further changes.

Secure access for remote collaborators is provided by SSH-based methods as well as by exchange of datasets using iRODS. The later is achieved by connecting local iRODS instances to a so-called federation.

6. Selection and Preservation

6.1. Selection of Data

The goal for data management is to become able to reproduce experiments. However, we do not strive for bit-wise reproducibility but for a repeatability of experiments. The former is almost impossible to achieve as factors such as available computers and compilers, especially in the future, are out of our control. A model experiment becomes repeatable when the following conditions are met: (a) source code of the model is preserved along with the code of tools used for pre- and post-processing; (b) input data is preserved; and (c) all settings relevant for the model run (e.g., name-lists) are preserved.

6.2. Method for Preservation

The data will be retained for ten years after the end of the project in agreement with DFG regulations. This is ensured by utilization of the tape archives available at the local computing centers; the costs are covered by participating Universities.

Beyond that, data used in publications will be made available to the public wherever the licenses allow this and the data are required to repeat the experiments. In contrast to the climate community, publication of model results from weather experiments is rather unusual and so far no suitable archive for multi-terabyte datasets is available. W2W will accompany the development of such an archive at the partner institution ZDV, JGU. In addition to that, services provided by local libraries at KIT[9] and LMU[10] will be used.

7. Data Sharing

7.1. Potential Targets for Sharing

Data created and collected within the project is of interest for (a) readers of W2W publications, (b) members of W2W for collaboration, and (c) for future members of W2W in a possible third phase.

7.2. Search Methods

Readers of publications will find data by means of persistent identifiers (e.g., DOIs[11]). Members of the project can use the iRODS data management system to search for metadata. DOIs or other persistent identifiers are not used for internal data and code, unless it has been published. Inside of the project, local instances of the iRODS system are connected to each other and make it possible to replicate data at all sites and to search for data located at other sites.

7.3. Restrictions

Within the consortium, data are freely shared without any restrictions. For sharing with externals, the following restrictions apply:

Data Type	**Restrictions**
External Data	In general, external data are provided by third-parties and shared by them. We do not re-share their data, unless we have explicit permission to do so and the third-party is not sharing the data (e.g., because it has not been archived).
Internal Data	Data created within W2W are shared for scientific purposes to make experiments repeatable. This is done when license agreements for (input) data and code permit it. Unless otherwise required, we recommend data publications under one of the creative common licenses[12].
External Code	Like data, code is shared by the third-parties who own it. For closed-source code like ICON or COSMO, this usually involves signing a license agreement with the code owner. External code is not re-shared, unless it is open source and the license permits it.
Internal Code	Program libraries or other software projects created within W2W are made

9 KIT open: https://www.bibliothek.kit.edu/cms/kitopen.php
10 Open Data LMU: https://data.ub.uni-muenchen.de
11 https://www.doi.org
12 https://creativecommons.org/licenses

Data Type	Restrictions
	available to the public as GIT repository by using code sharing platforms like github.com and gitlab.com. To gain maximal visibility, repositories are organized in a group called “wavestoweather”. In addition, persistent identifiers (e.g., DOIs) are obtained from open access code publishing services like Zenodo[13]. Usage of an open-source-license that ensures reproduction of the copyright noticed is recommended (e.g., MIT[14]). Code that is not published for its own sake, but in support of a publication (e.g., scripts for plots or Jupyter Notebooks), can be published without a repository only on a platform like Zenodo or as supplementary material at the journal. Scientists are encouraged to do so, but without a strict requirement.

As outlined above, restrictions are required wherever license agreements for data produced by external partners are in place. During the process of collecting data from third-parties, we will already ask for permission to share derived results to minimize restrictions.

In general external data and code are not exclusively available to W2W. For internally create data and code exclusive access is only required before the publication of scientific results. For all publications regardless of the form, a clear attribution to the project should be possible.

8. Responsibilities and Resources

8.1. Assignment of Responsibilities

The data management plan is created by the Z2 project and reviewed by the steering group. Revised versions are created in the same way. Responsibility for the technical implementation lies with the Z2 members at the participating sites: Robert Redl (LMU), Jörg Steinkamp (JGU), and Christian Barthlott (KIT). They act as interface between the project and the local computing centers and administration groups. Responsibilities for individual activities:

Activity	Responsibility
Data Collection	
• External Data	In general, data is centrally collected by Z2 staff. Data collected by project members is included and checked by Z2 staff before sharing within the project.
• Internal Data	Data will be collected by project scientists. Where necessary, with support of Z2 staff. Project scientists are responsible for metadata and documentation.
• External Code	Will be centrally collected by Z2 staff (e.g., ICON and COSMO), but also by project scientists.
• Internal Code	Will be collected by project scientists. Where necessary, with support of Z2 staff.
Metadata	
• Production	The same responsibility as for the data itself.
• Control	Quality control (e.g., completeness) is done by Z2 staff. A technical solution, which automatically checks availability of mandatory information, should be implemented.
Data Quality	The same responsibility as for the data itself. For external data we rely on quality-control mechanisms of our third-party partners.
Storage and	R.R., J.S., and C.B. coordinate the provision of the actual disk space. They also

13 https://zenodo.org
14 https://opensource.org/licenses/MIT

Activity	Responsibility
Backup	ensure that backups are regularly created. They do that in collaboration with the computing centers.
Archiving and Sharing	Individual project scientists with technical support of R.R., J.S., and C.B.

It is the responsibility of project PIs to ensure that project scientists respect the policies outlined in this document. PIs can ask for technical support from Z2 staff.

8.2. Resources

In addition to the granted financial support for the required hardware the following support is required for the implementation of the data management plan:

- A central component of the data management is the data management software iRODS. Expertise on this area is available in Mainz (J.S.). J.S. will contribute to training of project scientists as well as other members of Z2.
- iRODS is already in use at ZDV and KIT. At LMU a solution involving the administration group of the faculty is planed to ensure persistence beyond the end of the project.