Research Data Archiving Checklist

Guidance for PhD supervisors

Authors: TU Delft Data Stewards Team
Contact: datastewards@tudelft.nl

Background

In accordance with the TU Delft Research Data Framework Policy, all doctoral candidates who started on, or after 1 January 2019, are required to upload all research data and code underlying completed PhD theses to 4TU.ResearchData (or another suitable data archive) before graduating, unless there are valid reasons not to do so.

This document should be used by PhD supervisors as a final checklist about the activities and plans regarding archiving of the research data and/or code from a PhD project before signing Form B. Please only sign Form B after going through this checklist with your PhD candidate and you are satisfied that the research data and/or code is archived in accordance with the guidelines given here and the best practices in your research domain, if applicable.

Please note that:

- Data and/or code underlying scientific papers should have been archived at the time of the publication of the corresponding papers. This guidance document can be used for such data and/or code that has not been archived yet or for any other remaining data and/or code.

- Issues around research data management and data archiving should have been already discussed and addressed throughout the PhD project and the corresponding decisions should have been recorded in a Data Management Plan in consultation with the faculty Data Steward.

- This checklist can also be used for archiving research code/software, whenever applicable. For more detailed guidance, please check the FAIR software checklist by the TU Delft Digital Competence Center.

Terms marked in **bold** are provided with explanations or examples at the end of this document.

Research Data Archiving Checklist

This is a checklist that focuses on the activities and plans regarding archiving of the **research data** and/or code/software from a PhD project.

Check what data and/or code has been already archived

> Tip:
>
> - Check the Data Management Plan (DMP) for an overview of the data (Q3, Q8, Q10), data sharing (Q26 – Q32), arrangements regarding confidential and personal data (Q9, Q14, Q15, Q18, Q22-Q25), licenses (Q32), responsible contact after the PhD (Q34).
> - Discuss the questions on this checklist with your PhD candidate.
> - Contact your faculty data steward for advice, if needed.

1. If applicable, do you have a list of research data and/or code that has been already archived? (Yes / No)

> Tip:
>
> - Data and code underlying scientific papers should have been archived at the time of the publication of the corresponding papers.
> - Check the definition and examples of research data in the Explanations section below.

If No to 1: Please make sure to discuss this with your PhD candidate.

If Yes to 1, check if:

1. The list of archived research data and/or code is complete.
2. You can access all the data and/or code on the list.

If Yes to 1.2:

1.2.1. Check if the data and/or code properly are properly documented and complete.

Check what data and/or code can or should be published

2. Is there any research data and/or code that has not been archived yet? (Yes / No)

> Tip:
>
> - This could be data and/or code underlying scientific papers that has not been archived yet or any other remaining data or code underlying the PhD thesis.

If No to 2:

You can stop here.

If Yes to 2:

2.1. Is the remaining research data and/or code suitable for archiving? (Yes / No)

> Tip:
>
> - Certain types of research data might be unsuitable for archiving. Check examples and advice about what to do in such cases.
> - If the project includes personal data or confidential data, issues around management and archiving of such data should have been already discussed and addressed throughout the PhD project and the corresponding decisions should have been recorded in a Data Management Plan. If this is not case, please contact your faculty Data Steward for advice.

If Yes to 2.1:

Please ensure the data and/or code is archived appropriately by using the rest of this checklist.

If No to 2.1:

You can stop here.

Choose a proper data repository

Tip:

- It is advised to choose a **trusted data repository** with clear terms and conditions such as 4TU.ResearchData, zenodo, DANS, figshare.
- When choosing a repository, it is also important to take archiving costs into account. For instance, TU Delft researchers (including PhD candidates) can deposit up to 1TB of data and/or code free of charge per year to 4TU.ResearchData.
- Check detailed guidance about archiving data and software.
- If you are unsure about the choice of repository, you can contact your faculty data steward for advice.

Essential criteria

3. Is it one of the following repositories: 4TU.ResearchData, zenodo, DANS, figshare? (Yes / No)

If Yes to 3: Continue with 4.

If No to 3, check if:

3.1. It is recognized in the research community that you work with.
3.2. It provides persistent and unique identifiers like DOI, handle etc.
3.3. It offers standard licenses for data and/or code.
3.4. It uses common metadata standards for the data.

Assess the technical quality of the data

4. Check if the data files are well structured

> **Tip:**
>
> - Check here for tips for file naming.
> - Check the paper Good enough practices in scientific computing.

    4.1. Check if the files are properly named.
    4.2. Check if the files are organized with a proper structure.

5. Check if all the files can be opened (and are not corrupted).
6. Check if the data is stored in open or standard format for long term archiving.

> **Tip:**
>
> - Check the preferred file format at the 4TU.ResearchData or Library of Congress Recommended Formats for detailed info.

7. Check if there is a **README file** that can be opened in an open format, e.g. .txt or .md.
8. Check if the README file includes the elements listed below:

> **Tip:**
>
> - The elements listed below are mainly applicable to data while some of them are also applicable to code. You can find some README templates and guidance here:
>   - Software / Code README (TU Delft DCC)
>   - Data README (4TU.ResearchData)
>   - Data / Code / ML models READMEs (TU Delft AE)

    8.1. Title
    8.2. Author(s) names, affiliations, contact info and ORCID
    8.3. List of all the files and accompanying documentation
    8.4. Methodology of data collection and analysis, including (if applicable) any setting or protocol, assumptions and software (including version number) used to analyze the data
    8.5. Description of the data and/or code, the file naming and structure of data
    8.6. References to related articles, with DOI
    8.7. Licenses
    8.8. Recommended way of citing
    8.9. Any other information that helps readers to better understand the data and/or code

9. If applicable, check if any additional documentation is going to be published along with the data and/or code (e.g. a data dictionary, Electronic Lab Notebook, Readthedocs, GitHub Wiki, etc.).
10. For any type of **personal data** or **confidential data** that is going to be published with restricted access or as a metadata-only record, check if the following is included in the metadata or other accompanying documentation:

    10.1. The reason for data being confidential
    10.2. The data ownership
    10.3. The contact info for requesting access to the datasets
    10.4. The protocol for accessing the data

Explanations

Research Data

Research data is the evidence that underpins answers to research questions, and which is necessary to validate research findings. Data can come in various forms and types, characteristic to specific disciplines of research. For example, data can be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, modelling, interview or other methods, or information derived from existing evidence. Examples of research data include:

- Documents (text, Word), spreadsheets
- Laboratory notebooks, field notebooks, diaries
- Questionnaires, transcripts, codebooks
- Audiotapes, videotapes
- Photographs, films
- Protein or genetic sequences
- Test responses
- Slides, artifacts, specimens, samples
- Collection of digital objects acquired and generated during the process of research
- Database contents (video, audio, text, images)
- Models, algorithms, scripts
- Contents of an application (input, output, logfiles for analysis software, simulation software, schemas)
- Methodologies and workflows
- Standard operating procedures and protocols

Trusted data repository

4TU.ResearchData, zenodo and figshare are all trusted repositories that archive data according to the FAIR (Findable, Accessible, Interoperable, Reusable) principles by:

- making research data accessible, discoverable and available for the long term,
- providing persistent and unique identifiers like DOI to make data findable and citable
- offering standard licenses that determines terms and conditions regarding sharing and reuse
- using common metadata standards to help others identify and discover the data.

Yet, if there are other disciplinary or domain repositories that are commonly used and endorsed by your research community, those might be more suitable to archive the data resulting from the project. Re3data.org offers an overview of data repositories.

README file

A README file provides information about a dataset and is intended to help ensure that the data can be correctly interpreted, by yourself at a later date or by others when sharing or publishing data. A README file must be submitted along with the dataset file(s).

Personal data

Personal data - all information about an identified or identifiable natural person (the data subject). A person is considered identifiable if he or she can be identified directly or indirectly based on one or more items of personal data, for example, name and address, ethnicity, date of birth and IP-address. In general, it can be assumed that personal data include all data relating to a living person that makes it possible to identify this person or to distinguish him or her uniquely from other persons.

Confidential data

A few examples of confidential data:

- national security data (e.g. nuclear research)
- data falling under export control regulations
- confidential data received from commercial, or other external partners
- data related to competitive advantage (e.g. patent, IP)
- data which could lead to reputation/brand damage (e.g. animal research, climate change, personal data)
- politically-sensitive data (e.g. research commissioned by public authorities, research in social issues)