# ARCHIVING

## FREQUENTLY ASKED QUESTIONS

Luke Günther & Job Schepens
Project S
sfb1252-service@uni-koeln.de

## Contents



## What is "long-term archiving" and why is it necessary?

*Long-term archiving* refers to the process of **preserving digital content** (including both data and metadata) for an extended period of time, while ensuring its continued *accessibility*, *readability*, and *interpretability*.

To comply with the **DFG's funding requirements** for Collaborative Research Centres, we need to guarantee that all research data remains accessible "beyond the ultimate funding period".[1] Specifically, the DFG's Code of Conduct *Guidelines for Safeguarding Good Research Practice*[2] states that all *primary data*, i.e. raw research data used for publication plus all relevant documentation, should be stored **for at least 10 years**.[3] This is to ensure that all research outcomes can be traced back to individual contributors in the foreseeable future.

Project S is responsible for implementing the DFG's requirements within our CRC.

[1] Information Management in Collaborative Research Centres | dfg.de

[2] Deutsche Forschungsgemeinschaft (2025). Guidelines for Safeguarding Good Research Practice. Code of Conduct. https://doi.org/10.5281/zenodo.14281892

[3] Note that this 10 year timeframe starts with the *date of publication* of any studies relying on the respective primary (= raw) data. Even if the data itself is older, the 10 year period "resets" when you use it for any new publications.

# What does the CRC's archiving process look like?

Over the previous two phases, we have settled on an archiving process that is meant to be as straightforward and manageable as possible. There is **one archiving round per year**, announced in fall and to be completed by the end of the year.

The entire process can be broken down as follows:

### Data collection

Each project discusses what data they need to archive and then organizes the relevant files and folders into a specific *archive folder* on Sciebo. At least one person per project, the designated *metadata contact person*, is responsible for providing detailed information about the origin and purpose of the data in a separate *metadata spreadsheet*. (We recommend splitting this task between members.)

See below for the step-by-step tutorial.

### Review

Once this process is completed, **your metadata contact person informs us** that your data is ready to be archived. We then carefully review the information you provided and **extract the metadata** from the spreadsheet into a structured and standardized format.

### Storage

We submit this initial information package to the *Data Center for the Humanities* (DCH). The DCH does a **final data input check**, generates *checksums* to protect against unauthorized file changes, and assigns a permanent *Digital Object Identifier* (DOI). A short description of each archive is published on the DCH's website. The resulting *archival information package* is transferred to the *ITCC* where it is **stored on magnetic tape**.

The DCH has provided a separate info sheet detailing their involvement.

# What do I/we need to do?

### Step 1: Select data

Within your project, first **decide what data needs to be archived**, including:

- everything you have worked on within the past year, whether published or unpublished
- any data that has never been archived before

Please include everything, whether it has been used as part of a publication or not. Even if the data was previously uploaded to an external repository. Even if the data is still a work in progress.

We've provided some recommendations about the files you should consider.

If you are unsure, **ask us**.

### Step 2: Copy & organize data

Next, you should **copy all relevant files and folders** into your project's designated *archive folder* on Sciebo, located at `XX_general/XX_archive` (where `XX` stands for your project ID, e.g. `A01_general/A01_archive`).

Organize your files into clearly labeled subfolders, e.g. by year or by study ID/title.

### Step 3: Add documentation

Once your data has been placed in the archive, you need to document it. All metadata is located in a subfolder of your archive called `XX_metadata`. We provide two *metadata spreadsheets* (Excel files), one for experiments and another for corpus-based studies:

- Use the spreadsheet `metadata_experimental` if you have conducted any experiments with participants, for example in a lab-based setting. This includes: elicitation studies, eye-tracking, pupillometry, EEG/ERP, questionnaires, mouse-tracking etc.
- Use the spreadsheet `metadata_corpus` if you have compiled your own corpus/dataset, have done any fieldwork trips, or based your analyses on existing corpora.

If both types seem applicable to your project, please fill in both. We've shared the template files in the `S_FAQ` folder.

### Notes for continued projects

- Please review the current contents of your archive folder.
  - If you have continued working on a dataset or study, you can simply replace the corresponding folders in the existing archive. Check the existing metadata for accuracy and update it as necessary.
  - For any new data, place it in the archive folder and create corresponding entries in the metadata Excel sheets.
- You may start with a clean slate if you would rather do so, e.g. if your existing archive already contains a lot of (old) data but you aren't familiar with the folder structure. In that case, either:
  - delete all existing content and recreate the metadata subfolder (it's safe, don't worry) **OR**
  - move everything somewhere else – but please keep it in your project's team folder `XX_general` and give it a sensible name (e.g. `2024-12_XX-archive` and not `old archive` or `delete me?` 😉 )

### Notes for new projects

- If it doesn't exist, create a subfolder `XX_metadata` within your archive.
- Discuss which metadata spreadsheet fits your project best, fill it in, and move it to this subfolder.

**Ask us if *anything* is unclear.**

## How should the metadata sheets be filled in?

Both metadata spreadsheets bundle a number of tabs/tables (generally listed at the bottom of the screen).

- Each table contains several fields of information that are either *mandatory* (marked as `M`), *mandatory if applicable* (`MC`) or *optional* (`*`, more asterisks indicate a higher priority).
- Each field comes with a *title* (bold and blue), a *description* (grey) and information about the expected *format* of the value you provide (yellow).
- Some values need to be chosen from a dropdown list.

**Shared across both sheet types:**

- `Project Info` and `Contributors` contain general information about your project and a list of everyone who has contributed to the archived data with their contact information.
- In `Path information`, you describe which folders in your archive correspond with which experiment or study that you've described in the other tables. This is simply a path-ID mapping, e.g. `surveys/some-subfolder/2025-02_survey-about-something` belongs to the ID `survey_01`. You will be asked to assign these IDs in the tables relevant to your project type (experimental/corpus, see below).

**Only applicable to `experimental` sheet:**

- `Experiment (general)`: general information about the experiments you've conducted
- `Experiment (specific)`: detailed description of hypotheses, methodological approach, sampling, the materials you've used etc

In the general information, you will be asked to assign a unique ID to each experiment which you will use to reference specific experiments in the other tables.

**Only applicable to `corpus` sheet**:

- `Corpus`: general information about the corpora you've used or created
- `Study`: description of the analyses you've conducted based on that data
- `Fieldwork`: details about fieldwork trips (location, timeframe, languages, purpose, who went)
- `Sessions`: list of recorded sessions and interviews

In `Corpus` and `Study`, you will be asked to assign unique IDs to each corpus/study. Use these in `Fieldwork` and `Sessions` to assign trips and recording sessions to specific corpora/studies.

Do **not** manually modify the remaining tables called `filelist`, `variables` and `autoComplete`.

Lastly: The metadata sheets look more intimidating than they are, we promise. If you need help, just let us know as soon as possible.

## What is the role of the contact person?

We ask each project to designate a contact person. They are responsible for:

- coordinating the archiving process within their project,
- making sure that all data is accounted for,
- checking that the metadata spreadsheet is filled in correctly and
- communicating the result back to us.

If we have questions regarding the current status of your project's archive, we will first message that contact person.

## But I've already created my own metadata! Should I still fill in the metadata sheets?

Yes! 😄 We need the same standardized information from all projects. However, we love to hear that you've created your own documentation, especially if it adds details we didn't ask for or covers information that's specific to your research! Please **add it your project archive** as well.

## What data should be archived?

All data that was produced as part of the project. This includes:

- raw and processed datasets (e.g. text files, audiovisual recordings, annotations)
- scripts for processing/analysing said datasets with as much documentation as possible (software requirements, package dependencies/environments, usage examples, container configurations, version control etc.)
- study materials (e.g. questionnaries, specific tasks)
- any files documenting your planning, methodological approach and results

Data collected by **PhD students** for their dissertation (that is not part of the other project data) also needs to be archived. Simply create a subfolder like `lastname_phd`, copy your data and provide relevant information in the metadata sheet.

## Some of the data is shared between projects. What should we do?

Based on our experience, most datasets are small enough (in terms of storage space) that you could simply include them in both/all archives. If you have concerns about the file/folder size (or uploading it again would be cumbersome), only store the data in one project's archive and contact us with the details.

## Our data contains personal or otherwise sensitive information. What should we do?

If your data contains personal/personally identifiable information about third parties (such as study participants), please carefully consider whether you are allowed to process and upload it to Sciebo, taking into account:

- applicable consent forms
- Sciebo's Terms of Use
- EU data protections laws (GDPR)

Some personal data of third parties ***must not*** be stored on Sciebo, even if consent was given. According to their Terms of Use at the time of writing[4], this includes (following Art. 9 GDPR):

- data revealing the racial and ethnic origin
- data indicating the political opinion
- data revealing the religious or philosophical convictions
- data indicating the trade union membership
- genetic data
- biometric data for the unique identification of a natural person
- health data
- data on the sex life or sexual orientation of a natural person

**What you should do:**

- Please ***always*** inform us about the existence and nature of potentially or known sensitive data in your project.

[4] Terms of Use | Hochschulcloud NRW

- Err on the side of caution and do **not** upload the data to Sciebo. Instead, request an external hard drive from us and deliver the data to us in person. You can still document the files as if they were part of your archive folder.

## Who will have access to the data?

Nobody will have direct access to the data. The data is stored as a "dark" archive in the ITCC's tape library, so it's not published anywhere and is not meant for regular access. You or third parties may request access to the archive by contacting us or the DCH. The project leaders will then be contacted. (This hasn't happened so far.)