Managing digital research data

Basics, tips & tricks

Team Research Data Management

last updated: August 2024

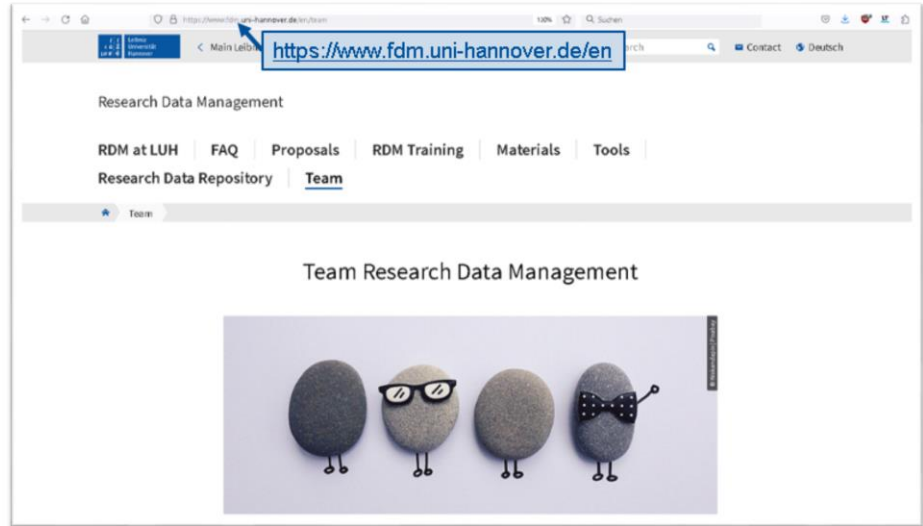*DOI: 10.5281/zenodo.13305129*

Hello and welcome to our online course "Managing digital research data - Basics, tips & tricks"!

Welcome!

Chapters

- Introduction
- Planning your data management
- Specifications, guidelines and laws
- Collecting, storing and processing data
- Documenting your data processing
- Protecting your data
- Archiving and publishing your data
- Supporting services and initiatives

https://www.fdm.uni-hannover.de/en

Research Data Management

RDM at LUH | FAQ | Proposals | RDM Training | Materials | Tools | Research Data Repository | Team

Team

Team Research Data Management

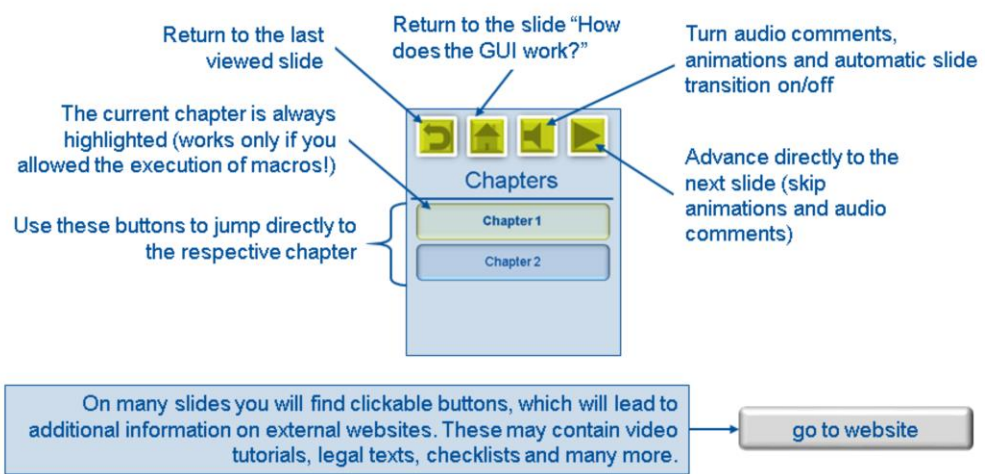visit the website of the Research Data Support Team

This course was created by the joint Team Research Data Management of LUH and TIB. Please visit our website to find out more about our training, counselling and support services. You will find extensive information on research data management in general, but we are also happy to provide LUH members with individual advice.

How does the GUI work?

Chapters

- Introduction
- Planning your data management
- Specifications, guidelines and laws
- Collecting, storing and processing data
- Documenting your data processing
- Protecting your data
- Archiving and publishing your data
- Supporting services and initiatives

Return to the last viewed slide

Return to the slide "How does the GUI work?"

Turn audio comments, animations and automatic slide transition on/off

The current chapter is always highlighted (works only if you allowed the execution of macros!)

Advance directly to the next slide (skip animations and audio comments)

Use these buttons to jump directly to the respective chapter

Chapters

Chapter 1

Chapter 2

On many slides you will find clickable buttons, which will lead to additional information on external websites. These may contain video tutorials, legal texts, checklists and many more.

go to website

Before we get started with the topic itself, here are a few words on how the graphical user interface works. We designed this presentation for autonomous learning. That means we want you to decide for yourself, which chapters you want to work through and in which order. For this reason, there is a list of main chapters on the left. A click on a respective buttons will bring you directly to the corresponding chapter overview. From there, you can also jump to a specific slide. To ease orientation the active main chapter is highlighted in green on the sidebar.

If you want to go back to the previously shown slide, click the green button showing a bent arrow at the upper left. The button next to it brings you back to this explanation slide at the beginning of the presentation. The third green button switches the audio comments on the slides and the related animations on or off. So, if our chatter starts getting on your nerves or if you just want to have a closer look at a slide before moving on, please don´t hesitate to click on it. The rightmost button brings you directly to the next slide, in case you want to skip animations and audio comments of the current slide.

The final remark concerns the grey buttons that you can find on many slides. Clicking on these will generally open a browser window showing a website or an online document with additional information.

Contents and aspiration of this course

Chapters

- Introduction
- Planning your data management
- Specifications, guidelines and laws
- Collecting, storing and processing data
- Documenting your data processing
- Protecting your data
- Archiving and publishing your data
- Supporting services and initiatives

i

Aspiration of this course: Giving an introducing overview over the most important aspects of the handling of digital research data. Addresses persons with little prior knowledge.

go to the advanced courses

Image by Menno de Jong: https://pixabay.com/de/photos/lemur-neugierig-halfaap-guck-guck-3295891

In this course, we would like to give you an introductory overview of the most important aspects of dealing with research data. It aims at people with little previous knowledge and goes into more breadth than depth. However, we will successively offer more in-depth courses on individual topics. You will find these on our website under "Training courses and events".

After a short introduction, we will first explain how you can plan data management at an early stage and what you need to consider. We then present important specifications, guidelines and laws as general framework conditions. The following chapter is about the collection, storage and processing of data, succeeded by a section dedicated to the appropriate documentation of these activities. The topic of protection against data loss and misuse also deserves an extra chapter. Finally, we explain how to archive and publish data. On the last slides, we will present further services and support for data management at your disposal.

Introduction

What are research data?

What is research data management (RDM)?
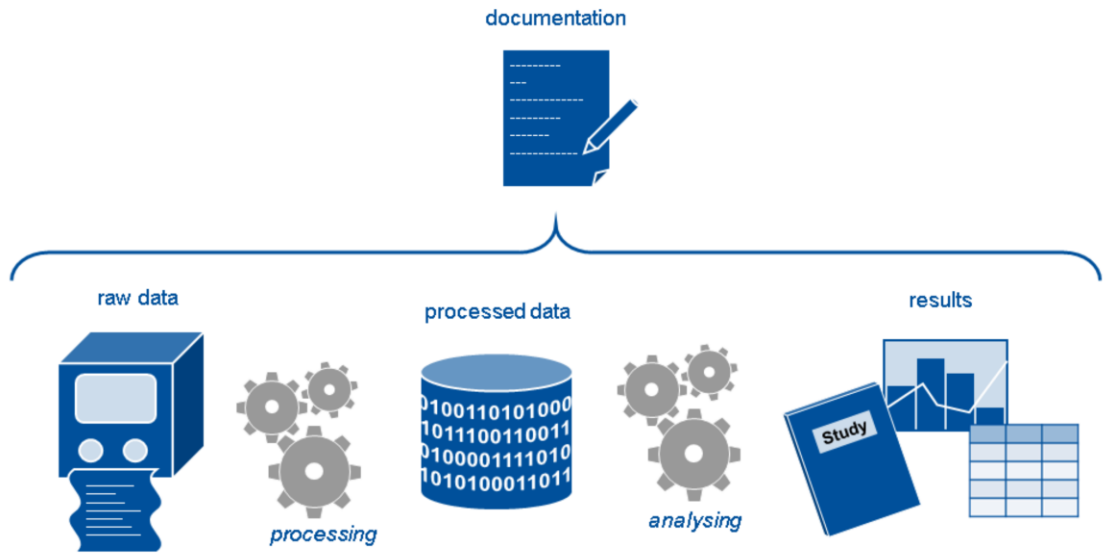
Why is data management important?

So, let's begin! In this introductory chapter, we will explain what the terms "research data" and "research data management" mean and why data management is so important.

What are research data?

What does the term "research data" actually cover? There are several definitions, but none that can claim general validity. When we speak of "research data" in this course, this is what we have in mind:

On the one hand, there is the so-called raw or primary data, which is data that is still complete unprocessed. For example, this can be measurement data, survey data or archive material.

This raw data is usually processed in several steps. For example, data is aggregated, filtered or converted. Others, such as interview recordings, are transcribed, pseudonymised and annotated.

In the next step, during analysis, data is often aggregated. For example, general statistics can be derived from many measurement series, interviews or questionnaires. The results of these analyses and their interpretation are usually published as research literature.

Raw data, processed data and analysis results are all research data. In addition, there is another category of particular importance: the documentation of everything you do with your data during the research process. Documentation includes descriptions of methods and workflows as well as logging processing steps and systematically collecting structured metadata about the files. We will deal with this in more detail in the chapter "Documenting your data processing".

What is research data management (RDM)?

planning

processing

administrating

structuring

documenting

selecting

archiving

publishing

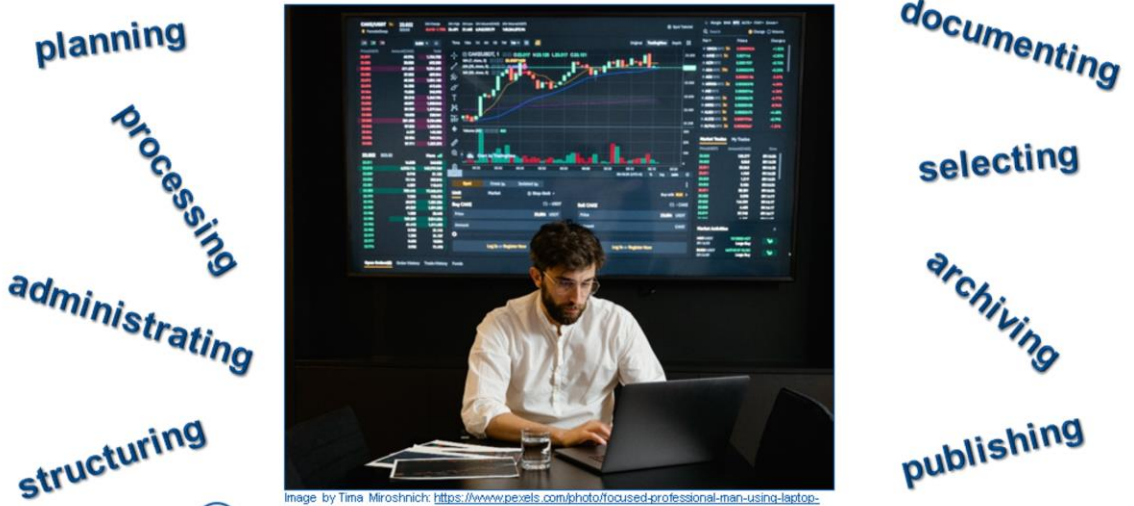Image by Tima Miroshnich: https://www.pexels.com/photo/focused-professional-man-using-laptop-7567529/

RDM: Conscious and systematic handling of data from the planning stage until the end of a project and perhaps even beyond.

Since different types of data are usually processed throughout the research process, the management of this research data is also something that accompanies all stages of scientific work. Strictly speaking, it begins before a project even starts, namely with the planning of how to handle data. And it doesn't end when a project comes to an end either because the resulting data is usually archived over longer periods of time or even published and must remain readable and usable during this period.

Research data management is therefore the conscious and systematic handling of scientific data. Since the term is somewhat unwieldy, it is usually abbreviated to "RDM".

Why is data management important?

- ✓ you keep an overview

- ✓ team work is easier

- ✓ you can safeguard high quality standards in research

- ✓ you save time and avoid stress

- ✓ you comply with official requirements

Chapters

The opinion is still widespread that a planned and systematic data management consumes valuable time and resources that should be spend preferably for the actual research work. However, this assessment turns out to be a naive assumption time and again. All too often, a doubled or tripled amount of the supposedly saved-up resources have to be spent afterwards in order to iron out the negative consequences of a negligent data management, at least partially. Please take a look at our reading tip.

- If you manage your data well right from the start, there are many advantages:

- You maintain an overview because you can reliably find data and information again.

- You work together much more effectively as a team because you have agreed on common standards and procedures.

- The quality of your data and the research results derived from it is assured because you have established meaningful review mechanisms.

- You save an enormous amount of time and energy, especially at the end of a project because your data is already prepared and documented. You have everything at your fingertips and can even reconstruct what you did at the very beginning.

- And finally, you comply with various formal requirements, such as those defined in laws, guidelines or funding conditions.

Planning your data management

What should I consider?

Adequate infrastructure

People and departments involved

Developing a data management plan (DMP)

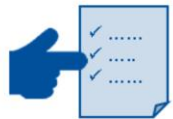Tools for developing a data management plan

The key to a good data management is timely and well thought-out planning. In this chapter, we explain which points you should take into account and which tools may be helpful.

What should I consider?

specifications, guidelines, laws

1) planning

2) collecting/ creating

3) processing

4) analysing

5) archiving/ publishing

01001101010
00101110011
00110100001
11101010101

people involved in the project: workflows & responsibilities
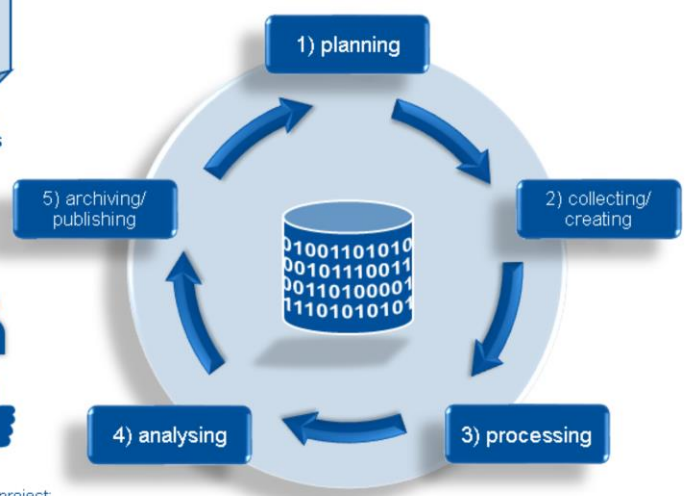
documentation

structure & overview

data storage & transmission

protection against loss and misuse

selection

Plan your data handling well ahead. Be aware that there are many aspects to consider, starting with the framework conditions like formal provisions. We will address this point in more detail in the following chapter. Also, be aware of all the people contributing to your project - either directly or indirectly – and the adjustments of workflows and responsibilities between them.

The circle chart in the centre symbolises the typical cycle that data in research projects go through. After the planning stage, the first step is generally the creation or collection of data. Following procession and analyses, you usually archive or publish your data. You may also delete some, for example for compliance with data protection regulations. The preserved data is available for re-use in new projects. Hence, the circle is closed.

During the distinct phases, you will have to find adequate solutions for different kinds of data handling challenges. What should I document where and how? How do I keep the order and overview? Do I dispose of sufficient suitable infrastructure for data storage and transmission? How can I adequately protect my data from loss and misuse? And how do I decide which data to delete and which to keep?

Those who do not ask themselves such questions before the first data are already on the table will run into manifold problems. We therefore highly recommend that you take the time to plan your project in detail well ahead. In the end, you will save a much greater amount of time just because you do not have to fight chaos in the entire course of your project!

Adequate infrastructure
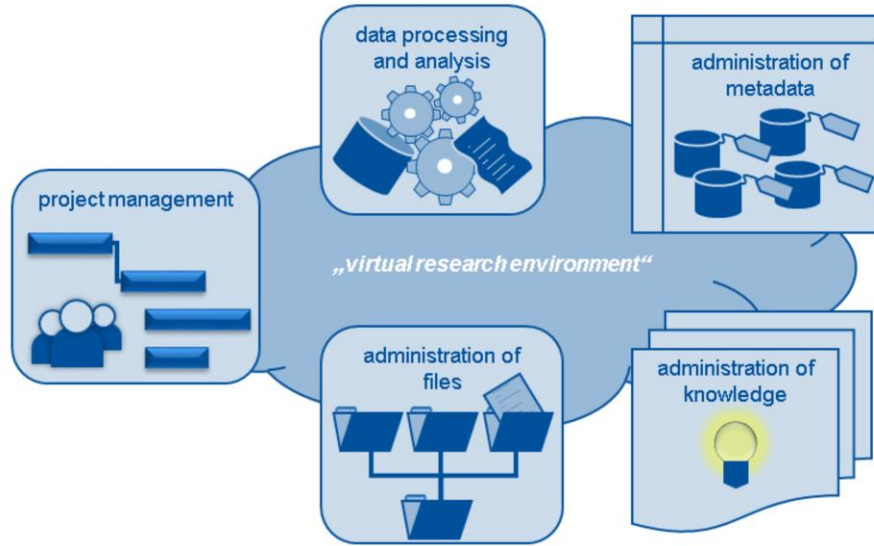
For many data management tasks, you will need suitable infrastructure, that is, hardware, software and perhaps online services. Ideally, the infrastructure components are linked and configured to work as an integrated system. In larger projects with many persons involved, these systems should optimally support teamwork, both, simultaneous and asynchronous. You may also hear the term "virtual research environment" for such infrastructure. Main components may be:

- A project management software used to plan time and resources and to define tasks and works packages

- Infrastructure for data processing and analysis that may comprise everything from licenses for special software to computing time on mainframe computers

- Data storage that meets the research requirements, especially regarding capacity, velocity and security. A file administration software running on top may complement such hardware.

- Some system to store and administer context information on your files, the so-called metadata.

- And finally, a system to record general knowledge, such as workflow descriptions, methods and the designs of experiments, but also protocols of team meetings and the like.

In this course, we will present a number of services available to you at our university. For special requirements, however, you may need to resort to external commercial or disciplinary service providers.
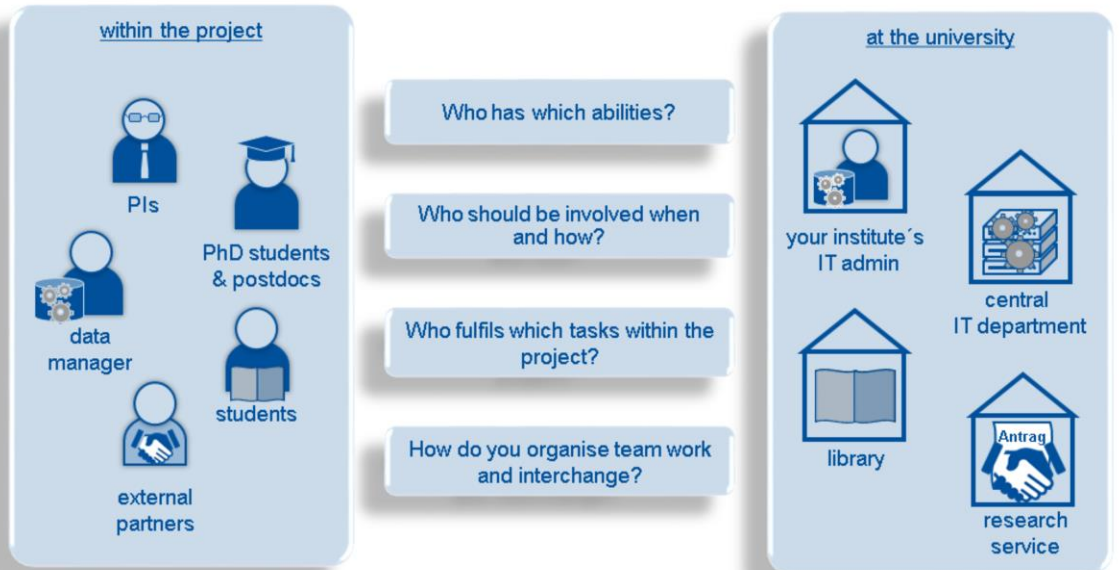
Apart from infrastructure, you should also put some thought in the people contributing to your project, either directly or indirectly. Among others, a project may involve the following players:

- "Principle investigators" who designed the project, applied for funding and head the research operations

- PhD students and postdocs working on individual topics or in certain research areas

- A data manager overseeing the technical infrastructure and supporting data management in general

- Student assistants working in individual subprojects or graduating in the scope of the project, and

- External partners like researchers from other institutions, service providers or commercial enterprises

Consider as well, which department at your university can support you and in which way. Your institute may employ its own IT administrator, who can help you setting up the infrastructure. The IT department provides a number of services centrally. The library offers counsel and support, in particular regarding questions on data publication, open access and metadata standards. Finally, the research service helps optimising content and wording of RDM chapters in funding proposals.

Ask yourself: Who has which competences and abilities regarding RDM? Whom should I involve at which point in time? Who assumes which tasks within the project? And how do I organise communication and collaboration between all persons involved as effectively as possible?

Developing a data management plan (DMP)

i DMP: A structured document containing detailed information on data handling.
→ May and should be constantly updated and complemented during the project

typical structure of a DMP:

- administrative information
- data collection and basic methodology
- storage, backup and security
- archiving
- sharing and publishing data
- resources und responsibilities

further reading

The Digital Curation Centre´s website on data management plans provides tools, checklists and guidance.

go to the DCC website

We recommend that you record your ideas in writing in a data management plan, abbreviated as DMP. A DMP is simply a structured document containing detailed information on data handling. You can and should regularly update and complement your plan in the course of a project. It is hence completely normal if your first draft is still incomplete, or if some of your ideas turn out to be unfeasible later on. The structure of a DMP is not standardised but conforms to the requirements of the respective project. A typical structure may comprise the following chapters:

- Administrative information
- Data collection and methodological basics
- Storage, backup and security
- Archiving
- Sharing and publication of data
- Resources and responsibilities

Please find extensive additional info on the subject at the website of the British Digital Curation Centre.
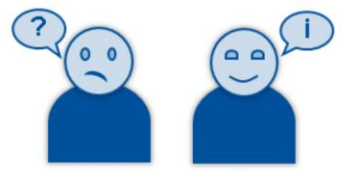
Tools for developing a data management plan

DMP

templates, checklists, examples (can be found in disciplinary web portals and on the websites of some funding bodies)

individual counselling by RDM support staff

go to the website of the Research Data Support Team at LUH

argos (en)

DMPonline (en)

RDMO (de/en)

commented online editors

There are many more templates, checklists and examples of data management plans. Some are disciplinary or tailored to the requirements of a certain funding organisation. Just have a look at the relevant websites of your research community or your preferred funding body.

Using an online tool can be of great help, in particular when working jointly on a DMP with several colleagues. Such tools walk you step by step through the chapters of a pre-set questionnaire. They explain for each chapter, which questions to answer and what to consider. In most cases, you enter your answers in a text box, but sometimes you can also choose between predefined options. When finished, you can export all of your inputs compiled into a single document.

One of the oldest and best known tools is DMPonline, provided by the British Digital Curation Centre. For the last couple of years more and more universities also deploy the Research Data Management Organizer, abbreviated as RDMO and developed in Germany. LUH is currently testing RDMO, but does still not offer a productive version. However, you can use the RDMO instance of the information portal forschungsdaten.info linked on this slide. This is a trustworthy instance complying with German data protection standards. At a European level, the development of Argos is showing progress. Argos has the potential to become a standard tool in the future, especially for EU-funded projects. Some disciplinary associations provide DMP tools as well.

If you require further counsel on DMP tools or templates, if you want feedback on your DMP draft or if you have any other question, don´t hesitate to contact our Research Data Support Team.

Specifications, guidelines and laws

Laws frequently affecting research practice

RDM is good scientific practice!

Guidelines for handling research data at LUH

Funder requirements

Why funders require RDM

Recommendations for RDM statements in funding proposals

Depending on discipline and research methods, different kinds of guidelines, policies and laws may be relevant for your project. Some are strictly binding while others are of recommending nature. In this chapter, we present some of the most important regulations with a direct impact on data management.

Laws frequently affecting research practice

Chapters

- Introduction
- Planning your data management
- Specifications, guidelines and laws
- Collecting, storing and processing data
- Documenting your data processing
- Protecting your data
- Archiving and publishing your data
- Supporting services and initiatives

Image by succo: https://pixabay.com/de/photos/hammer-waage-gericht-justiz-recht-802301

i We have compiled the most important information on legal issues on our website on the subpage "Legal & ethical aspects in handling research data".

to subpage „Legal & ethics"

§ Privacy and data protection laws

to the official version of the GDPR (Official Journal of the EU)

to NDSG (niedersächsisches Vorschrifteninformationssystem, in German)

for detailed information (in German) see:
Advanced course "Umgang mit personenbezogenen Forschungsdaten. Rechtliche Grundlagen, Methoden und Hilfsmittel."

go to the advanced course

§ Intellectual property and patent laws

to the German copyright law (Bundesamt für Justiz)

to the German patent law (Bundesamt für Justiz)

further reading (in German):
website of the Federal Ministry of Education and Research (BMBF): Urheberrecht in der Wissenschaft: Was Forschende und Lehrende wissen sollten. At LUH, Dezernat 4 offers counsel on patent applications.

to the BMBF website   to the Dez.4 website "Patents and Startups"

Obviously, laws are strictly binding. If a project processes personal data, all persons involved should know and respect the relevant provisions of the European General Data Protection Regulation and of the Data Protection Law of Lower Saxony. We provide an advanced course on this subject, which you can find on our website.

Laws on intellectual property often come into play when research is industry-oriented. The Federal Ministry of Education and Research compiled frequently asked questions related to intellectual property rights in teaching and research along with the respective answers. If you have questions on patent applications, you can get counsel at Dezernat 4.

You are also welcome to take a look at our website on legal and ethical aspects, where we compiled further information and points of contact.

RDM is good scientific practice!

RDM is especially relevant for these 6 out of 19 guidelines:

Guideline 7: Cross-phase quality assurance

Guideline 10: Legal and ethical frameworks, usage rights

Guideline 11: Methods and standards

Guideline 12: Documentation

Guideline 13: Providing public access to research results

Guideline 17: Archiving

Guidelines for Safeguarding Good Research Practice
Code of Conduct
DFG

i You can read the guidelines themselves and accompanying disciplinary comments in the DFG portal "Research Integrity" (English version not finished, yet).

to the DFG portal

! Please note as well:
- DFG guidelines on the handling of research data
- disciplinary guidelines, requirements, policy papers etc.

You can find both on this website:

to the DFG website „Handling of Research Data"

Chapters
- Introduction
- Planning your data management
- Specifications, guidelines and laws
- Collecting, storing and processing data
- Documenting your data processing
- Protecting your data
- Archiving and publishing your data
- Supporting services and initiatives

Apart from laws, there is the so-called good research practice. If you grossly disregard these standards, your reputation as a scholar may suffer greatly. What's more, you can be excluded from funding and suffer additional penalties. The code of conduct "Guidelines for Safeguarding Good Research Practice", edited by the German Research Foundation, is effective since August 2019. It comprises 19 guidelines, six of which explicitly refer to an adequate handling of research data. These are:

- Guideline 7: Cross-phase quality assurance

- Guideline 10: Legal and ethical frameworks, usage rights

- Guideline 11: Methods and standards

- Guideline 12: Documentation

- Guideline 13: Providing public access to research results and

- Guideline 17: Archiving

Please familiarise yourself with this code of conduct and comply with its regulations and recommendations in your research practice. In addition, you should also know the "DFG Guidelines on the Handling of Research Data" from 2015. On the same website, you can find a – somewhat chaotic – list of disciplinary guidelines and other RDM-related documents. If there is a specific policy for your discipline that surpasses the general DFG requirements, reviewers are encouraged to measure your proposal by these disciplinary standards. It is therefore a good idea to review the list for possibly relevant documents before submitting a proposal.

Guidelines for handling research data at LUH

principles:

- how to handle research data:
  - protect against losses
  - process for sustainable (re-)use
  - document
  - archive (long-term)
  - → head of project is responsible!

- publish research data according to the FAIR principles in (disciplinary) repositories (request)

- develop project-internal RDM policies and data management plans (recommendation)

- integrate RDM into teaching → faculties (recommendation)

§ You can find the official LUH website with the guidelines here:

got to guidelines

Many research institutions have their own policies or guidelines by now. At LUH, the recently updated "Guidelines for handling research data at Leibniz University Hannover" are effective since 2017. They comprise four principles of different commitment levels. Principle 1 says that the respective project directors are responsible to ensure that research data are

- protected against loss

- processed for sustainable use

- documented and

- preserved long-term

Principle 2 requests that you publish your research data in compliance with the FAIR principles in repositories, preferably in disciplinary repositories. We will explain the FAIR principles and describe best practice in publishing data later on. The third principle recommends developing project-internal RDM policies and data management plans. Finally, the fourth principle consists of a recommendation to the faculties to intensify integration of the RDM topic in academic teaching. Please familiarise yourself with these guidelines.

Funder requirements

the classic: „Oh oh, we still have to write something about research data..."

Application guide

------ --- ------ ----- - --
------ -- --- -- ------ -- --------

Data handling

------ --- ------ ----------- -- ---
-- ------ -- -------- -----------

submission deadline: day after tomorrow!

i Dezernat 4 is happy to support when it comes to drafting RDM-related proposal chapters. The sooner you come up to us, the more efficient our counselling can be. Please do not expect text blocks for copy & paste!

to the Dezernat 4 website (competences overview)

Funders generally demand that proposals contain a convincing RDM concept, and they specify on the expected information. Unfortunately, word still hasn´t spread everywhere that the respective proposal chapter is actually important relevant for the assessment – and that it is not enough to just copy and paste a standard template text. Many applicants therefore tend to delay drafting this chapter until submission deadline is imminent. Hence, last-minute requests at Dezernat 4 still occur way too often. In these cases, adequate counselling is barely possible, since RDM requirements may vary greatly between projects and disciplines. This is why we urge you to get to us early on when you are working on a proposal containing a RDM chapter or a data management plan. You will do a favour to us and to yourself. We are happy to help with the wording as well, but please, don´t expect ready-to-paste text templates.

Why funders require RDM

researcher — **produces** → **research data & results** ← **wants to check** / **wants to re-use** — (disciplinary) community of fellow researchers

Studie

i funders' key concerns:
- make sure that results can be validated
- avoid multiple funding

important prerequisites:
- data are accessible long-term
- data are readable and comprehensible long-term

video tip

This video perfectly illustrates a "worst practice scenario"…

Karen Hanson, Alisa Surkis, Karen Yacobucci (2012): Data Sharing and Management Snafu in 3 Short Acts. CC-BY 3.0 unported, doi: 10.5446/31036

to the video in the AV-Portal of TIB

Why do funders require RDM concepts anyway? In short, they want to maximise the impact of the deployed funds and at the same time make it easier to assess the quality and validity of research results.

On one side, there are the researchers, who produce research data and results based on these data. On the other side, there are their colleagues from the wider disciplinary research community. These people are interested in validating and comprehending the published results, drawing upon the data. Furthermore, they may want to build on the existing data and re-analyse them in the scope of their own studies, which may address different research problems. Both, validation and re-use are only possible if the community has access to the research data, if the data are readable and usable, and if an adequate documentation facilitates their comprehension. Funders want you to describe convincingly how you are going to process and document your data and how you will share them with others because by these means you can safeguard research quality. They also help to avoid unnecessary expenses caused by generating the same data multiple times in different projects. This video, in turn, illustrates very well a "worst practice" scenario.

Recommendations for RDM statements in funding proposals

commitment to relevant policies
- of the funder (e.g. good scientific practice (DFG), H2020 Open Data Pilot)
- the own university (for German universities, see list on forschungsdaten.org)
- the own discipline (see overview on the respective DFG website, at the bottom)

Policy

in case of joint projects: intention to write a project-internal RDM policy

DMP

intention to develop a data management plan

intention to elaborate data according to the FAIR principles and to make accessible for re-use open access

information on used (IT) infrastructure: what is available at your own university and beyond?

further reading

We compiled further tips and infos on our website. We also provide a manual for drafting internal RDM guidelines (in German).

go to the proposal infos   see the manual

Science Europe issued an extensive RDM guide. It also contains DMP tips for both, applicants and reviewers .

go to the Science Europe guide

When drafting a RDM chapter for your proposal, pay attention to the relevant requirements of the respective funding body or funding programme. The following statements, however, are almost always appropriate:

- Get informed on RDM-related policies and guidelines relevant to your project and proclaim that you will respect these regulations. The funding body, your own research institution and your disciplinary community may all have issued such policies.

- In case of large-scale joint projects, it is a good idea to develop a project-internal RDM policy, which defines general principles binding all partners and subprojects alike. If you intend to draft such a document, mention it in the proposal. Also, have a look at our how-to guide for internal policies.

- Even if the funder does not strictly require you to submit a data management plan with the proposal, announce that you will write one. This is usually a helpful and reasonable measure.

- Mention that you will process your data following the FAIR principles before making them available for re-use under "open access" conditions. Promise to publish your data by the end of the funding period at the latest. If this is not possible, state the reasons why you need to restrict access or delay publication. We will come back to the FAIR principles in the chapter "Archiving and publishing your data".

- Dedicate another paragraph or two to the IT infrastructure you are planning to use in your project. Explain, how the components meet your requirements, for example regarding capacity and security. Alternatively, state that the existing infrastructure does not meet your requirements and refer to the appropriate proposal section in which you apply for the funding of new or additional infrastructure.

Collecting, storing and processing data

Re-using existing data

Data storage and transmission

LUIS services for data storage and backup

Structured deposition and file naming

Handy tools to make live easier...

Special services und programmes
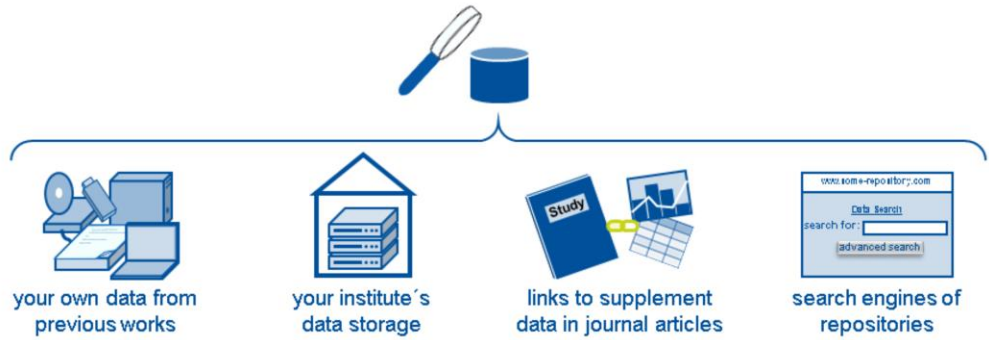
Sharing data (also with external partners)

This chapter addresses the implementation of data management in research practice. You will also learn about available methods, services and tools.

Re-using existing data

your own data from previous works

your institute´s data storage

links to supplement data in journal articles

search engines of repositories

www.some-repository.com
Data Search
search for:
advanced search

! Check before generating new data:

- Are there any existing data you could include in your own analysis (e.g. to broaden your database or for reference and comparison purposes)?
- If the answer is yes, do you have access to these data and all necessary permissions to process them?
- If the answer is yes, can you read and understand the data?
→ At this point you already notice what is important when making your own data available for re-use later on…

In the previous chapter, we pointed out that most funding bodies expect you to make your data available to others for re-use. Reversely, you should always check whether data that you need for your project already exist elsewhere and if you could use them. You can start your search at places like the following:

- Perhaps you yourself possess data from previous projects that you could incorporate well into the new undertaking.

- There may also be other researchers working in your team or at your institute, who deposited suitable data, for example at the institute's internal data storage. This occurs quite often in projects with frequently changing staff, such as Research Training Groups comprising several time-shifted cohorts of PhD students or large-scale long-term projects.

- Sometimes you may come across journal articles providing links to supplement material.

- And last, not least, you can use the search databases of data repositories. They work like a library´s OPAC, with the difference that the search result consists in a list of datasets instead of books and articles. We will have a closer look at this point later on.

If you found data relevant to your project, check the following:

- Can you access the data and are you permitted to re-use them?

- If the answer is yes, can you read and comprehend the data?

At this point, you already get an idea of what will be important when sharing your own data later on…
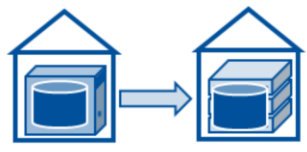
Data storage and transmission

! Make sure you dispose of adequate infrastructure and sufficient capacities _before_ generating any data!
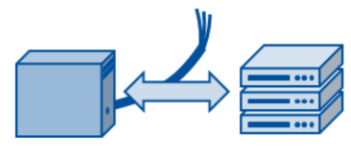
sufficient storage capacity?

adequate security levels of storage systems?

is an automated daily backup guaranteed?
(two copies in different locations)
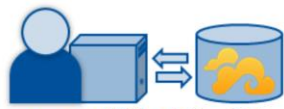
sufficient bandwidth?

Still before starting to collect data, ensure that you dispose of sufficient appropriate infrastructure. In particular, ask yourself the following questions:

- Does storage capacity suffice?

- Will there be an automatic daily backup? It is crucial that the backup is not located in the same building as the original. Otherwise, both copies could get lost in the event of a fire, for example.

- Does bandwidth suffice? This questions is relevant if have to transmit large volumes of data via the internet or an intranet. This may be the case during backups or when sharing data with external partners.

- Finally, you should ask yourself whether your storage systems comply with security standards, if you are working with sensitive data.
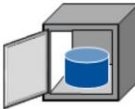
The central IT department of LUH provides a number of services for data storage and backup, which you can use free of charge:

- A personal cloud storage based on the software seafile is set up automatically for all LUH members the moment they activate the "single-sign-on" in their account manager. Once activated, you can use you personal storage to grant others access to chosen folders, including people outside LUH. This way, you can exchange or collaboratively work on data. You can also use seafile to synchronise a folder on your local hard drive or in a network drive with your cloud storage, thereby creating a constantly updated backup. For additional information, please refer to the service description.

- For projects, you may want to use a "Projektablage". This is whole bundle of services, which includes another seafile cloud storage. However, there are some functional differences in comparison with your personal cloud storage. They are outlined in the service description.

- If you want to store data safely for a long time, you can use the LUIS data archive. The archive is especially suitable to hold data that does not qualify for publication for legal or ethical reasons but still needs to be preserved according to good research practice.

- If you want to publish data but cannot find a suitable disciplinary repository, you can upload them to the LUIS data repository.

- If you store your data on servers of your institute during a project, clarify whether your institute uses the LUIS backup & restore service. In case it does, you can rest assured that your data is automatically backed up every night.

Structured deposition and file naming

date YYMMDD	...	editor	version
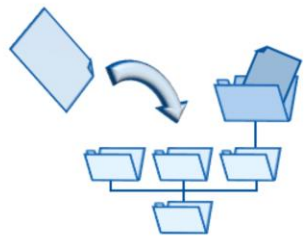
Example: 210519_WorkshopRDM_slides_draft_VS_v03.pptx

recommendations:
- consistently name files and folders according to a defined scheme
- use name components that provide meaningful information about the file´s content
- start with the component most often used for sorting
- avoid spaces and special characters

Once you established, where you would store your data in the course of your project, the next question to address is how you want to store them. In order to maintain orientation, it is crucial to organise them systematically in a structured manner. You should set up a basic folder structure right at the start of your project. This is what we recommend:

- Name your files and folder consistently, following a defined scheme.

- Use name components that feature meaningful information about the file and arrange the components in a fix order. In this example, the date comes first while the editor´s initials and a version number are last. The keywords in between tell something about the file´s content. However, completely different name components may be appropriate in other use cases.

- The same holds true for the components´ arrangement. Always place the component in the first position that is likely to serve most frequently as sorting criterion. If the name contains a date, it is best to use a numeric format that puts the year in front, followed by month and day. That way, alphabetic sorting corresponds to the correct chronological order.

- Another crucial point is: Avoid spaces and special characters in file and folder names. You may otherwise run into manifold problems when using programmes unable either to interpret these characters correctly or to recognise them as part of the name or path. The only exception is the underscore.

Handy tools to make live easier…

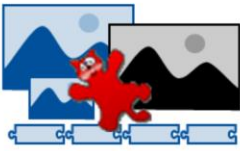systematically name lots of files at once using RenameMaster

go to download page

clean up tables using OpenRefine

go to download page

use PDF24´s OCR to make scanned texts searchable

go to download page

Document
Once upon a time there was a scanned text that could also be read by machines...

edit many images at once following a defined pattern (e.g. changing names, colours or size) using IrfanView

go to download page

Find more tools in the "tool box" on our website (currently in German only).

go to tool box

Manual data processing may absorb much time in everyday research. It is often possible to automatize such processing steps at least partly with the help of appropriate tools. Here are some free tools that may safe you a lot of time when carrying out routine tasks:

- If file names still do not follow your naming convention, it is unnecessary to rename them manually one by one. This is what tools like RenameMaster can do for you. Use them to change, delete or complement name components, or to automatically add a counter etc. with just a few clicks.

- When filling out spreadsheets it depends much upon the discipline of the editors whether they insert data in the correct format, which is a serious disadvantage. Cells intended to hold numeric values might contain text instead, or there is no consistency in using either points or commas as decimals characters, resulting in erroneous calculations. A manual correction can be rather time-consuming when dealing with numerous or large spreadsheets. OpenRefine lets you automate many clean-up operations at least partly, helping you to quickly transform your data into a usable format.

- If you are working with analogue documents such as archive material, you often scan or photograph them as a first processing step. Machines however, cannot read or search the resulting image files. Luckily, many programmes such as the free tool PDF24 nowadays include an optical character recognition feature to recognise text in raster files. The programme identifies all characters in an image file in a matter of seconds. You then have the option to save the text information along with the raster image in a pdf file.

- If the need arises to edit a larger number of images following an identical sequence of processing steps, you may want to use the batch feature of IrfanView. Use cases may include the creation of thumbnails or converting coloured photos into greyscale images.

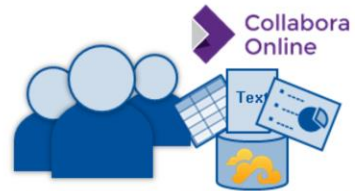By the way, we recently added a toolbox section to our homepage where we present some more useful tools.

Special services und programmes

Scientific Computing (using the LUIS computing cluster)

go to service description

Collabora Online

multi-user online editing of documents (also simultaneously) using Collabora

go to service description (German)

LICENSE — **Bug Creator 3.0**

LUIS central software and license acquisition

go to software catalogue (German)

data generation or processing by external service providers (e.g. transcription service, call centre, aerial photography etc.)

You may need additional special hardware, software or services to perform further analyses on your data and to document the respective processing steps.

You can use the LUIS computing cluster to carry out large-scale computing operations, such as complex simulations. Please reserve the necessary capacities in good time.

When working in a team, it is a common scenario that several people jointly create and edit documents. Merging different version manually, however, is as tedious as it is error-prone. As an alternative, move documents to your seafile cloud storage. From there, open them in Collabora Online. Collabora enables simultaneous editing of the same file by several users. Note, however, that compatibility with Microsoft Office is limited.

If you need special research software, have a look at the LUIS software catalogue. LUIS already purchased licences centrally for many of the numerous programmes listed there. They may also obtain additional licenses on request.

Consider as well, that sometimes it might be sensible and even cheaper to outsource certain processing steps to external service providers. Such services may comprise the transcription of audio recordings, telephone surveys or the taking aerial photographs, for example. When delegating jobs to third parties, make sure to conclude appropriate order processing contracts. If you desire additional counsel related to this subject, we are happy to broker contact with the legal department and the data protection office.
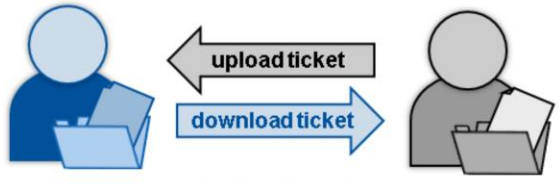
Sharing data (also with external partners)
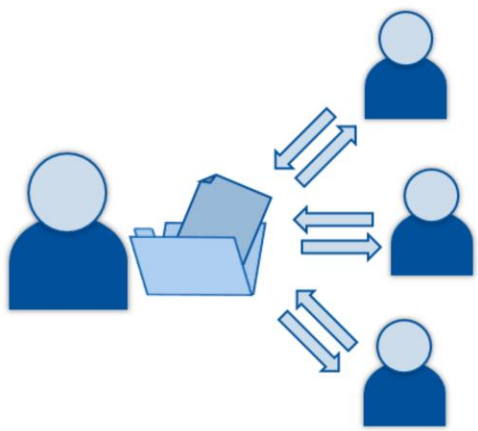
creating share links to files and folders
in the "Projektablage" or "Cloud-Seafile"

go to Seafile user manual

upload ticket

download ticket

LUIS Download Ticket Service
(as an alternative to big e-mail attachments)

go to service description (German)

Chapters

Introduction

Planning your data management

Specifications, guidelines and laws

Collecting, storing and processing data

Documenting your data processing

Protecting your data

Archiving and publishing your data

Supporting services and initiatives

Projects are often joint ventures, involving researchers from several institutes and working at different places. In these cases, performing effectively as a team requires easy sharing of data and joint access to files.

We already presented both versions of the seafile cloud storage. In particular, they both enable you to set permanent joint access rights for individual folders or even entire libraries. You can also grant external partners not invited to a "Projektablage" the right to access chosen folders via upload and download links. You can configure these links to expire automatically at a set date.

If you need to send or receive files up to 20 gigabytes large, there is another sharing option: LUIS provides a ticket service for creating upload and download links and sending them via e-mail. This way you avoid inflating the recipients' mailboxes. The service is available to you if you already have an account for a "Projektablage" or if you are using a LUIS mailbox.

Documenting your data processing

Why documentation is so important

The significance of metadata

General and specific context information

How metadata is generated and which forms they can take

Documentation – possible instruments

We already mentioned the importance of good data documentation in the previous chapters. In this section, we will introduce you to important concepts and possible tools for this task.

Why documentation is so important

data interim storage, model „Asse"…

Without good documentation you are in danger of …

- …failing to find specific data
- …being unable to retrace data creation and processing steps
- … being unable to interpret data due to lacking context information
- …mixing up files (outdated or contending versions)
- …not being able to interchange data with others or to combine them with data from other sources

An adequate documentation is good research practice!

Lack of or inadequate documentation can cause plenty of unwanted surprises, especially when an important deadline is approaching. This is because without good documentation, you are in risk of

- failing to find specific data,
- being unable to retrace data creation and processing steps,
- being unable to interpret data due to a lack of context information,
- mixing up files, for example when accidentally working in outdated or contending versions, and
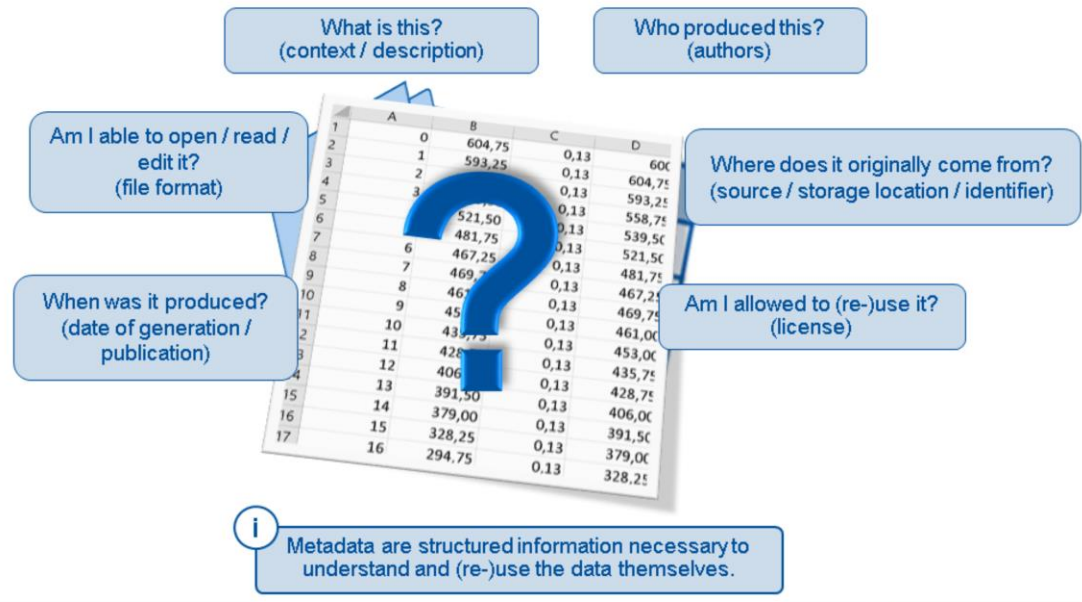- not being able to interchange data with other people or merge it with data from other sources.

Hence, please do not forget: adequate documentation is part of good research practice for a reason!

The significance of metadata

What is this? (context / description)

Who produced this? (authors)

Am I able to open / read / edit it? (file format)

Where does it originally come from? (source / storage location / identifier)

When was it produced? (date of generation / publication)

Am I allowed to (re-)use it? (license)

i Metadata are structured information necessary to understand and (re-)use the data themselves.

Does this sound familiar? Somewhere, you come across an unidentified file that may contain important information. However, you can't do anything with the data because you don't know the context. For example, you will ask yourself:

- What is this that I have in front of me?
- Who produced this?
- Where does it originally come from?
- Can I even open, read and edit it?
- When was it created? And
- Am I allowed to reuse it?

The answers to all these questions are metadata. Metadata is structured information that you need to understand and use the data itself.
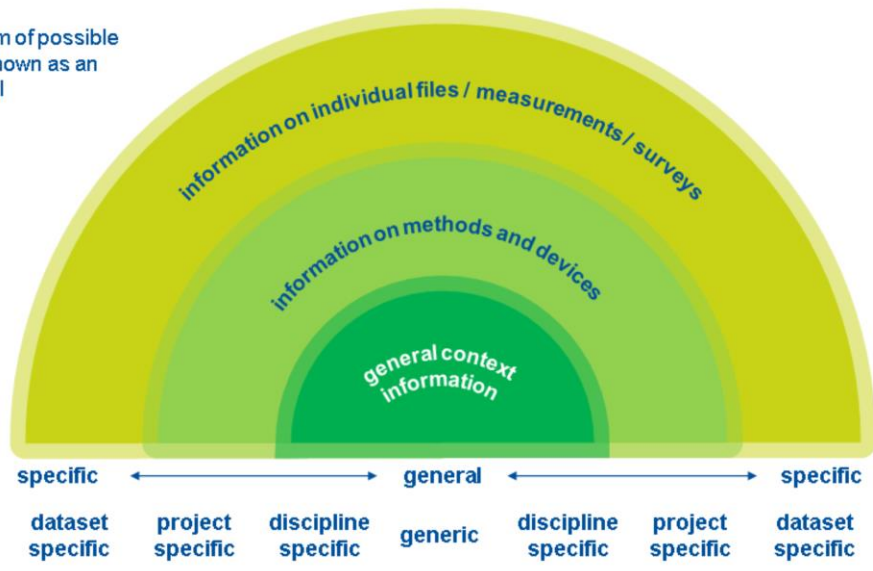
General and specific context information

the spectrum of possible metadata shown as an onion model

information on individual files / measurements / surveys

information on methods and devices

general context information

specific ← → general ← → specific

dataset specific | project specific | discipline specific | generic | discipline specific | project specific | dataset specific

Metadata may refer to very different things as illustrated with this onion model. At its core is the general information that is always important and hence should always be recorded. The answers to the questions in the previous slide, about authors or the project period, for example, would constitute such general information.

In this model, the metadata becomes more specific towards the margins. Some information may only be relevant in the context of a particular discipline or within a certain project. This is especially true for information that relates to specific methods and tools used to generate data.

The margins would be the place for details on individual files, measurements or survey processes, for example the file size, the location, day and time of an interview, the calibration parameters of a measurement device or weather conditions at the time of an observation.
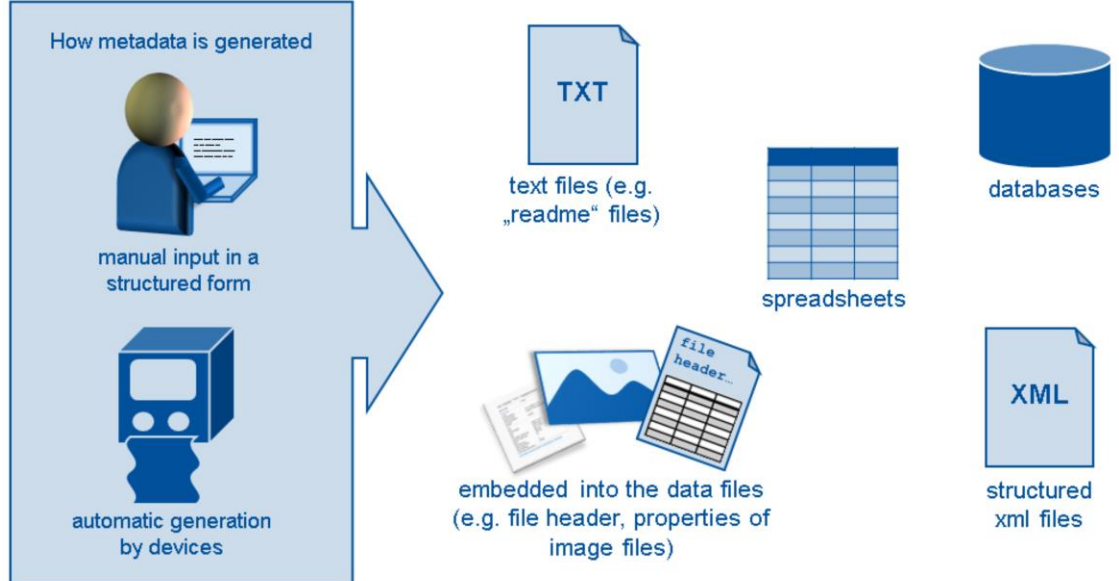
As you can see, very different metadata can be relevant for different projects, methods and data types. This is why it is important to determine which mandatory or optional context information to record and in which format, before creating any data. This is especially true for joint projects with several working groups frequently exchanging data.

How metadata is generated and which forms they can take

How metadata is generated

manual input in a structured form

automatic generation by devices

text files (e.g. „readme" files)

spreadsheets

databases

embedded into the data files (e.g. file header, properties of image files)
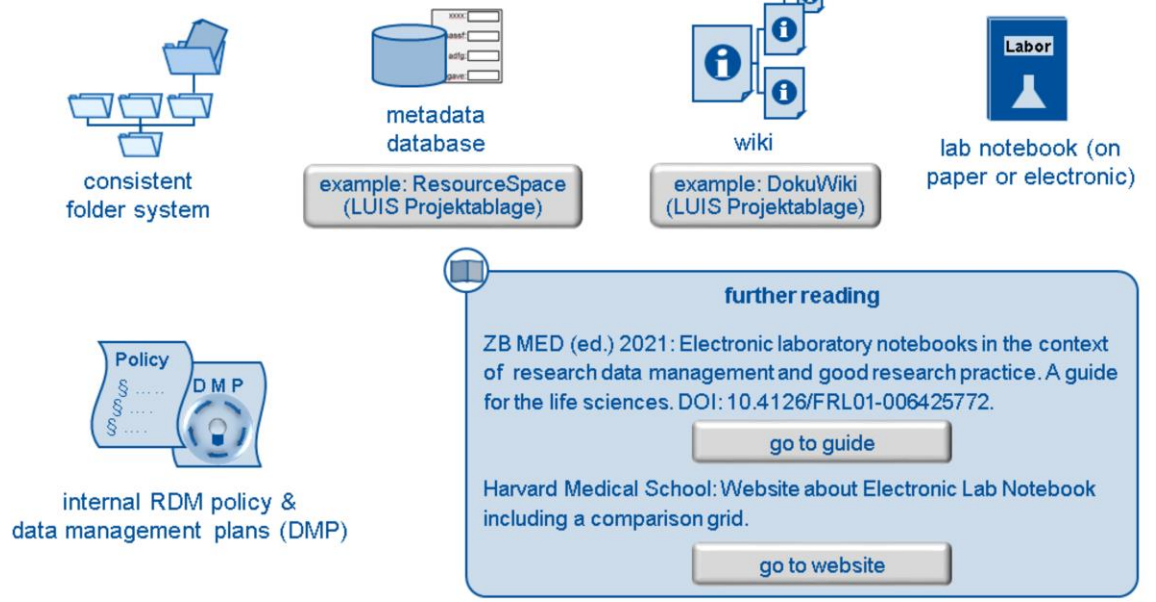
structured xml files

TXT

XML

Metadata is either recorded manually, for example in a table or database, or it is generated automatically by devices. For example, digital cameras not only store the images themselves, but also information on the camera model, the timestamp or even GPS data of the location. Many measuring devices also provide information on the device model and the applied settings in addition to the actual measurement data. Accordingly, metadata can exist in a wide variety of formats, for example as a manually edited README text file, as a spreadsheet or in a database. Some file formats also embed metadata directly in the data files. You know this from digital images when you look at the file properties, or from files with file headers. For metadata optimized for machine readability, the XML format is common.

Documentation – possible instruments

consistent folder system

metadata database

example: ResourceSpace (LUIS Projektablage)

wiki

example: DokuWiki (LUIS Projektablage)

Labor

lab notebook (on paper or electronic)

Policy

DMP

internal RDM policy & data management plans (DMP)

further reading

ZB MED (ed.) 2021: Electronic laboratory notebooks in the context of research data management and good research practice. A guide for the life sciences. DOI: 10.4126/FRL01-006425772.

go to guide

Harvard Medical School: Website about Electronic Lab Notebook including a comparison grid.

go to website

Various kinds of tools are suitable for documentation in general, some of which we would like to introduce briefly. On a very simple level, consistent folder systematics can already constitute a form of documentation, because the name components of folders and files should already contain information about the contents. This is usually not sufficient, of course.

Systems with an underlying database suit the purpose of a systematic storage and management of metadata especially well. If you have LUIS set up a "Projektablage" for you, the service ResourceSpace is automatically included. Originally developed for image databases, this software may also be used for metadata management in general. Albeit some adaptations will be necessary.

For knowledge management and the collection information in form of descriptive continuous text, the use of a wiki is a good idea. In simple terms, a wiki consists of interlinked web pages. You know this principle from Wikipedia. A wiki does not have to be publicly accessible, however. It may as well exclusively serve internal purposes. The LUIS offers the wiki software DokuWiki as part of the "Projektablage" bundle.

In natural sciences, it is common practice to record important context information on experiments in lab notebooks. Classically, these are physical paper books. However, the use of electronic lab notebooks is on the rise. They combine some features of a wiki with metadata management. If you consider using an electronic lab notebook in your project, the articles linked in the "further readings" box may help you decide, which system is best for you.

And finally, data management plans and internal policies form part of the project documentation as well. After all, this is where you document the principles, means and methods of your data management.

Protecting your data
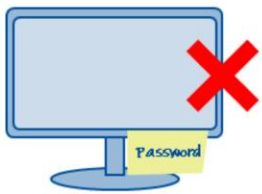
Handling passwords: NOT like this, please!

Generating secure passwords

Rights management

Encrypting files and folders

Encryption using VeraCrypt

Physical access restrictions

Protecting data from unauthorised access is becoming an issue of increasing importance. In this chapter, we present methods and tools you have at your disposal that may help you with this task.

Handling passwords: NOT like this, please!

Password

Do not stick your password on your screen!

An: all@project
Betreff: password

Dear all

The password is „ABC123"

Never send passwords as plain text via e-mail!

One password to rule them all...

Do not use the same password for different accounts!

tool tip

You can manage your passwords in a reasonably secure way using the password manger KeePassXC (not recommended for highly sensitive passwords). In this case you only have to memorise one master password

Tip: Save your password database to your personal LUH cloud storage, which can be accessed from different devices.

go to download page

In most cases, access authorisation work with passwords. A technical security measure is hence ineffective if users do not take good care of their passwords. Here are some tips that may sound trivial but that many people nevertheless ignore all too often:

- Please no not stick a post-it with your password written on it to your screen! By doing so, you enable everybody with access to your office to decrypt your data or tog log into your accounts! That holds true not only for your colleagues but theoretically also for the cleaning service, the facility manager or burglars. In principle, it may actually be more secure to write down passwords instead of storing them digitally. But please, keep them in a safe, lockable place, such as a vault.

- Never send passwords as plain text via e-mail! Not only is it easy to read along unencrypted mails during transmission. Sometimes people also carelessly forward an e-mail without checking beforehand whether sensitive information not intended for the recipient is still lurking somewhere in the prior correspondence.

- Last, not least, avoid using the same password over and over for multiple accounts. You are otherwise risking the misuse of all of your accounts once the password ends up in the wrong hands. There is no doubt that, in practise, it is completely unrealistic to memorise unique passwords for everything. Whenever dealing with sensitive data or critical infrastructure, however, you should create individual passwords at least for these accounts.

Password managers like KeePass are a practical solution for the administration of less sensitive login details. When using such a programme you only have to memorise a single secure master password.

Generating secure passwords

User: JohnDoe

Password: ****************************

further reading

ConnectSafely (May 2020): Tips for Strong, Secure Passwords & Other Authentication Tools. Website

go to website

tips for secure passwords:

- at least 20 characters
- hard to guess (do not use birthdays etc.)
- phrase instead of a single word
- no obvious sense (no common proverbs etc.)
- altered spelling (e.g. „tr!ck" instead of „trick")
- combining several languages

Avoid language specific characters, such as ö,ä,ü,ß etc. in passwords. These characters may not be available on devices with the keyboard layout of another country. You may thus be unable to enter your password!

Theoretically, a password´s security depends solely on its length. This hold´s true, at least, in case of so-called brute force attacks that work by simply testing all possible combinations of characters. At the current state of information technology, a length beyond approximately 20 characters allows for so many possible combinations that very few attackers dispose of the necessary patience and computing power. This is why passwords for sensitive accounts should be at least 20 characters long.

In practice, however, attackers will try out more likely character combinations first. Such combinations may be popular passwords like "123456", or words listed in an electronic dictionary. In case of pinpointed attacks, researched data related to a certain person and his or her personal environment may also be tried out. Such data comprise, for example, birthdays, phone numbers, number plates and nicknames. Combinations of such data and the reverse character order of these words and numbers are unsafe as well. You should hence refrain from using researchable personal information in passwords!
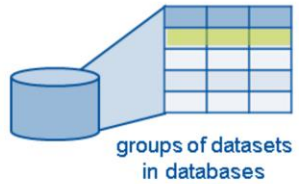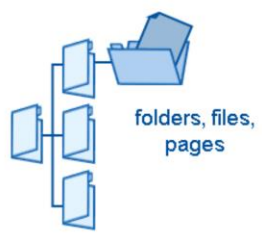
You may find it easier to memorise a long password if you think of a pass phrase, say, an arrangement of several words. The phrase will be much harder to guess if it does not make sense at first glance. Therefore, a popular proverb is less suitable. You can also modify the spelling of words, for example, by replacing letters with similar looking special characters. Finally, you can combine different languages by mixing German and English words, for instance. If you speak Breton, by chance, that´s even better… Please find additional information on the ConnectSafely website linked here.

As a final remark: Don´t use language-specific special characters in passwords! In the worst case, you may be unable to enter your password when using a device with a foreign keyboard layout!

Rights management

Some data management systems let you define access rights. Such rights may apply to individual files and folders or to entire folder trees, but also to specific groups of datasets in a database, for example. The level of detail for user rights settings vary according to the software or the file system. At the simplest level, administrators can usually choose whether to allow reading only or reading and writing. Or they may simply deny access at all. When editing rights are finer-grained, a user may have permission to edit a file but not to delete it, for instance.

In most cases, it is a good idea to leave the setting of user rights to IT professionals, since erroneous configurations may cause loss of data or security breaches. Consider this point when planning your project and allocating resources.

Permissions may be set per individual user. Larger systems with lots of users, however, generally apply group policies. Such policies set the rights for a whole group of users, meaning that every group member inherits all the rights defined in the policy. A single person may nevertheless belong to several groups and hence inherit different sets of permissions.

Many rights management systems also provide a feature to log file access attempts. With the help of the log file, it would be possible to retrace a compromised account in case of security incidents.

Encrypting files and folders

! The more sensitive the data, the more important is _continuous_ encryption!

recording devices

individual files and folders

mobile storage media

laptops

workspace PCs

server
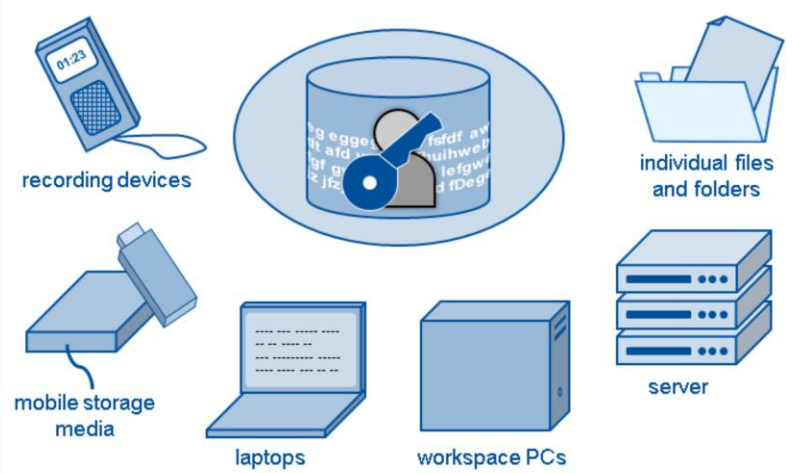
further reading

Aseem Kishore (21 July 2019): How To Encrypt Zip Files. Online article at Help Desk Geek.

[go to online article]

WinRAR documentation.HELP!: Archive name and parameters dialog: general options. Set password.

[go to documentation.HELP!]

Dave McKay (23 October 2023): What Is Encryption, and How Does It Work? Online article at How-To Geek.

[go to online article]

Chapters

- Introduction
- Planning your data management
- Specifications, guidelines and laws
- Collecting, storing and processing data
- Documenting your data processing
- Protecting your data
- Archiving and publishing your data
- Supporting services and initiatives

Some data are subject to legal or ethical requirements that demand special security measures to protect them from unauthorised access. This holds true for personal data and for data relevant for patent applications, for example. In other cases, you may have guaranteed by contract to protect certain data from disclosure. The most important technical measure to protect such sensitive data is the encryption of files and entire drives. Before collecting any sensitive data, please think carefully about storage locations in the course of a project, both, temporary and long-term. Analogous to perishable food requiring an uninterrupted cold chain, make sure that your sensitive data is encrypted continuously.

If you use recording devices, such as voice recorders, these should support encryption. Small, mobile devices are especially likely to get lost, and you don´t want a thief or a casual finder to be able to read the sensitive data they are holding. The same is true for usb sticks and external hard drives, by the way.

Data on laptops with unencrypted hard drives are easy prey as well. The login password of your operating system does not protect you, by the way, because it is simply possible to boot the device via a live CD, or to remove the internal drive and connect it to another computer. Laptops are quite frequently stolen or just forgotten in the train…

The loss of workspace PCs is less likely, but break-ins into university building certainly do occur. If you store your original data or the back-up copies on servers of your institute or of the central IT services, however, these servers are generally located in well-protected rooms.

Last, not least: Please don´t forget individual files or archives that you just send quickly as an e-mail attachment or that you upload to a commercial file hoster. You should avoid both when dealing with sensitive data. If there really is no way around, at least make sure to encrypt the files beforehand! As describes in the linked articles, you can do that quite easily by setting a password when creating an archive file using programmes like 7zip or WinRAR.

Encryption using VeraCrypt

! Generally speaking, encryption by software is more reliable than encryption by hardware!
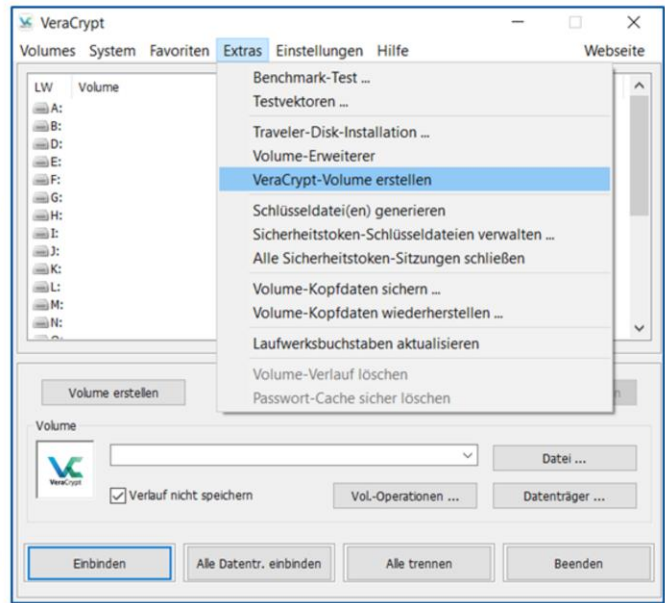
tool tip

VeraCrypt is a powerful, user-friendly, open source programme capable to securely encrypt individual folders, partitions and entire drives. It is available for all common operating systems.

[go to download page]

Techlore`s YouTube video explains well the advantages, features and configuration choices of VeraCrypt.

[go to video]

Many up-to-date devices like Laptops, smart phones and external SSD drives provide built-in encryption via an internal microprocessor. This feature is not always active by default, however. In the past, serious security breaches in hardware-driven encryption were discovered more than once. Sometimes, not even firmware updates could solve the problem.

Encryption by software is hence more reliable, at least in general. The Windows operating system provides the programme Bitlocker for this purpose, but it is only available in the Pro-version. The MacOS counterpart is FileVault. However, we recommend using the free and open source software Veracrypt instead, which runs with all common operation systems.

VeraCrypt lets you encrypt partitions or entire drives, as well as individual file containers. Such a container works like a folder that and contain an unlimited number of files and subfolders. Otherwise, it behaves like an ordinary file that you can move, send or delete as usual. Techlore´s YouTube video tutorial explaining Veracrypt´s features and configuration options is very instructive. Just follow the link.

Physical access restrictions
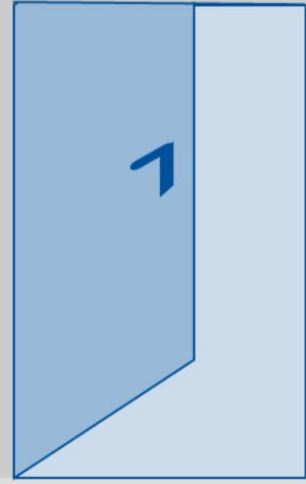
- Which media contain the originals or copies of the data?
- Where are these media located?
- Who has legally access (including janitors, cleaning service etc.)?
- How easy would it be to gain access illegally (burglary)?

Choose storage locations whose security and access restrictions are adequate to the sensibility of your data!

further reading

A recent example:
Maïa de La Baume (2020): Thieves use lockdown as cover for EU Parliament burglaries. (online article from 1 July 2020 at POLITICO.eu)

go to article

Most hackers trying to gain unauthorised access to your data will launch their attacks via the internet. However, even without any network connection, attackers may go as far as to obtain direct physical access to the storage media. If you are working with data that could be of interest for ruthless and highly skilled attackers, put some thought into the physical access situation to your drives and servers. In particular, you should ask yourself the following questions:

- Which media hold the originals or backup copies of my data? Don´t forget e-mail attachments, external media, prints on paper and backup server.

- Where are these media located? An additional security backup on your private usb stick in your pocket may quickly turn into a security breach. When using a cloud storage, try to find out about the servers' location and security measures. If you are unable to retrieve this information, do not use the service.

- Who has authorised physical access? If you store your data on the internal hard drive of your workspace PC persons with access may include your colleagues, the facility manager or the cleaning service. In case of central servers, at least the facility´s IT administrators and maintenance personnel have access.

- How easy would it be to gain access illegally by breaking into the rooms containing the storage media? A windowless server room with doors of steel is certainly safer than a ground level office with large windows.

If such considerations sound somewhat paranoid to you, why don´t you have a look at the article linked in the "further reading" box? After a series of burglaries at the EU parliament apparently targeting data storage devices, the question emerged whether the parliament´s own security service might be involved…

Additional organisational measures

HOW TO PROTECT YOUR DATA

Train and sensitise all persons involved in data processing.

Avoid "free" commercial cloud storages, file hoster etc.

R.I.P.

See for a professional (!) physical destruction of defect media and media no longer needed.

(i) You can hand over electronic media designated to be destroyed to Sachgebiet 12 (IuK) in the main building "Welfenschloss" (R318A).

If evitable, do not store data on mobile devices. If you must, remove the data from these devices as soon as possible.

Check regularly if everybody follows the rules. Define workflows and responsibilities!

further reading

Surveillance Self-Defence (2019): Your Security Plan. Online article, 1 October 2019

go to article

Apart from the technical provisions, you can further enhance the security of your data by applying simple organisational measures. It is of uttermost importance that all persons involved in processing of sensitive data are sensitised and receive adequate training. For them, participating in suitable courses by the Research Data Support Team or the Data Protection Office could be mandatory.

Don´t upload your data to locations completely outside your sphere of control. If you use "free" cloud storage services like Dropbox, Google Drive and so on, you have no chance to monitor where your data goes and who has access to them. Please avoid using these services! If there really is no way around, only upload data that is already encrypted. You could use file containers or programmes like Boxcryptor and Cryptomator for this purpose.

If you need to discard defective storage media that once held sensitive data, ensure their physical destruction by professionals! If you don´t, there is a risk that someone is retrieving the media from the electronic scrap in order to restore the data. Suitable equipment and knowhow provided, this isn´t even particularly difficult in many cases. You can hand over defective media to the colleagues of Sachgebiet 12 in the main building. They will see for a proper destruction.

Avoid saving data to mobile storage media, because these media are especially likely to get lost or to be stolen from you. If you don´t have a current backup, you may suffer a loss of important data, but apart from that, the thief or a casual finder may misuse the data. Encryption minimised the latter risk but cannot rule it out completely.

Last, not least, establish regular controls in your projects to ensure that everybody sticks to the rules when handling sensitive data. Define workflows and responsibilities, preferably in writing. A data management plan or a project-internal policy can be suitable documents to record such agreements. You could also work out a special security plan following the manual linked in the "further reading" box.

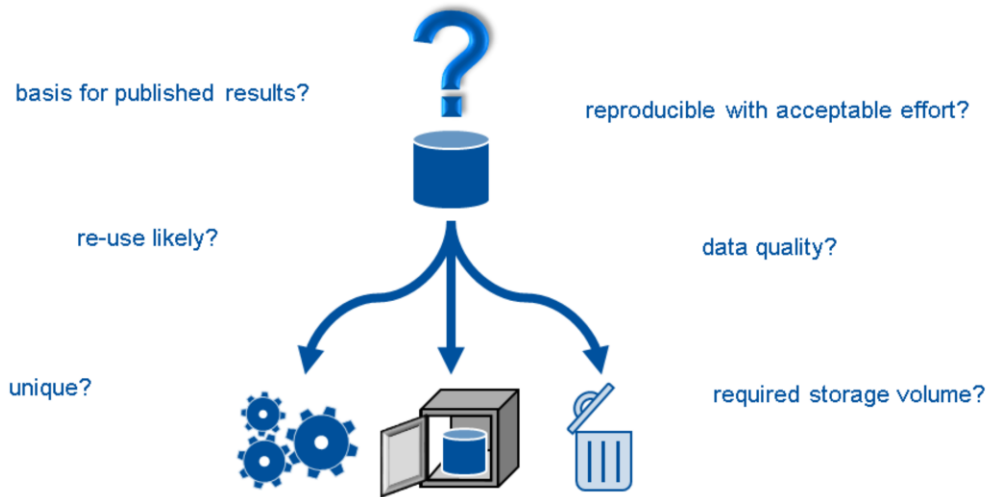# Archiving and publishing your data

Chapters

In line with good research practice, you should store relevant data long-term and make them publicly available, if possible. In this chapter, we explain options and methods for archiving and publishing research data professionally.

Many researchers ask themselves whether they really have to keep all data at all times. This question can clearly be answered in the negative. It is neither practical nor sensible to keep everything all the time, and you are not expected to do so. However, you should always decide consciously and with good reason, which data to keep for how long and which to delete. The following criteria may provide some guidance:

- Is there a scientific publication based on the data? In this case, keeping the data for no less than ten years is generally mandatory.

- Is a re-use likely? If you cannot imagine any scenario in which you or anyone else would ever be interested in the data again, it may not be necessary to keep them.

- Is your data unique and unreproducible in the same form? This would be the case with weather records or opinion polls, for example. In case of doubt, it is preferable to keep such data.

- Is your data reproducible with reasonable effort? For example, simulations should always produce the same results if the applied software, algorithms and input data are identical. In such cases, it may be acceptable not to keep the resulting data. However, it might be necessary to archive the used software environment instead.

- Of which quality are your data? By tendency, we recommend keeping the raw data and the final products. In contrast, you may want to delete file versions that represent intermediate stages or contain errors. However, please keep the metadata! This way it remains traceable that the data once existed, and the reason for the deletion is on record as well.

In case of data-intensive research, even keeping those data of potential long-term value may sometimes be impossible because of limited storage capacities. Although extending these capacities should be the first option, at some point, there may be no way around deleting larger or older files in favour of smaller or newer ones with equally relevant content.

Refining data according to the FAIR principles

Chapters

Findable (others can find my data)
- rich metadata in searchable public registries (e.g. the search databases of repositories)
- permanently retrievable via a unique identifier (e.g. a DOI)

Accessible (others have access to my data)
- preferably online access via standard protocols (e.g. http(s) / ftp)
- transparent access conditions

Interoperable (my data is compatible to other data and can be processed by machines)
- systematically elaborated and documented according to disciplinary standards
- machine-readable data and metadata in common and preferably open file formats
- references to other related data

Re-usable (other can utilise my data for their own purposes)
in addition to all points listed above:
- adequate documentation and precise attributes (data is comprehensible)
- distinct license (conditions for re-use are defined in a legally binding manner)

further reading

Angelina Kraft (2017): The FAIR Data Principles for Research Data. Article in the TIB blog, 12 September 2017.

[go to TIB-Blog]

GoFAIR: FAIR Principles (Website).

[go to GoFAIR website]

We have already briefly mentioned the FAIR principles on several slides. FAIR is an acronym made up of the first letters of Findable, Accessible, Interoperable and Reusable. These terms describe the form in which data should be available in order to optimise long-term usability. Let's find out what the actually mean.

"Findable" means that others can find your data. What´s the point in preserving data if hardly anyone knows where they are stored or that they even exist? That´s why you should upload the metadata into publicly searchable directories, for example into the search database of a repository. By assigning a so-called "persistent identifier" to your data, they can be found reliably via a permanently valid link for many years to come. The "Digital Object Identifier", abbreviated as DOI, is a particularly well known Identifier.

Your data are "accessible" when others can actually take hold of them. If possible, they should hence be retrievable online via a standard transmission protocol such as http or ftp. Under certain circumstances, it may be necessary to restrict access. In such a case it is nevertheless important to clarify, why access is restricted, who may still retrieve the data and which are the conditions for data sharing.
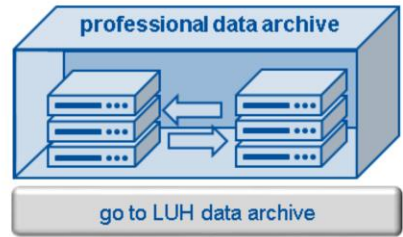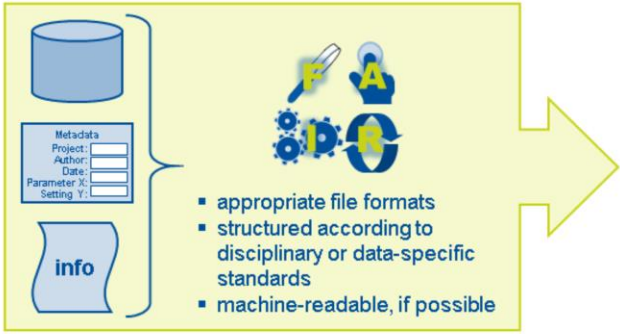
Data are "interoperable" if they can be combined with similar data from other sources or processed by other programmes than those they were generated with. The same applies to metadata. This is why we recommend choosing open and widely used file formats whenever possible. You should also structure and format data and metadata according to established standards, if such standards exist. We will elaborate on this point on the next slide.

Data intended to be "reusable" by yourself as well as by others not only has to comply with all the points mentioned so far. What is more, the data´s accompanying documentation needs to be comprehensible even for people who were not involved in the collection process. Clear conditions for data re-use should be defined in a license.

Correctly archiving data

Metadata
Project:
Author:
Date:
Parameter X:
Setting Y:

info

- appropriate file formats
- structured according to disciplinary or data-specific standards
- machine-readable, if possible

professional data archive

go to LUH data archive

i what a data archive should offer:
- servers located in appropriate rooms
- redundant data storage
- periodic replacement of media
- maintenance by professionals
- bit stream preservation guaranteed for at least 10 years
- access limited to data owners

i You can find lists of file formats appropriate for long-term storage all over the internet, e.g. on this website of Duke University Libraries:

go to list

The Research Data Alliance (RDA) compiled a catalogue of disciplinary metadata standards.

go to catalogue

! CDs, external USB drives and sticks etc. are NOT appropriate!

The FAIR principles not only apply to published data but also to data archived for verification purposes, for example. Besides the data themselves the same holds true for the related metadata and documentation materials. Make sure that they are all available in long-term readable file formats, structured according to disciplinary or data-specific standards and documented in a comprehensible way. You can check for suitable file formats and relevant standards on the websites linked in the box .

When archiving your data long-term, please use a professional data archive like the one provided by LUH. Such an archive should meet the following requirements:

- The servers should be located in suitable rooms with special fire protection and safety standards.

- All data should be stored redundantly so that a backup exists at all times.

- Storage media should be replaced regularly as they wear out over time.

- The infrastructure is maintained by appropriately trained IT personnel.

- The so-called bit stream preservation, which means an error-free preservation of the data, should be guaranteed for at least 10 years. And:

- Access to the data should be limited to data owner.

Please note that CDs, hard drives, flash drives and so on are not suitable for long-term archiving. These media degenerate over time, even and especially if they are not in use. Files stored on such media are in danger of becoming faulty or even unreadable after some years.

Publishing high-quality data opens up opportunities!

increased impact on the scientific discourse

special trust in the integrity of your research

others follow your example

additional citations when others re-use your data

additional impact and recognition (if data are relevant and of high quality)

gaining profile beyond one´s own disciplinary community

further reading

Max Planck PhDnet: Incentives to adopt open science practices in your daily research. Online article, 30 June 2020.

go to website

further reading

open-access.net: Open Access and Research Data. Undated online article.

go to website

Chapters

Introduction

Planning your data management

Specifications, guidelines and laws

Collecting, storing and processing data

Documenting your data processing

Protecting your data

Archiving and publishing your data

Supporting services and initiatives

We strongly recommend that you elaborate your research data according to the FAIR principles and make them publicly available whenever there are no legal restrictions. By now, most funders and research organisations expect you to do so anyway, but apart from these external requirements, data publication is also in your own interest. If your data is of high quality and relevant to a wider circle of researchers, sharing them publicly will bring you additional attention and recognition.
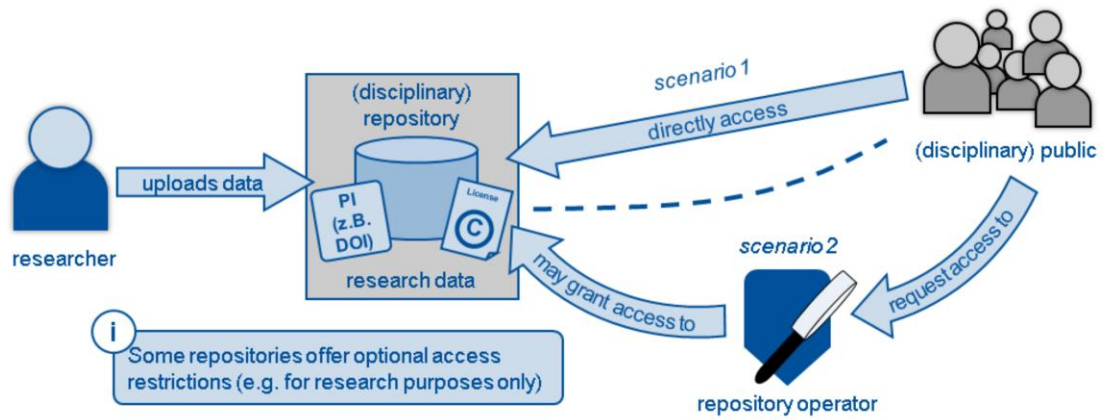
Data can and must be cited in the same way as text publications. Anyone using them without indicating their source violates the principles of good research practice, apart from potential copyright issues. The data also draws attention to your related articles. In effect, others will cite you more frequently, which, in turn increases your influence on the scientific discourse. Since the availability of data enables objective verification of your results, trust in the integrity of your research will increase. You may even assume a pioneering role in disciplines where data publications are still uncommon. If you prepare your data in an exemplary manner and make them available, others will follow your example. Many data are relevant across disciplines and re-usable in different contexts. Hence, your publications also constitute an opportunity for gaining reputation beyond your own disciplinary community. This may sometimes lead to exciting interdisciplinary cooperations.

The websites linked in the further reading boxes describe further benefits of practicing open science in general as well as current positions of academia and research funders.

Making data available via a repository

If you intend to publish data, we recommend using a repository specialised in your discipline or the type of data you are going to upload. A repository or data centre stores and manages files and related metadata. It works like an online archive, where you can find - and often download - relevant data sets via a search database filled with metadata. We will explain where and how to look for a suitable repository on the following slide.

Good repositories assign a so-called persistent identifier to each data set. This is a permanently valid link reliably leading to a landing page from which the data - or at least the metadata - are retrievable long-term. Particularly well known are the "Digital Object Identifiers", abbreviated as DOI.

Conditions for a re-use of your uploaded data by others should be clearly defined. This is why public data are generally subject to a standardised licence, for example, a Creative Commons licence.

In case of access-restricted data, interested parties may have to sign an individual user agreement instead. Such documents contain provisions on confidentiality and data protection measures. Some disciplinary repositories automatically restrict access as soon as personal data is involved, or they let data providers define to whom data access may be granted and under which conditions.

Accordingly, third parties may access data in one of two ways: In scenario one the data is completely public and anyone can access them directly. In scenario two, data access is restricted and requires a formal request. In this scenario, anyone interested in the data would contact the repository operators who would then check whether the respective person qualifies. As an example, he or she may have to demonstrate a plausible research interest. If the person qualifies, access to the data is granted after signing a user contract.

Choosing a suitable repository

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

go to re3data.org

How is access regulated (open, restricted)?

How is data re-use regulated (e.g. Creative Commons licenses)?

Do the data get persistent identifiers (e. g. a DOI)?

Is the repository certified (e. g. Data Seal of Approval)?

Does the repository have its own policy?

! Also pay attention to these points:

- Who is allowed to *upload* data?
- Which data and metadata formats are accepted?
- Which (perhaps optional) services do they offer?
- How much do they charge for hosting your data?
- → Preferably, clarify these points BEFORE the start of your project. You will then be able to allocate or apply for the necessary funds, and you can generate your data straight away in the required formats.

i LUH provides its own institutional data repository
go to service description

Which repository is suitable for publishing your data? The re3data.org website now lists more than two and a half thousand repositories from all over the world. However, the majority allow data upload only for members of specific institutions or projects. When browsing the search results, note the small icons that help assessing quickly which requirements a repository meets. The re3data website provides further details when hovering your mouse over the icons. However, setting the filter options according to your wishes right away is probably the easiest way to find what you are looking for.

If you look for a place for publishing your own data, make sure that your repository of choice assigns a persistent identifier and regulates data use in a clear and legally binding manner. Consider the following points as well:
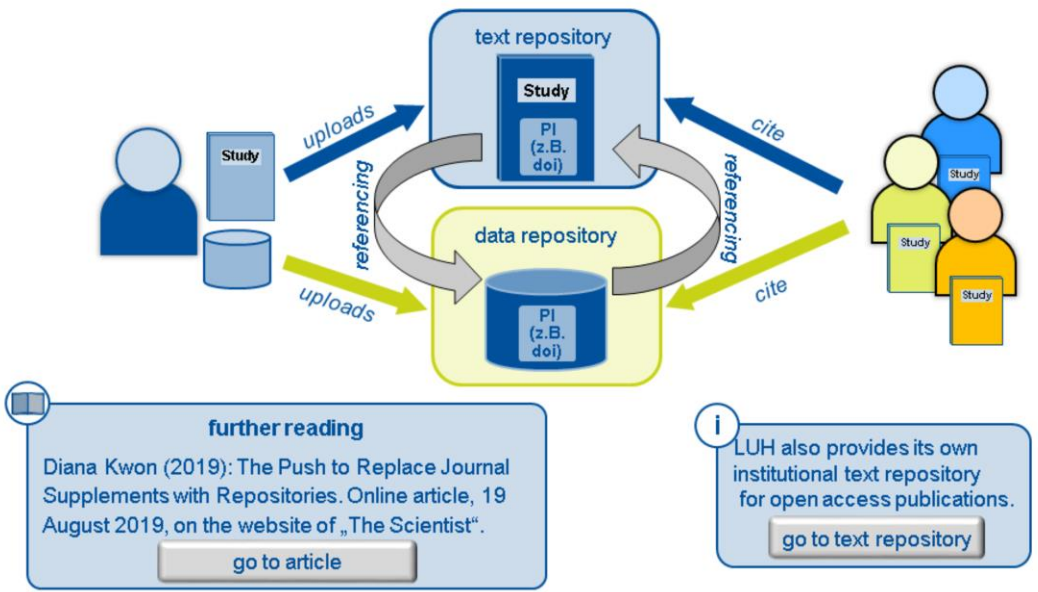
- Who can actually upload data? As we mentioned, upload is often restricted to a certain this group of people.

- Which data and metadata formats does the repository accept? If you search for suitable repositories before starting to collect data, you can store your data and metadata in the required format straight away, which will spare you to effort to convert them later on.

- Which services does the repository provide? Repository operators may offer more than just storing and publishing data. For example, they may check data quality or see for a conversion of files into current formats over time.

- What are the costs for data upload? They vary greatly and do not only depend on volume and the guaranteed preservation period. Calculations may also consider additional curation services. Remember that many funding bodies cover such costs if you include them in your grant application, so try to obtain quotes from service providers in due time.

If you couldn´t find a suitable disciplinary repository, of course you can always opt for upload to the generic LUH repository free of charge.

Linking journal articles with related data

text repository

Study

PI
(z.B.
doi)

uploads

referencing

cite

referencing

data repository

PI
(z.B.
doi)

uploads

cite

Study

Study

Study

Study

further reading

Diana Kwon (2019): The Push to Replace Journal Supplements with Repositories. Online article, 19 August 2019, on the website of „The Scientist".

go to article

i LUH also provides its own institutional text repository for open access publications.

go to text repository

If a text publication bases on published data, both publications should refer to each other in their metadata. Ideally, you upload your data to a data repository and receive a persistent identifier. There are also repositories for text publications that function in the same way. Hence, you should get an identifier for your text publication as well. This is also the case with many repositories operated by publishers. If you did not get an identifier from them, you can also publish your articles in the LUH text repository as a secondary publication. In that case, you will definitely receive a DOI. This is legally possible, however, if you already ceded your copyright as an exclusive right to a publisher.

In the metadata of your published data set, you can now specify the identifier of the associated text publication. This way, anyone who comes across the data will also find the article containing analyses and interpretations. Likewise, you can specify the identifier of the data set in the metadata of the article so that readers can get directly to the underlying data. Your data are now just as citable as your articles. A citation should always include the persistent identifier in order to ensure reliable findability of corresponding articles or data sets for many years from now.

As a final remark: Many publishers provide their own repositories where supplement data is stored along with the related articles, so text and data are stored in the same place. We do not recommend choosing this option, however, as the long-term availability of the data is not always guaranteed. The article linked in the further reading box elaborates on this issue.

Supporting services and initiatives
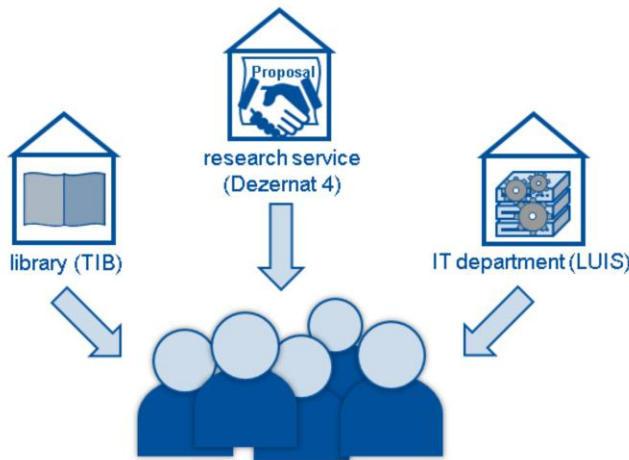
The Research Data Support Team of LUH

External information and support

RDM working groups and initiatives - national and international

This course offers a general overview over different aspects of research data management but does not go into detail. Consequently, it may have *raised* more questions than it *answered*. But don´t worry, there are numerous sources of information and support that you can consult. In this final chapter, we would like to present some of them.

You already know that this course is provided by LUH´s Research Data Support Team. Let´s have a closer look at our services in general. Our team comprises RDM experts from the library, say the TIB, from the Research Service at Dezernat 4 and from the IT department LUIS. Apart from public training courses about diverse RDM-related subjects, we are also happy to organise training sessions on demand for a closed circle of participants, such as members of a specific institute or larger project. Of course, you can always request individual counselling as well. Our principle foci of competence are:

- Documentation and publication of data

- Data management plans and RDM policies

- RDM statements in grant applications

- RDM-related services and infrastructures available to you at LUH

- Questions related to legal issues, in close cooperation with the legal department and the data protection office, and

- Practical implementation of data management measures in research processes

Please find further information, guides, events and more at our website. Just have a look!

External information and support

WISSENSCHAFTSBEREICHE

Chemie	Geisteswissenschaften	Geowissenschaften
Informatik	Ingenieurwissenschaften	Lebenswissenschaften
Mathematik	Physik	Psychologie
Sozial- und Verhaltenswissenschaften	Sportwissenschaften	Wirtschaftswissenschaften

go to overview at forschungsdaten.info (German)

ⓘ Disciplinary services may provide the following:

- counselling of projects from a disciplinary perspective
- discipline-specific info materials
- discipline-specific services, tools and infrastructure (e.g. a DMP tool, data processing, quality checks, archiving and publishing data)
- networking in the disciplinary research community
- developing and defining disciplinary standards and policies
- representing the interests of your discipline towards policy makers and funding bodies

Get information on services for your discipline and current developments!

There is much more help and support beyond the university. For example, you can find general information related to RDM at the portal forschungsdaten.info. We already placed links to this website on several previous slides. It also includes an overview over disciplinary services and initiatives that may offer the following services:

- Counselling for projects from a disciplinary perspective
- Disciplinary information material
- Disciplinary services, tools and infrastructures
- Disciplinary networks and communities
- Development and definition of disciplinary standards and guidelines, and
- Representation of your discipline´s interests towards policy makers and funding bodies

Find out about available services relevant to your discipline, but also keep an eye on newly emerging trends and developments, regarding community organisation or disciplinary standards and best practice, for example.

RDM working groups and initiatives - national and international

DINI/nestor-AG Forschungsdaten
(RDM support staff at research institutions
in German-speaking countries)

go to the working group's website (German)

Research Data Alliance (RDA)
(international network of RDM specialists
with a special focus on technical aspects)

got to RDA website

NFDI consortia

The federal government and the state governments jointly fund the development of a national research data infrastructure (NFDI). Its core are disciplinary consortia composed of researchers and infrastructure providers. These consortia are tasked with developing disciplinary standards, establishing services and organising the research community.

All researchers are invited to participate in the work and to join in the discussions!

LUH's NFDI website   official NFDI website   DFG`s NFDI website

In addition to the disciplinary services, there are cross-disciplinary initiatives and working groups. Here are some examples:

The joint working group of DINI and nestor is a network of the RDM support staff at research institutions in German-speaking countries. It not only serves its members as a platform for discussion and exchange but also organises cross-institutional training courses and conferences, and issues guides and other information material.

The Research Data Alliance targets a similar audience, but focuses largely on technical aspects, such as infrastructure and data standards. The RDA is a global network with continental and national chapters.

Of particular importance to researchers are the disciplinary consortia that emerged in the scope of the national research data infrastructure, NFDI. These consortia receive long-term funding from the federal government and the states and are expected to establish disciplinary standards, to develop services and to organise the research community. This is where the course is currently set for research practice for many years to come. Researchers are hence well advised to participate in these projects and discussions early on. So, try to keep up with the latest developments on this matter.

Thanks for your attention and have a nice evening!
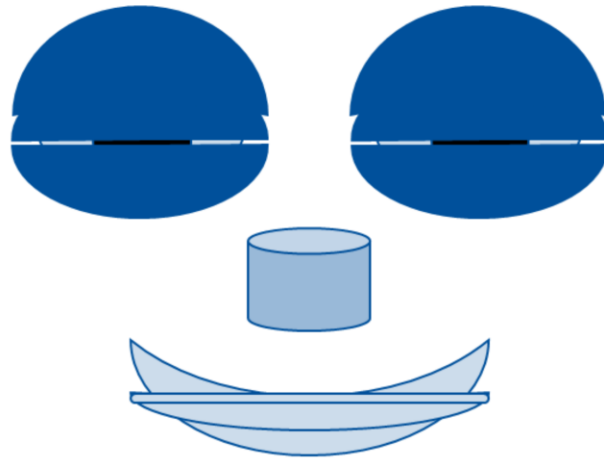
Wow, you finally made it to the end of this presentation! Congratulations, you should now have a solid basic knowledge on the most important aspects of research data management.

We hope that this course was comprehensible and that it met your expectations. If you have further questions or seek individual counsel, please don´t hesitate to contact our Research Data Support Team.

We plan to offer more courses in this asynchronous online format in the month and years to come, so pay our website a visit from time to time. As for today, we say goodbye and wish you much success in your research. Have a nice evening!