Research Data Management
A Practical Introduction

Hartmut Schlenz, Torsten Bronger, Michael Selzer,
Britta Nestler, Leo Riem, Salome Enahoro

Contents

1. Introduction

The professional management of research data is becoming increasingly important in science and is now strongly recommended or even required by many third-party funding bodies (DFG, BMFTR, BMWE, EU, etc.) for the approval of funding for new research projects. The main motivation for this is to create a comprehensive collection of research data, the production of which is primarily financed by tax revenue. The focus here is on good scientific practice and the traceability of research results. On the other hand, it is about enabling meaningful reuse, for example as training data in the field of machine learning (ML) or in general in applications involving artificial intelligence (AI).

Another aspect is the conservation of resources, whether financial or for the protection of the environment and the climate. Conducting experiments and computer simulations may require a lot of energy and thus contribute to increased CO_2-emissions into the atmosphere. Science can make a contribution here by avoiding the repeated generation of existing data.

The aim is to generate FAIR data, i.e. research data that is findable, accessible, processable and reusable (FAIR: Findability, Accessibility, Interoperability, and Reusability). In order to achieve this goal, it is necessary to establish a comprehensive infrastructure (hardware, software, etc.). The German Federal Government is supporting this development, for example by financing the National Research Data Infrastructure (NFDI) and the thirty consortia it comprises, each of which is geared to the needs of different scientific and technical disciplines.

With this introduction to research data management (RDM), we would like to provide practical assistance for all stages of RDM to anyone who generates and/or wants to use research data, regardless of the type of data. After reading this best practice guide, you should be able to understand essential RDM structures and handle your data confidently, from creation and analysis to long-term storage.

Research data is a valuable resource and it is worth handling it carefully and sustainably. With this in mind, we hope you enjoy reading this guide.

Hartmut Schlenz, Torsten Bronger, Michael Selzer, Britta Nestler, Leo Riem, and Salome Enahoro

Jülich and Karlsruhe, July 2025

Acknowledgements

We would like to express our sincere gratitude to the German Research Foundation (DFG) for its support within the framework of the National Research Data Infrastructure (NFDI). The funding is specifically for the NFDI4ING consortium, and in particular for the CADEN archetype. Without this funding, this handbook would not have been possible. Special thanks go to H. Schlenz for his support under funding codes D.B.B01953 (until September 2025) and D.B.C02632 (from October 2025).

We would also like to thank the many colleagues who provided us with valuable feedback during the initial presentation of this handbook at the international CoRDI 2025 conference at RWTH Aachen University, which has contributed to the further improvement of this best practice handbook. We welcome further feedback from the community in the future.

Preface to Version 1.1

In this version 1.1 of our best practice handbook on research data management, a number of updates, additions, and corrections have been made, for example on the topic of RDMO (Chapter 4). The chapter on the electronic laboratory notebook JuliaBase (Chapter 5) has been redesigned. The section on Coscine (Chapter 9) has also been expanded and updated.

This handbook has been published on the Zenodo platform and is available in German at `https://doi.org/10.5281/zenodo.18468239` as a PDF file and in English at `https://doi.org/10.5281/zenodo.18468308`.

Hartmut Schlenz

Jülich, January 2026

2. The life cycle of research data

Research data refers to all information that is generated and processed during scientific work. It can be recorded and documented either analogously, i.e. traditionally with paper and pencil, or digitally with the aid of computers. The information is then usually stored as files in a wide variety of formats. In experiments and measurements, for example, research data can be generated as simple text files (ASCII format), as more complex tables in proprietary formats (e.g. MS Excel), as images and graphics (e.g. in JPG format), as audio or video files (e.g. in MP3 or MP4 format) and as source code from software (e.g. Python or C code). A distinction must be made between the actual data and the metadata that describes it. We will discuss this important difference in more detail in the following section. The following figure 2.1 illustrates a typical life cycle of research data. Electronic laboratory notebooks (ELN) play a central role [**ZB**] in the life cycle of research data, from the collection of data and metadata, processing and data analysis, to publication, archiving and data sharing. The use of an ELN [**ZB**, **Briney**, **Corti**, **Putnings**] is an essential part of best practices for modern and comprehensive research data management FDM. We will therefore discuss this topic in detail and repeatedly in the following chapters of this guide.

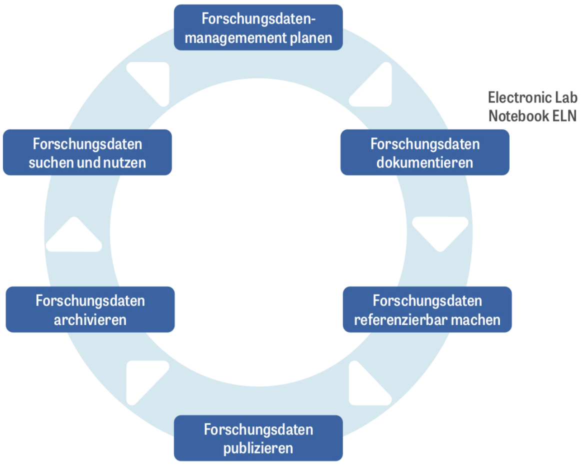


Figure 2.1.: The life cycle of research data [**ZB**].

Summary:

Research data is all analogue and digital information that is created, processed and stored during scientific work.

2.1. Data and metadata

First of all, what exactly is research data? This question is not easy to answer, as scientific research fields can be very diverse and heterogeneous, and so can the data they generate. In practical terms, even a single measured value (a number with an associated dimension; e.g. a pressure of 1 bar) is already research data, even if this example only contains a single data point, which may have been determined with great effort. Large research data sets can be generated in medicine, for example when scanning human brains using Nuclear magnetic resonance spectroscopy NMR. Huge data sets are also possible, which, after their creation, can only be stored and further processed at their place of origin and cannot be transferred due to their size alone. One example is measurement data from particle collisions at the CERN (European Organisation for Nuclear Research) near Geneva, where the structure of matter is being investigated. In general, we distinguish between primary and secondary research data. Primary data is the raw data that is initially generated, while secondary data is primary data that has been further processed. Primary data can be further subdivided. The following table illustrates this classification:

Table 2.1.: Primary and secondary data.

Primary data	Secondary data
Observation data	Processed raw data
Experimental data	Collected data
Simulated data	

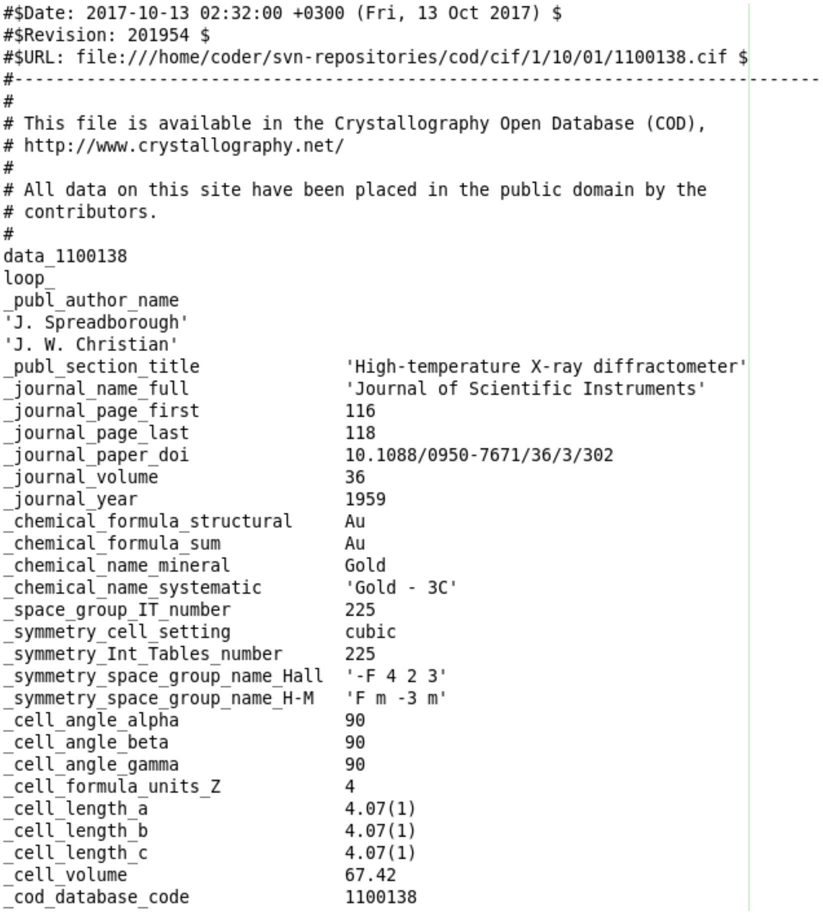
Observation data can be weather data, such as the continuous recording of temperature and air pressure. Experimental data is usually generated in a laboratory and simulated data using special software on a computer.

Metadata form a separate category. We also refer to this as "data about data", i.e. additional information about how primary or secondary data was created. A popular example of metadata is EXIF data. EXIF is a standard format of the Japan Electronic and Information Technology Industries Association for storing metadata in digital images. If photos are processed with image editing software, the image data (EXIF data) associated with an image or photo and recorded by a digital camera can also be displayed and used to search for specific images in an image database (*Find all images taken with a lens focal length of 50 mm, etc.*).

A comparable format in crystal structure research is the CIF format (Crystallographic Information File), which has been established for decades and is used to describe the structural properties of a crystalline material and their experimental determination in detail. This is done using defined conventions that enable the generation, reproduction and retrieval of structural data even many years later. Figure 2.2 shows a simple example of an excerpt from the CIF file for gold (Au), taken from the freely available Crystallography Open Database COD. In this case, simply entering the element name *Au* is sufficient to obtain the desired information and the CIF file from the COD via (`http://www.crystallography.net/cod/`) and use it for further processing. Information is provided about who determined the structural data, when, and in which journal the data was originally published. The CIF file also contains information about the chemical composition, symmetry information, the number of the data set within the COD database, and much more. In short, metadata is a structured, digital form of documentation.

Summary:

Primary data is raw data, and secondary data is data that has been processed further. Metadata is a structured, digital form of documentation of primary and secondary data.



```
#$Date: 2017-10-13 02:32:00 +0300 (Fri, 13 Oct 2017) $
#$Revision: 201954 $
#$URL: file:///home/coder/svn-repositories/cod/cif/1/10/01/1100138.cif $
#------------------------------------------------------------------------------
#
# This file is available in the Crystallography Open Database (COD),
# http://www.crystallography.net/
#
# All data on this site have been placed in the public domain by the
# contributors.
#
data_1100138
loop_
_publ_author_name
'J. Spreadborough'
'J. W. Christian'
_publ_section_title              'High-temperature X-ray diffractometer'
_journal_name_full               'Journal of Scientific Instruments'
_journal_page_first              116
_journal_page_last               118
_journal_paper_doi               10.1088/0950-7671/36/3/302
_journal_volume                  36
_journal_year                    1959
_chemical_formula_structural     Au
_chemical_formula_sum            Au
_chemical_name_mineral           Gold
_chemical_name_systematic        'Gold - 3C'
_space_group_IT_number           225
_symmetry_cell_setting           cubic
_symmetry_Int_Tables_number      225
_symmetry_space_group_name_Hall  '-F 4 2 3'
_symmetry_space_group_name_H-M   'F m -3 m'
_cell_angle_alpha                90
_cell_angle_beta                 90
_cell_angle_gamma                90
_cell_formula_units_Z            4
_cell_length_a                   4.07(1)
_cell_length_b                   4.07(1)
_cell_length_c                   4.07(1)
_cell_volume                     67.42
_cod_database_code               1100138
```

Figure 2.2.: CIF file for gold (Au).

2.2. The creation of FAIR research data

The most important criterion for the creation of FAIR research data is that this data must be findable and accessible. It does not matter whether the data is primary or secondary. Furthermore, data must be interoperable and reusable. We will return to these criteria repeatedly in the following chapters of this guide. Here, we will first briefly outline the criteria that must be met for research data to be FAIR.

2.2.1. Findability

In order for research data to be found, the associated metadata must first be made publicly accessible. Data must also be uniquely identifiable using standard mechanisms, for example by assigning permanent, digital identifiers PID: persistent and unique identifiers. We explain the practical assignment of such PIDs in the chapters on electronic laboratory notebooks (ELNs) and knowledge graphs. Existing naming conventions for the respective discipline should be observed, or, if these do not yet exist, new naming conventions should be created. This will also make it much easier to search databases or repositories using keywords at a later stage. Ideally, existing metadata standards should be taken into account or new standards created. We will also provide detailed and practical information on this in the following chapters.

2.2.2. Accessibility

At the outset, it must be determined which data is to be released or published. It is not always desirable or advisable to make all data from a research project freely available, especially if patents are to be filed for developed processes or materials. This also applies to projects that are financed exclusively with public third-party funds. Once this decision has been made, consideration must be given to how the data is to be made technically accessible. What infrastructure and software should be used for this purpose? A popular and relatively simple option is to upload it to freely accessible repositories such as Zenodo (https://zenodo.org/). Zenodo is an online storage service that can be used primarily for scientific data sets, but also for science-related software, publications, reports, presentations, videos, etc. The service is funded by the European Commission and operated by CERN. There are other repositories that can be used, which we will discuss in more detail in the chapter on permanent data storage. If access to research data and its metadata is to be restricted, the necessary regulations and protection mechanisms must be defined and created.

2.2.3. Processability

Data can be processed more easily if it and its metadata comply with established standards. These can be common file formats such as tables in CSV format. The CSV (comma-separated values) file format describes the structure of a text file that can be used to store or exchange simple structured data. Free and general file formats such as simple text files (ASCII code) should be preferred. The use of proprietary file formats automatically limits the possible uses for a number of potential users. For image files, the file formats JPG, PNG, BMP or TIFF are very popular (depending on the application), and for audio and video files, the formats MP3 and MP4 are commonly used. However, this list is not exhaustive. This guide was written using the LaTeX typesetting system in the freely available TeX format, which can be viewed and edited with any text editor.

2.2.4. Reusability

It may be useful to specify certain licenses for the reuse of research data (e.g. Creative Commons CC BY 4.0 International) so that once published data can be used as freely and globally as possible, if desired. Also make it clear whether use is limited in time or permanent and whether any embargo periods must be observed. This may be the case in industrial collaborations or research networks, especially if a project has not yet been completed. Ideally, you should also document how the data quality has been checked and ensured. We will discuss this point in more detail in the following chapters on data quality and data tracking. This is a core topic of research data management as a whole and should be given careful consideration.

Summary:

FAIR research data must be discoverable, accessible, processable and reusable. This requires well-structured metadata and unique identification of data sets. It must be specified where and how research data, the associated metadata, the necessary documentation and, if applicable, software (source codes) are stored.

3. Legal aspects

As the authors of this guide are not lawyers but scientists, no legally binding information can be provided here. The following sections contain a summary of freely available legal information, which should be regarded as non-binding guidance. In case of doubt, the legal department and/or data protection officer of the respective research institution should be consulted before critical systems are put into operation or the processing of research data is not secure and clearly compliant with the law. The following sections explain in more detail in which cases particular attention must be paid to the legally compliant handling of research data.

3.1. Copyright when using third-party data

Usage and copyrights as well as rights to third-party intellectual property and patent rights must be checked before using third-party data. The legal obligations apply across the board to the recording, documentation, storage, archiving and reuse of data used [**DFG**]. Information itself is not protected by copyright. The same applies to theses and academic opinions, so that they cannot be subject to copyright monopoly and free academic discussion is guaranteed [**Lauber**]. However, research data may well be protected by copyright or ancillary copyright laws. According to §2 Abs. 1 UrhG, copyright protection may be considered if a "personal intellectual creation" exists (§2 Abs. 2 UrhG). For this, a work must exhibit particular individuality [**Lauber**]. Purely manual, routine work is therefore excluded, as are professional practices. The scientist must have had explicit creative freedom. For detailed information, please refer to [**Lauber**, **Baumann**].

3.1.1. Text and data mining

The legislator defines text and data mining as the *automated analysis of individual or multiple digital or digitised works in order to obtain information, in particular about patterns, trends and correlations* (§ 44b Abs. 1 UrhG) [**Brehm**]. The law thus applies in principle to all objects protected by copyright (texts, graphics, images, audio recordings, music, data, databases, etc.). Of course, objects that are not protected by copyright can also be the subject of text and data mining. Copyright permission is generally not required for this. Extensive databases that were very costly to create are, however, protected independently of the copyright protection of the content itself. This description is limited to scientific text publications [**Brehm**].

Extensive information can be found in the guidelines by Elke Brehm [**Brehm**]. There she describes in detail the conditions under which text and data mining for scientific purposes may be carried out in scientific publications on the basis of so-called restrictions and/or contracts, and what risks may be involved.

3.2. Processing of personal data

Data protection law applies when personal data is processed. According to the General Data Protection Regulation (Art. 4 Nr.1 DSGVO), this is any information relating to an identified or identifiable person [**Lauber**]. Since the introduction of the DSGVO, there has been ongoing discussion about what is and is not permitted in photography and in general when handling image data. A personal reference can also be established in photos with faces that have been made unrecognisable if identification is possible based on the background, clothing and posture of the persons depicted, as well as accompanying information about the time and place the photo was taken [**Lauber**]. In the case of medical research data, the metadata may contain information about the researchers involved in addition to patient or test subject data. The DSGVO also applies here, as this is also personal data. In this context, when using electronic laboratory notebooks (ELNs), care must be taken to ensure that the personal data of researchers stored in public institutions is not used to monitor performance. Ideally, a works agreement between the works council and the employer should specify exactly how such data is to be handled [**ZB**, **Corti**, **Johannes**]. According to § 75 Abs. 3 Nr. 17 BPersVG, the introduction and use of technical equipment intended to monitor the behaviour or performance of employees is subject to co-determination by the works council [**Bremecker**]. Data processing systems are also generally considered to be technical equipment within the meaning of co-determination. The technical equipment does not have to be intended to monitor the behaviour and performance of employees. According to the case law of the Federal Administrative Court in Leipzig (BVerwG), this condition is already fulfilled if the technical equipment is objectively suitable for monitoring the behaviour or performance of employees. It is therefore sufficient if the equipment is initially intended for other purposes and there is no intention to monitor. The technical possibilities (hardware and software) for potentially possible monitoring or control alone make co-determination necessary. According to the legal opinion of Paul C. Johannes [**Johannes**], the freedom of scientific research and its participants should be given particular priority. According to Article 5 III of the German Basic Law (GG), scientific freedom is considered particularly worthy of protection. In the next chapter, we will discuss the technical and digital possibilities of using an ELN in detail, and it will quickly become clear that performance monitoring is indeed possible with an ELN.

3.3. Legal framework for the transfer of data

The conditions under which research data is released for reuse should be as unrestrictive and transparent as possible [**Lauber**]. Since copyright cannot be completely waived under German law, which also applies to research data, licence agreements must be used. Comprehensive, royalty-free rights of use are granted to users through so-called free licences. Repositories such as Zenodo use model contracts for this purpose. Creative Commons licences are also widely used. The European Commission recommends the licence types CC-BY and CCO [**EU**]. The following table summarises the most important information about these two licence types. Further possible licences are described in more detail in [**Lauber**]. The Open Data Commons (ODC: `https://opendatacommons.org/`) can also be considered as a licence model for research data.

Table 3.1.: Creative Commons licences.

Licence	Permitted:	Condition:
CC BY	Reproduction, distribution, Creation of adaptations and their reproduction and distribution for commercial and non-commercial purposes	Attribution: Name of the creator; mention of the licence type and reference to licence text via URI/hyperlink; URI/hyperlink to the licensed material; copyright notice, reference to disclaimer; Notice if licensed material has been modified.
CCO	partial waiver of copyright; as not possible under copyright law, wide-ranging granting of rights of use.	generally no attribution required.

Summary:

Legal requirements must be observed in all areas of research data management, especially when using third-party data, processing personal data, and sharing research data. If in doubt, always consult the legal department or data protection officer of the respective institution.

4. The data management plan

The data management plan (DMP) is primarily a formal document that describes how research data is handled throughout its entire lifecycle (see Chapter 2), from project preparation, during the project, and beyond, for example as a planning basis for further projects. The DMP should be a natural part of the project planning and be updated during the project. Most third-party funders now expect a DMP as part of a project application (see Chapter 1), and there are initial verified cases in which a project application was rejected due to a missing DMP. An overview of the topic of DMPs can be found on the website of (`https://forschungsdaten.info/`), among others. However, it is still up to each individual to decide how much effort they want to put into creating and maintaining a DMP. A very comprehensive but also quite complex method is the use of RDMO (Research Data Management Organiser; `https://rdmorganiser.github.io/`). RDMO is free software that can be used for planning, implementing and managing research data management. Cooperation networks are also important for effective use. All RDMO users and institutions that have their own instance can seek advice and support from the RDMO community. This cooperation network enables RDMO to take into account requirements and feedback from the specialist sciences, as well as ensuring exchange and coordination with infrastructure initiatives for data management.

Figure 4.1.: RDMO-Software.

A demo instance is available at `https://rdmo.aip.de/` for initial testing. However, this software must be installed on a server and the accounts are managed centrally. Ideally, the operating institute has its own RDMO instance with administrator rights. The costs and effort involved in operating and maintaining RDMO are unavoidable. A quick start guide for using the RDMO software is available as a PDF file at the following link: `https://rdmorganiser.github.io/docs/Schnellstartanleitung_v2024.pdf`

However, experience has shown that even when the necessary infrastructure is provided, relatively few scientists take advantage of this option. It is much easier to create a DMP as a simple text file that can be modified gradually and easily. For most third-party funding providers, this simple solution is sufficient and is accepted for project applications. We will therefore briefly outline the information required for writing a simple DMP below and illustrate it with examples. This description is based on the recommendations of the DFG (`https://www.dfg.de/antragstellung/forschungsdaten`). The following questions should be answered as precisely as possible to ensure that a DMP is comprehensible. At the same time, these questions help with project planning and make it easier to structure projects, because wherever data is generated and processed in a project, corresponding experiments or simulations must also be carried out. When planning data flows, a sensible timeline for the entire project automatically becomes apparent.

4.1. Data description

How is new data generated in your project? Is existing data reused? Which data types, in the sense of data formats (e.g. image data, text data or measurement data), are generated in your project and how are they further processed? To what extent do these occur or what data volume is to be expected?

Example:

According to research in common data repositories, no current or suitable research data is available for reuse in this project. The data generated in the project will enable further insights in the field of XY. The data sets generated will be created by the project team using various analysis methods, primarily REM, XRD and TEM. The data will mainly be textual, tabular and image data. Where possible, this will be stored in open formats (textual data txt, rtf, pdf; tabular data csv; image data tiff). During the project period, analyses and evaluations will be carried out using the freely available programming language Python and its established open source libraries. The expected data volume will not exceed 100 GB.

4.2. Documentation and data quality

What approaches will be used to describe the data in a comprehensible manner (e.g. use of existing metadata or documentation standards or ontologies)? What measures will be taken to ensure high data quality? Are quality controls planned and, if so, in what form? What digital methods and tools (e.g. software) are required to use the data?

Example:

The research data and scripts generated are published in the JülichDATA repository. In accordance with the FAIR principles, the data in the repository is described by metadata based on the DataCite schema (including abstract, free keywords and DDC classification). The electronic laboratory notebook eLabFTW is used for data documentation. The repository adds a persistent identifier (DOI) to the metadata, which makes the data set uniquely referenceable. Files and directories are named according to a uniform scheme; for example, dates are formatted according to ISO 2014: [YYYY]-[MM]-[DD]. The scheme is defined at the start of the project together with all project participants. The tabular data (CSV) is documented using one or more tabular data packages. This specification documents the data, the individual columns (variables), their permitted data types and value ranges, and the relationships between columns (even across individual files). The formalised data description and documentation enables tool-based quality control to be carried out on a regular basis (e.g. all values in specific columns are within the permitted value ranges). The data is validated and its quality assured by regularly measuring standard samples with known properties at the various characterisation facilities. The data and scripts can be used with open-source standard tools. There are no costs for specialised software to read/edit/execute the files.

4.3. Storage and technical security during the course of the project

How will the data be stored and backed up during the project? How will the security of sensitive data be ensured during the project (access and usage management)?

Example:

The expected maximum data volume of 100 GB will be provided by a scientific storage area (hereinafter referred to as the *project network drive*) at the Jülich Research Centre. During the project period, data and metadata will be stored on the project network drive. The drive will be integrated as a network drive by all project staff via their respective operating systems. The project network drive is subject to an automated, regular, file-based backup routine by the computer centre. The data is backed up regularly and automatically on a file server. In the event that data and scripts are created locally on the project group's workstations, employees synchronise them once a week with the project network drive to prevent data loss. Open source tools established for the respective operating system are used for this purpose. Since no sensitive data is collected, there is no separate access and usage management. Access to the project network drive is managed centrally by the Research Centre Jülich's computer centre (only members of the project group have access). Access to the data by third parties is not required during the project period.

4.4. Legal obligations and framework conditions

What legal peculiarities exist in connection with the handling of research data in your project? Are any effects or restrictions to be expected with regard to subsequent publication or accessibility? How are usage and copyright aspects as well as ownership issues taken into account? Are there any important scientific codes or professional standards that should be taken into account?

Example:

There are no special legal considerations with regard to the data. The reuse of software from other authors is cited in accordance with good scientific practice.
Note: At this point, further legal safeguards may be used, as described in Chapter 3.

4.5. Data exchange and permanent accessibility of data

Which data are particularly suitable for reuse in other contexts? What criteria are used to select research data for reuse by others? Do you plan to archive your data in a suitable infrastructure? If so, how and where? Are there any retention periods? When will the research data be available for use by third parties?

Example:

The collected data and scripts are suitable for reuse by third parties. Therefore, metadata and some data are published in the JülichDATA repository, which is owned by the research centre. The publication includes all raw data and scripts generated, as well as final versions of text data and tables. Documentation is also included with the publication. Interim results of processing and analysis steps that can be generated from the provided data and scripts are not part of the publication. The publication follows the recommendations of the Open Access Policy and Research Data Policy of the research centre. The use of the electronic laboratory notebook eLabFTW and the repository JülichDATA ensures that several of the points addressed in the FAIR principles are met. For example, metadata is indexed in comprehensive reference systems and search engines (e.g. BASE, DataCite Search, OpenAIRE) via standardised interfaces (such as OAI-PMH). This increases the visibility of the research results. The metadata created is checked by the repository's editorial team. Furthermore, a DOI is assigned to data sets that are to be published. In accordance with the guidelines for ensuring good scientific practice, the data will be made publicly available by the repository for at least ten years, i.e. without access restrictions. Separate archiving, independent of publication, is not planned. A blocking period is not required. Publication will take place as soon as possible, but no later than within the last three months of the project period.

4.6. Responsibility and resources

Who is responsible for the adequate handling of research data (description of roles and responsibilities within the project)? What resources (costs, time or other) are required to implement adequate handling of research data in the project? Who is responsible for curating the data after the end of the project?

Example:

Mr/Ms XY is primarily responsible for handling the research data obtained in the project. This person will ensure that the DMP is complied with and updated. The data and scripts will be continuously documented and processed during the project period and finalised in the last three months. At the end of the project, all data intended for publication will be published at the specified location. No further curation of the data will take place beyond the project duration.

Summary:

Writing and continuously maintaining a DMP may initially seem like unnecessary extra work, or the advantages may not be immediately apparent. In fact, an efficiently written DMP (see above) is a useful aid in planning and implementing a project. A DMP also makes it much easier to generate FAIR research data.

5. Data collection, data storage and documentation

5.1. Electronic laboratory notebooks

Paper lab notebooks are increasingly being replaced by electronic lab notebooks in laboratories. This transition is not just about replacing paper with a digital application. Equally important is the ability to integrate the electronic form of the laboratory notebook into the overall system of digital research data management (RDM) [**ZB**]. In the life cycle of research data, the ELN plays an important role in the documentation phase. When considering the purchase and use of an ELN, a few fundamental questions should be asked at the outset:
What is the overall workflow for research data in the life cycle? Which IT applications or tools should be used at at which stage? What function should an ELN fulfill in this overall context?
In addition to the design of institutional research data management, an ELN can also make a significant contribution to good scientific practice, as its use makes research processes and research results more traceable. Compared to the traditional paper form, the use of an ELN offers a number of decisive advantages that make work in the laboratory or with computer simulations significantly more efficient and also increases data security [**ZB**]:

- Direct integration/linking of existing digital data (e.g. measurement results, image, video, audio files, texts, tables).
- No loss of information due to illegible handwriting.
- Search and filter functions.
- Functions for collaborative work (rights and role management).
- Team and laboratory organisation.
- Creation and use of templates (templates, e.g. for repetitive processes).
- Embedding in a networked digital research environment, interfaces, import and export functions, connection to repositories, long-term archiving, etc.).
- Faster and easier publication of data sets.

5.2. Popular electronic laboratory notebooks

The following two tables list a number of currently popular free (open source) and commercial electronic lab notebooks. These tables are not exhaustive. New ELNs are constantly being developed, often tailored to the needs of specific disciplines. Further information on existing ELNs can be found in the ELN Finder of the University and State Library of the TU Darmstadt (https://eln-finder.ulb.tu-darmstadt.de/home), on Wikipedia (https://en.wikipedia.org) or via a variety of other sources on the Internet. There are currently new developments in which large language models (LLMs) are being designed and programmed for use with electronic laboratory notebooks with AI. However, as these developments are still in their infancy and cannot yet be considered mature, we would like to refer you to the relevant literature [**Jalali**] at this point. If LLMs become more widespread in electronic laboratory notebooks in the future, which is not entirely uncritical and unproblematic in terms of data security in closed ELN systems, this could open up new possibilities for accelerated data collection and the retrieval of existing data. The transfer of data directly to an AI for data analysis and further use could also be simplified and accelerated in this way. Initial approaches to this already exist, including in the ELN Kadi4Mat described below, although LLM is not yet used there.

Table 5.1.: Free electronic laboratory notebooks (open source).

Name	web address
Juliabase	`https://www.juliabase.org/`
eLabFTW	`https://www.elabftw.net/`
Kadi4Mat	`https://kadi.iam.kit.edu/`
Chemotion	`https://chemotion.net/`
SampleDB	`https://github.com/sciapp/sampledb`
Pasta ELN	`https://github.com/PASTA-ELN/desktop`
Herbie	`https://www.hereon.de/herbie`
elog	`https://elog.psi.ch/`
Indigo ELN	`https://github.com/epam/Indigo-ELN-v.-2.0`
openBIS	`https://openbis.ch/`
LabCloud	`https://www.labcloudinc.com/`
OSF	`https://osf.io/`

Table 5.2.: Commercial electronic laboratory notebooks.

Name	Web address
Labfolder	`https://labfolder.com/de/`
eLabNext	`https://www.elabnext.com/`
RSpace	`https://www.researchspace.com/`
LabArchives	`https://www.labarchives.com/`
CERF 5.0	`https://cerf-notebook.com/about-cerf-5-0/`
Uncountable	`https://www.uncountable.com/`
Benchling	`https://www.benchling.com/`
Labstep	`https://www.labstep.com/`
SciNote	`https://www.scinote.net/`
Findings	`https://findingsapp.com/`
Hivebench	`https://scolary.com/tools/hivebench`
Find Molecule	`https://findmolecule.com/eln/`
SciCord ELN	`https://scicord.com/`
Labguru	`https://www.labguru.com/`
BrightLab	`https://www.researchstash.com/resource/brightlab/`
Docollab	`https://www.docollab.com/`
LabTwin	`https://www.labtwin.com/de/`
Mbook	`https://mestrelab.com/software/mbook/`

We expressly accept no responsibility for the accuracy and security of the web links (URLs) listed in the tables or for the content of the linked websites.

5.3. Electronic lab books in practice

5.3.1. JuliaBase

Does your scientific institute or working group generate a large number of samples, and does your team need a tool to keep track of them? JuliaBase was developed at just such an institute. It is a database solution for samples, their processing and characterisation, with the following features:

- Completely open source with a free software licence
- Browser-based interface that also works on mobile devices
- High flexibility for adaptation to existing production and measurement equipment and workflows
- Finely graded access control
- The ability to manage more than one department separately in a single database
- Connects to your LDAP server for user management
- Sample splits are tracked cleanly so that data from parent and daughter pieces is always visible

- Support for preliminary evaluation of raw data and data visualisation
- Automatic notification of changes to samples
- Grouping by sample series, topics and tags
- Complex searches made easy, e.g. "find all samples with infrared measurements that were deposited together with a sample on a glass substrate with a conductivity greater than 10–6 S/cm; oh, and only from this year and manufactured by John"
- Export to spreadsheets (via CSV files)
- Automatic tabular lab books
- REST API for direct connection of a measurement setup to the database
- Fully translatable; the core is currently available in English and German
- Layout can be adapted to corporate identity
- Mature code base since 2008
- Compliance with modern web and security standards

JuliaBase takes the approach that the database should adapt to existing workflows and not vice versa. However, the flexibility of JuliaBase comes at a price: Python code must be created for each type of process that you want to integrate. Typically, this only amounts to up to 100 lines of code for each process, and JuliaBase even contains code for typical processing and measurement setups that can be used as a starting point. Nevertheless, this work is necessary.

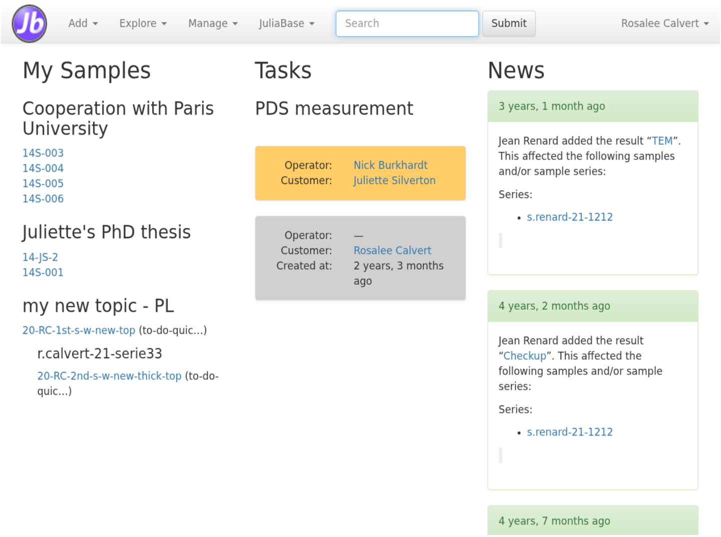
A tour of JuliaBase

A demo of JuliaBase is installed at `https://demo.juliabase.org` so you can play around with it a bit. You can add samples, processes, tasks, etc., view sample data sheets or a lab book, and much more. You can log in with different accounts to try out different permissions (roles).
The demo site is the JuliaBase installation of the fictional *Institute of Nifty New Materials (INM)*. It is a very small institute with only six employees. All accounts have the password "*12345*".

The demo accounts:

Sean Renard (s.renard) is the chief scientist and director of this institute. Accordingly, his JuliaBase account allows him to see all samples, but he also has other privileges.
Nick Burkhardt (n.burkhardt) has been an operator at INM for a very long time. He is responsible for the PDS (photothermal deflection spectroscopy) setup, a measuring device. He performs measurements for researchers. He would never allow another person to use his PDS.
Hank Griffin (h.griffin) is also an operator. He is responsible for the solar simulator, another measuring device. He performs measurements for researchers, but after appropriate instruction from him, other people can also use the device.
Eddie Monroe (e.monroe) is the third operator. He is a technician who operates a deposition system – in his case, cluster tool deposition. Such processes are used to produce samples. Here, too, other members of the institute can use this system after receiving appropriate training.
Rosalee Calvert (r.calvert) is a permanent scientist and prepares samples herself in the 5-chamber

coating system. She is currently the only one who uses this system. She then measures the samples in the solar simulator. Her current project is a collaboration with the University of Paris. Juliette Silverton (j.silverton) is a doctoral student who has a lot to do. Therefore, she cannot carry out the sample preparation and measurements herself, but has to leave this to others. She makes intensive use of the JuliaBase order list function to commission this work. Below, we take a closer look at the typical workflows of these individuals. **Rosalee: Typical normal user** Log in as r.calvert, i.e. as the typical normal user (Figure 5.1). In the main menu, you can see Rosalee's *My Samples*. This list does not normally contain all of a user's samples, but only those that are currently of interest to him/her. Nevertheless, this list can become quite long and is therefore organised by topic and sample series. You can click on the icons in front of the series or topic names to expand and collapse sections that you want to hide. A sample usually belongs to exactly one topic. This helps to organise the samples. (In Rosalee's list, all samples belong to the topic of collaboration with the University of Paris.) On the one hand, topics can be given meaningful names that clearly indicate the purpose of the samples. But more importantly, topics determine who can see the sample. The most important guideline in JuliaBase is: you can only see samples from your topics. People can be in any number of topics at the same time, but a sample is in exactly one topic. It can change during its lifetime. Senior team members may have permission to see *all* samples. Sean Renard is one such person.



My Samples

Cooperation with Paris University

14S-003
14S-004
14S-005
14S-006

Juliette's PhD thesis

14-JS-2
14S-001

my new topic - PL

20-RC-1st-s-w-new-top (to-do-quic...)

r.calvert-21-serie33

20-RC-2nd-s-w-new-thick-top (to-do-quic...)

Tasks

PDS measurement

Operator: Nick Burkhardt
Customer: Juliette Silverton

Operator: —
Customer: Rosalee Calvert
Created at: 2 years, 3 months ago

News

3 years, 1 month ago

Jean Renard added the result "TEM". This affected the following samples and/or sample series:

Series:

- s.renard-21-1212

4 years, 2 months ago

Jean Renard added the result "Checkup". This affected the following samples and/or sample series:

Series:

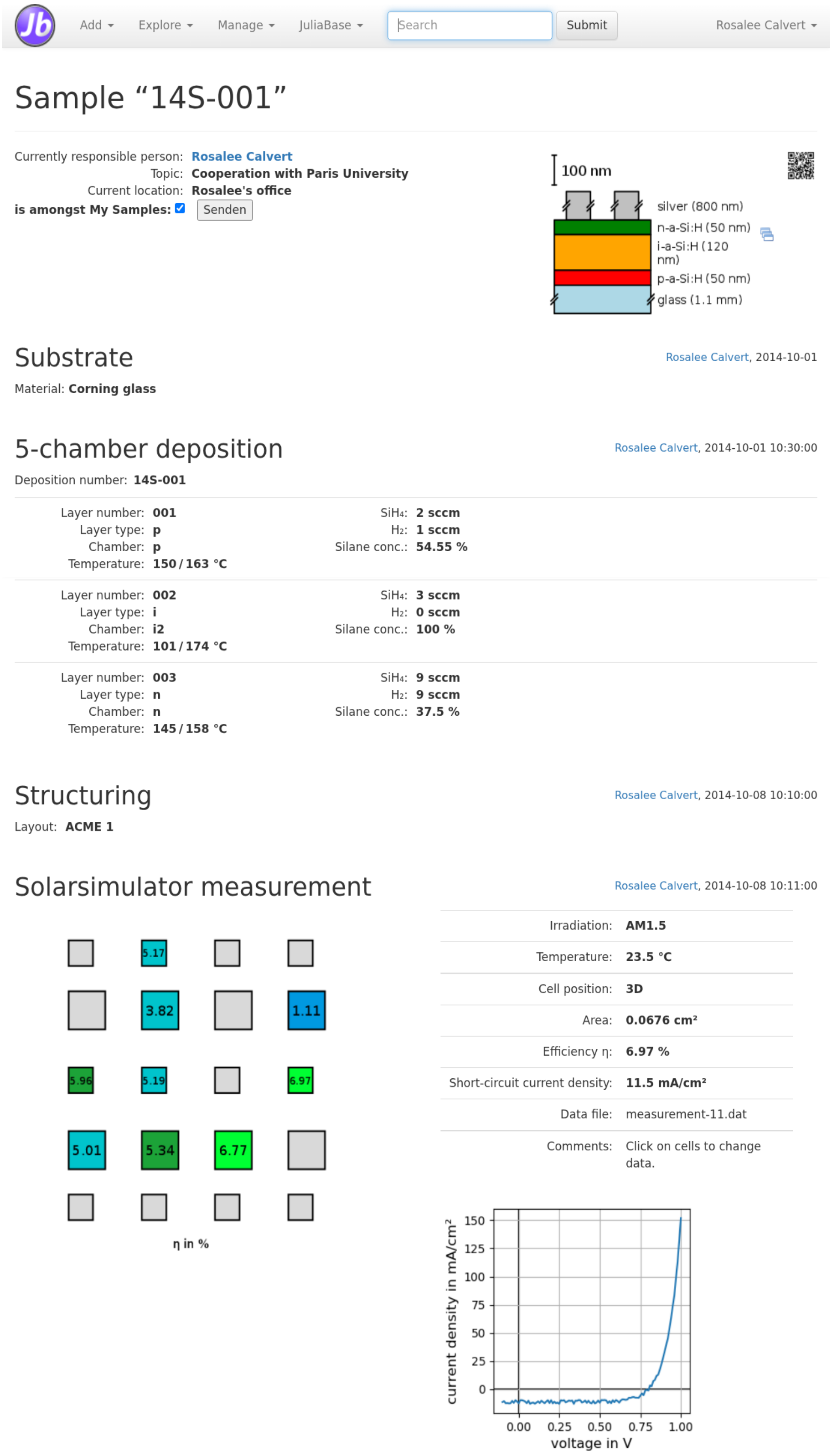
- s.renard-21-1212

4 years, 7 months ago

Figure 5.1.: Daily work.

Sample data sheet:

View an example of a sample by clicking on *14S-001* (Figure 5.2). You will see the sample's *data sheet*. At the top, you will find some general information about the person currently responsible and the topic. Then you will see a list of all the work that has been done with this sample, in chronological order. It starts with the substrate, continues with the deposition of the silicon layers, and ends with a measurement in the solar simulator. Each of these steps is referred to as a "process" in JuliaBase. Even the starting substrate is a process, albeit not in the literal sense. Each process has an operator and a timestamp. You can collapse processes by clicking on the heading. The main work involved in adapting JuliaBase to a new institute is creating all the processes that the institute needs in Python code. The solar simulator measurement at the bottom of the sample data sheet is a good example of why the effort is worthwhile: just click on the coloured squares and you will immediately see how the data and display change. Most other ELNs lack such features. This high degree of adaptability and



Sample “14S-001”

Currently responsible person: **Rosalee Calvert**
Topic: **Cooperation with Paris University**
Current location: **Rosalee's office**
is amongst My Samples: ☑ Senden

Substrate

Rosalee Calvert, 2014-10-01

Material: **Corning glass**

5-chamber deposition

Rosalee Calvert, 2014-10-01 10:30:00

Deposition number: **14S-001**

Layer number:	**001**	SiH_4:	**2 sccm**
Layer type:	**p**	H_2:	**1 sccm**
Chamber:	**p**	Silane conc.:	**54.55 %**
Temperature:	**150 / 163 °C**		
Layer number:	**002**	SiH_4:	**3 sccm**
Layer type:	**i**	H_2:	**0 sccm**
Chamber:	**i2**	Silane conc.:	**100 %**
Temperature:	**101 / 174 °C**		
Layer number:	**003**	SiH_4:	**9 sccm**
Layer type:	**n**	H_2:	**9 sccm**
Chamber:	**n**	Silane conc.:	**37.5 %**
Temperature:	**145 / 158 °C**		

Structuring

Rosalee Calvert, 2014-10-08 10:10:00

Layout: **ACME 1**

Solarsimulator measurement

Rosalee Calvert, 2014-10-08 10:11:00

Irradiation:	**AM1.5**
Temperature:	**23.5 °C**
Cell position:	**3D**
Area:	**0.0676 cm^2**
Efficiency η:	**6.97 %**
Short-circuit current density:	**11.5 mA/cm^2**
Data file:	measurement-11.dat
Comments:	Click on cells to change data.

Figure 5.2.: Sample data sheet.

flexibility is JuliaBase's main strength. Let's scroll back to the beginning. You will see a schematic cross-section of the sample. This is also an extension for the INM (which can be reused by other institutes). If you click on the cross-section, you will receive it as a PDF. This also applies to all plots in JuliaBase.[1ex] **Editing samples:**

You can edit a sample by clicking on the pencil icon next to the sample name. Editing a sample only affects the data at the top of the data sheet. In particular, no processes are affected.

Adding a process:

In addition, there is a cog icon at the top of the sample data sheet that is used to attach a new process to the sample. When you click on it, you will be asked what type of process you want to add or whether you want to split a sample into parts.[1ex] It is generally not a good idea to delete items from a database. However, due to high demand from users, JuliaBase offers the option of removing samples and processes at a later date. The rules are very strict, however: you can only delete processes that you can edit and that are less than an hour old.
Sample splitting:

To split a sample into pieces, click on the gear icon and select *Sample splitting*. You can then enter the new names for the sample pieces. When you view the data sheet for a sample, you will also see all the processes of its "ancestors".
The generic process:

Result processes, often simply called results, are a practical ad hoc way to attach generic data to a sample's data sheet. If you want to add a measurement result for which no specific process has been programmed yet, or if you want to add a diagram, image or comment, then create a result. This is the Swiss Army knife method when nothing else fits.

Advanced search:

Rosalee wants to see her best samples (Figure 5.3). To do this, select *Search for Things – Advanced Search* from the main menu. Now carry out the following steps and click on "Submit" after each step:

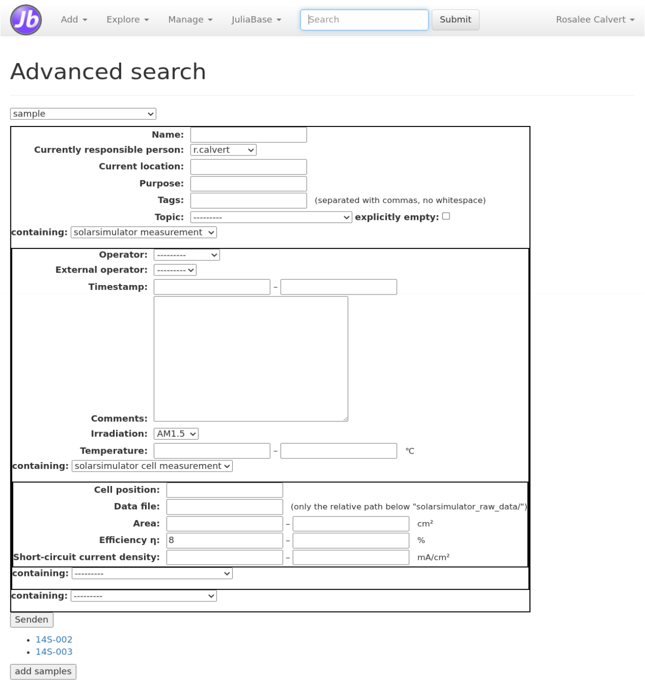
1. Select *Sample* from the drop-down menu.
2. Enter *calvert* in "*currently responsible person*" and select "*solar simulator measurement*" from the drop-down menu.
3. Select *AM1.5* in "*Irradiation*" and select "*Solar simulator cell measurement*" in the inner drop-down menu.
4. Enter the value 8 for efficiency η.

You will get the result shown in the image above this text: Two of your samples meet the criteria, namely *14S-002* and *14S-003*. This means that at least one solar simulator measurement was made on both samples under AM 1.5 irradiation, with at least one cell having an efficiency of more than 8 %

. You can bookmark the results of complex search queries and call them up again as often as you like. Each time, you will receive new results for your old search criteria.

Data export:

Rosalee needs the data in her spreadsheet programme. So click *Send* again. You can then select the



Advanced search

sample

Name:
Currently responsible person: r.calvert
Current location:
Purpose:
Tags: (separated with commas, no whitespace)
Topic: --------- explicitly empty: □
containing: solarsimulator measurement

Operator: ---------
External operator: ---------
Timestamp: –
Comments:
Irradiation: AM1.5
Temperature: – ℃
containing: solarsimulator cell measurement

Cell position:
Data file: (only the relative path below "solarsimulator_raw_data/")
Area: – cm²
Efficiency η: 8 – %
Short-circuit current density: – mA/cm²
containing: ---------

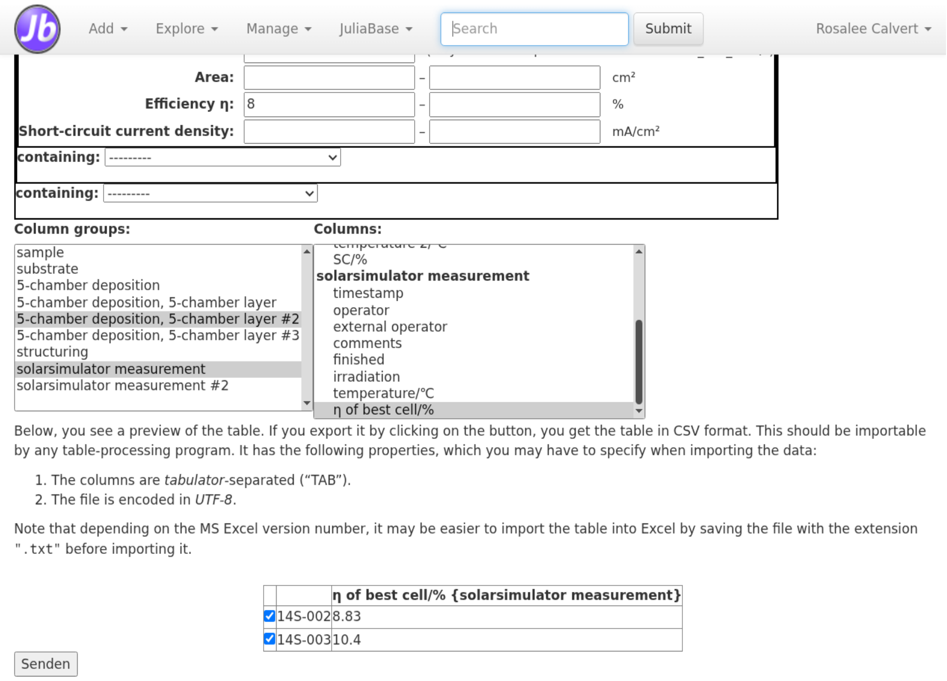
containing: ---------

Senden

- 14S-002
- 14S-003

add samples

Figure 5.3.: Advanced search.



Area: – cm²
Efficiency η: 8 – %
Short-circuit current density: – mA/cm²
containing: ---------

containing: ---------

Column groups:
sample
substrate
5-chamber deposition
5-chamber deposition, 5-chamber layer
5-chamber deposition, 5-chamber layer #2
5-chamber deposition, 5-chamber layer #3
structuring
solarsimulator measurement
solarsimulator measurement #2

Columns:
SC/%
solarsimulator measurement
timestamp
operator
external operator
comments
finished
irradiation
temperature/℃
η of best cell/%

Below, you see a preview of the table. If you export it by clicking on the button, you get the table in CSV format. This should be importable by any table-processing program. It has the following properties, which you may have to specify when importing the data:

1. The columns are *tabulator*-separated (“TAB”).
2. The file is encoded in *UTF-8*.

Note that depending on the MS Excel version number, it may be easier to import the table into Excel by saving the file with the extension `".txt"` before importing it.

	η of best cell/% {solarsimulator measurement}
☑ 14S-002	8.83
☑ 14S-003	10.4

Senden

Figure 5.4.: Data export.

processes to be included in the export. Select the second layer of the 5-chamber deposition and the first solar simulator measurement. Click *Submit*. Now you can select the fields from these processes that you want to include in the export. Select "*SiH4/sccm*" (this is the silane gas flow) from the layer and "*η of the best cell/%*" from the solar simulator measurement. Click "*Send*". It should then look like Figure 5.4. The table contains all the data that will be exported. Click "*Send*" one last time, and you can download this table as a CSV file, which you can open with your favourite spreadsheet programme.

Adding samples:

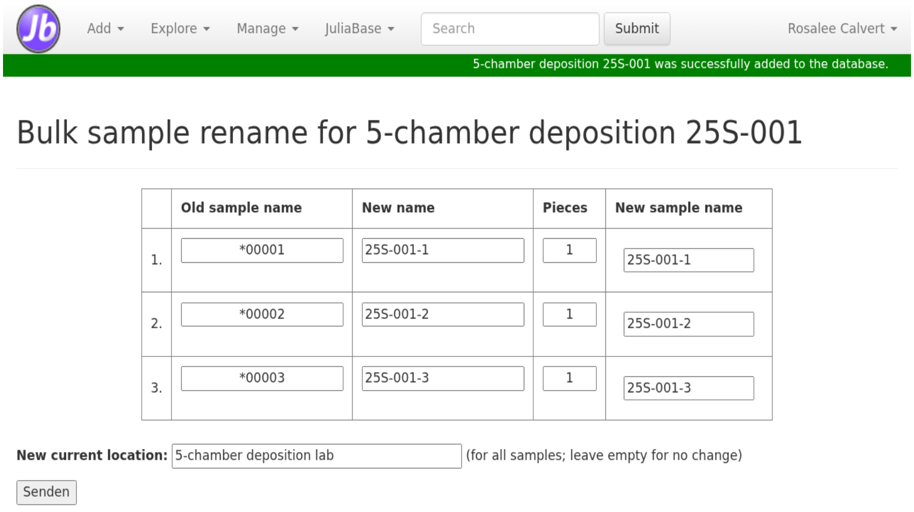
In the main menu, you can click on "*Add things – Samples*" to add samples. Note that this page is very institute-specific. Your institute probably does not use something like substrates, and certainly does not have something like a cleaning number. In any case, you must enter the number of samples and their current location. Add a few samples, but do not rename them yet. Fresh samples have a temporary name in JuliaBase. It looks like **00034*, i.e. an asterisk followed by a five-digit number. Never use these names on sample labels or in publications. They should be replaced with a real name as soon as possible. Rosalee's samples are named after the first silicon deposition.

Tabular lab books:

Open the lab book for the five-chamber deposition in the main menu. You will see six depositions from October 2014. Select one of them. JuliaBase displays a page containing only the details of this deposition. Click on the *gear icon* at the top of the page to create a new run using the current one as a template.

Add new deposition process:

Rosalee uses old separation runs as templates because the runs do not vary greatly from one another.



5-chamber deposition 25S-001 was successfully added to the database.

Bulk sample rename for 5-chamber deposition 25S-001

	Old sample name	New name	Pieces	New sample name
1.	*00001	25S-001-1	1	25S-001-1
2.	*00002	25S-001-2	1	25S-001-2
3.	*00003	25S-001-3	1	25S-001-3

New current location: 5-chamber deposition lab (for all samples; leave empty for no change)

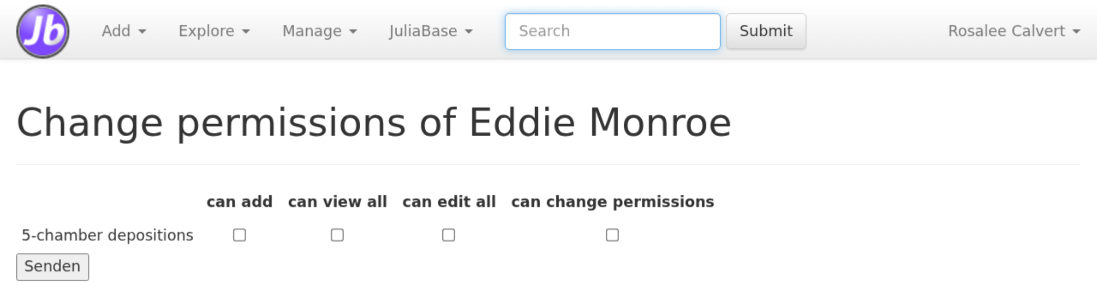
Senden

Figure 5.5.: Sample renaming after a new deposition process.

This allows her to add new runs without much fuss. On the page for the new separation, she only has to select the samples for separation (these are the recently added samples with these *... names), adjust a few things that were different in this run, and click "Submit". Now, it is common practice at the *Institute for Nifty New Materials* (INM) to give the sample the same name as the separation run. (JuliaBase also allows other naming policies.) Therefore, immediately after adding the deposition, you will be redirected to a page where you can review and change the new sample names. JuliaBase suggests the name of the deposition for all samples (in the case of the screenshot, for three samples). However, since the names must be unique, Rosalee adds . . . -a, . . . -b, and . . . -c to the names (see screenshot, second column). Click "Submit," and you're done! The newly saved samples appear with their correct names under "My Samples" on the main menu page. Of course, your institution may have a different workflow that does not require renaming. In this case, you can simply omit the renaming page in your own JuliaBase customisation.

Change permissions for processes:

As already mentioned, Rosalee is responsible for experiments on the 5-chamber deposition. But let's assume that Eddie also wants to do such depositions and is given an introduction? Then he should also have permission to add such deposits to JuliaBase (Figure 5.6). Rosalee goes to "*Miscellaneous – Permissions for processes*" in the main menu, selects Eddie from the drop-down menu and clicks on "*Submit*". She ticks the first two checkboxes and clicks "Submit" again. Eddie now has the following additional permissions:



Change permissions of Eddie Monroe

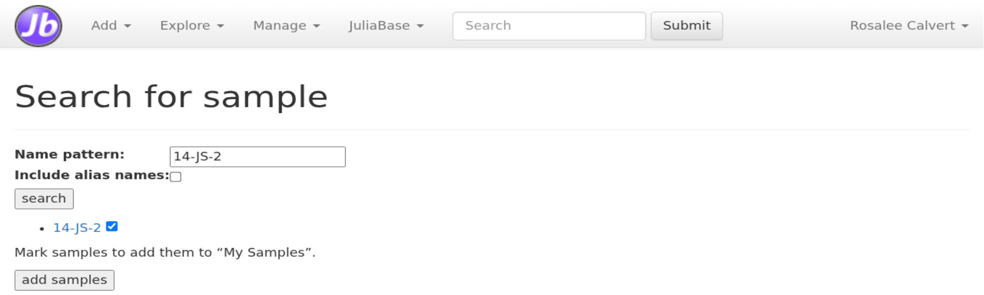
	can add	can view all	can edit all	can change permissions
5-chamber depositions	□	□	□	□

Figure 5.6.: Changing permissions for processes.

- He can add new 5-chamber separations.
- He can edit his own 5-chamber separations (those for which he was the operator).
- He can view all 5-chamber separations. This means, in particular, that he can view the lab book for this system.

Sample owner:

Sometimes samples need to change owners, e.g. when someone leaves the institute. Let's assume



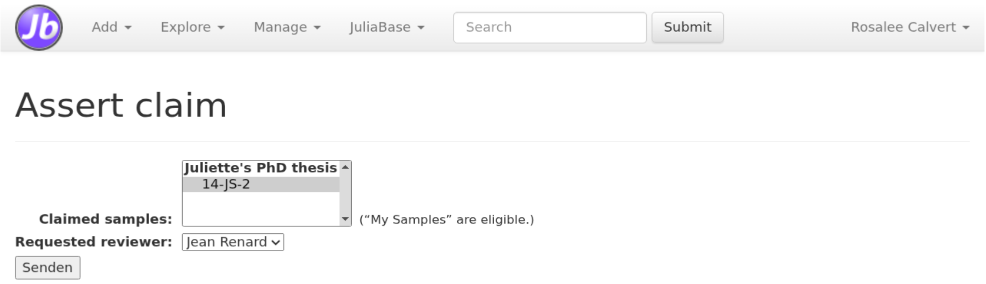
Search for sample

Name pattern: 14-JS-2
Include alias names: □

- 14-JS-2 ☑

Mark samples to add them to "My Samples".

Figure 5.7.: Sample search.



Assert claim

Claimed samples: Juliette's PhD thesis
14-JS-2 ("My Samples" are eligible.)

Requested reviewer: Jean Renard

Senden

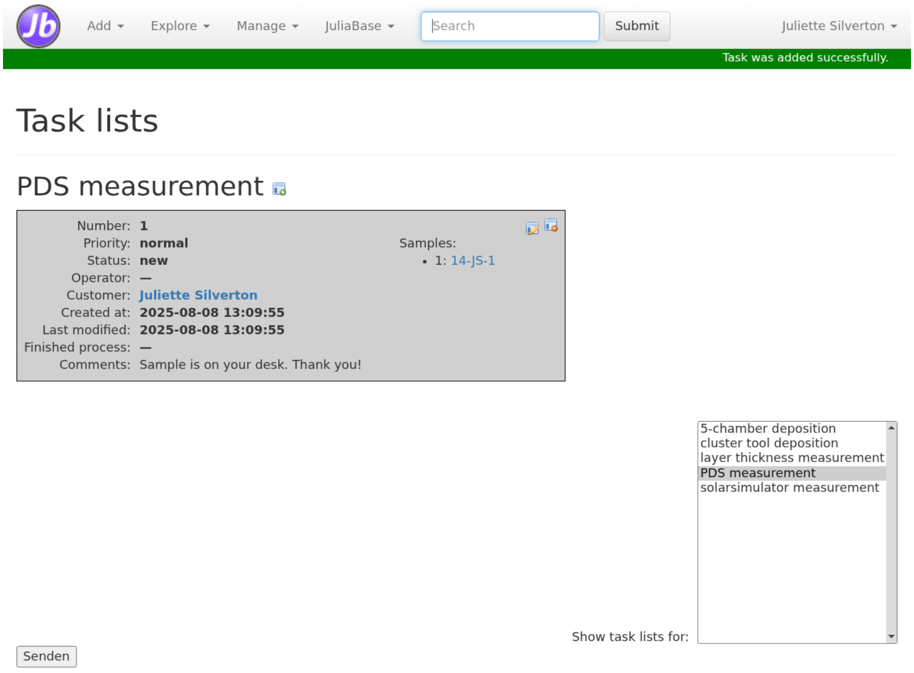
Figure 5.8.: Sample owner.

there is a sample from Juliette that Rosalee would like to take over. Rosalee has already found this sample using the search function (Figure 5.7). In principle, Juliette could assign Rosalee as the person currently responsible for the sample, but Juliette may not have time for this, or may no longer be working at the INM. It could also be a case of samples that have been imported into the database from old data and do not yet have an owner. In any case, in these situations, JuliaBase offers the option of essentially taking samples for yourself via "sample claim" (see Figure 10). Rosalee still has to select an authorised reviewer – her boss Sean Renard volunteers – who then decides on the sample claim.[1ex] **Juliette: The work distributor**

Juliette has a lot to do and cannot take care of mundane tasks such as sample preparation and characterisation herself. She therefore assigns tasks to other people and analyses the results. Log out and log back in as *j.silverton/12345*.

Adding a task:

Let's assume that Juliette wants to have a PDS measurement performed for her sample *14-JS-1*. To



Task was added successfully.

Task lists

PDS measurement

Number:	1
Priority:	normal
Status:	new
Operator:	—
Customer:	Juliette Silverton
Created at:	2025-08-08 13:09:55
Last modified:	2025-08-08 13:09:55
Finished process:	—
Comments:	Sample is on your desk. Thank you!

Samples:

- 1: 14-JS-1

Show task lists for:

- 5-chamber deposition
- cluster tool deposition
- layer thickness measurement
- PDS measurement
- solarsimulator measurement

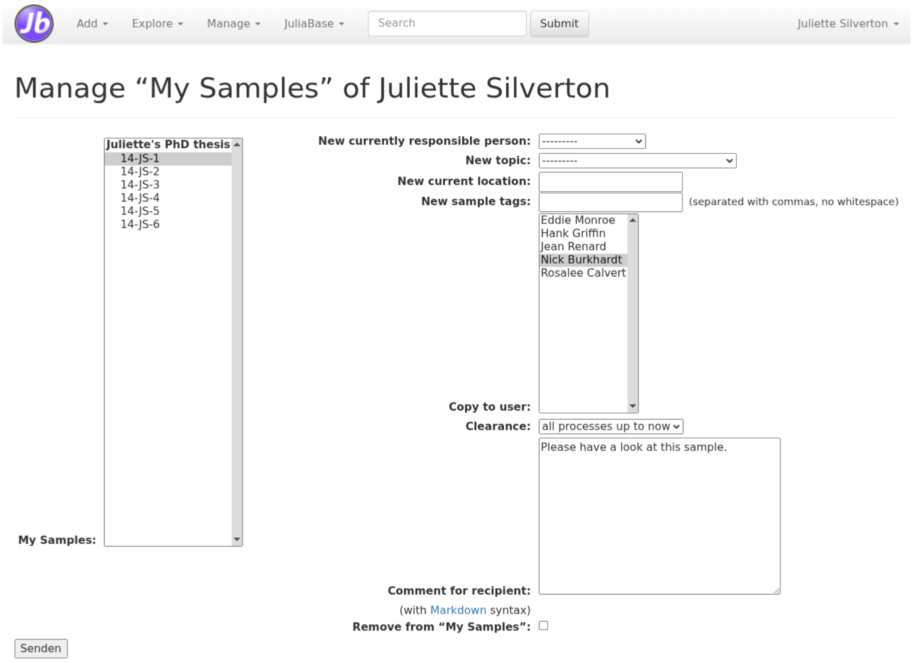
Senden

Figure 5.9.: Adding a task.

do this, go to "*Miscellaneous – Task Lists*" in the main menu (Figure 5.9). First, select all processes that are relevant to you: Select "*PDS measurements*" and click on "*Send*". Now add a new task for the PDS operator (that is Nick Burkhardt) by clicking on the "*plus symbol*" for "*PDS measurements*". Select sample *14-JS-1*, click on "*Send*" and you're done. You can see the new task in the list of tasks.

Sending a sample to another user:

Juliette wants to show Nick the sample *14-JS-1* so that he can take a look at it. Of course, Nick could search for the sample himself, but since the sample is related to Juliette's doctoral thesis and not to Nick, he cannot view the sample data sheet. To send the sample to Nick, click on the "pencil" icon next to "My Samples" on the main menu page. Select sample *14-JS-1* on the left-hand side. Then select Nick in the multiple selection "*Copy to user*" on the right-hand side and enter, for example, "*Please take a look at this sample*" under "*Comment for recipient*". Finally, set "Share" to all previous processes for the sample, because Juliette wants Nick to be able to see the entire data sheet for *14-JS-1*.



Manage "My Samples" of Juliette Silverton

My Samples:
Juliette's PhD thesis
14-JS-1
14-JS-2
14-JS-3
14-JS-4
14-JS-5
14-JS-6

New currently responsible person: ---------
New topic: ---------
New current location:
New sample tags: (separated with commas, no whitespace)
Copy to user: Eddie Monroe, Hank Griffin, Jean Renard, Nick Burkhardt, Rosalee Calvert
Clearance: all processes up to now
Comment for recipient: Please have a look at this sample.
(with Markdown syntax)
Remove from "My Samples": □

Senden

Figure 5.10.: Sending a sample to another user.

Nick: View sent sample

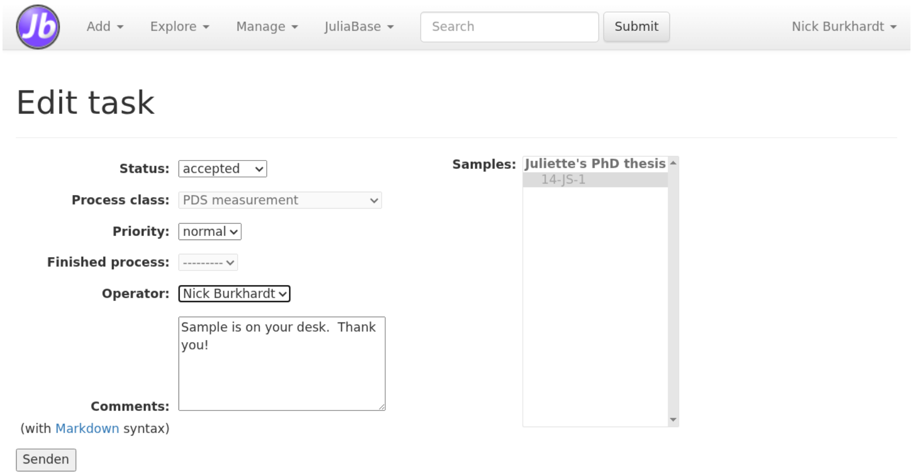
Now log in as *n.burkhardt/12345*. You will see *14-JS-1* under "*My samples*" and can view the data sheet. The sample has been sent successfully!

The news feed:

Nick was notified about the sample submission in the *Main menu – Miscellaneous – News feed*. There he can also see that Juliette has created a new order for a PDS measurement. The *news feed* contains all important news for the respective user: changes to their samples, new samples in their topics, samples transferred to them, new orders and much more. The *news feed* is not actually intended to be displayed in the browser. You can do so, but it is a little cumbersome. Instead, use an RSS feed-enabled programme such as Thunderbird. This programme can also show you which entries in the feed are really new.

Tasks:

Since Nick has read that Juliette has submitted a new PDS task, he visits the "*Task Lists*" page himself



Edit task

Status: accepted

Process class: PDS measurement

Priority: normal

Finished process: ---------

Operator: Nick Burkhardt

Comments: Sample is on your desk. Thank you!

(with Markdown syntax)

Samples: Juliette's PhD thesis
14-JS-1

Figure 5.11.: Task for Nick.

(Figure 5.11). As explained above, the first time you visit this page, you must select the *PDS* and click on "*Send*" so that Nick can see the PDS tasks. Usually, more than one person works on a system such as the PDS. Sometimes people are absent (holidays, illness, etc.). Therefore, it is not clear from the outset who will actually complete a job, and the job must be explicitly accepted and/or assigned to someone. To do this, click on the *pencil icon* to edit it. Set the *status* to "*accepted*" and transfer the task to Nick himself. Juliette will be notified of this. When Nick performs the measurement, he can set the status of the task to "in progress" and then to "completed". A completed task can even be linked to the specific PDS measurement. Some of these steps are optional. They depend on the workflow at your institute.

Sean: The team leader

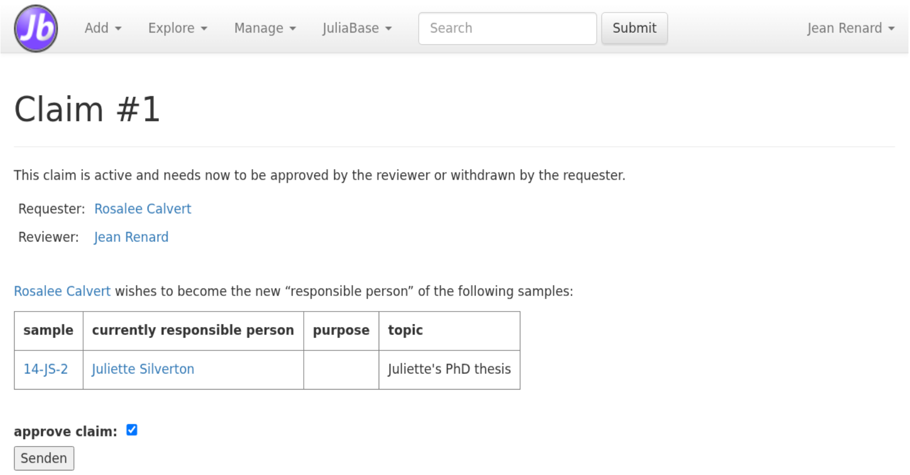
Log in as *s.renard/12345*. Sean is the team leader and has extended rights. These are:

- View all samples
- Create new topics
- Change memberships in all topics
- Grant and revoke permissions for all equipment and experiments
- Approve or reject sample requests

The *Institute of Nifty New Materials* (INM) has only two levels: the team leader and the rest. You can add additional levels in your organisation, and you can define permissions in other ways. However, in our experience, complex permission rules tend to be more of a hindrance than a help.

Approve a sample claim:

Navigate to "*Main Menu Miscellaneous – Sample Claims*". At the bottom of this page, you will see



Claim #1

This claim is active and needs now to be approved by the reviewer or withdrawn by the requester.

Requester: Rosalee Calvert
Reviewer: Jean Renard

Rosalee Calvert wishes to become the new “responsible person” of the following samples:

sample	currently responsible person	purpose	topic
14-JS-2	Juliette Silverton		Juliette's PhD thesis

approve claim: ☑

Senden

Figure 5.12.: Approve sample claim.

Rosalee’s claim for Juliette’s samples (Figure 5.12). Click on it. Sean can now review the claim in detail and approve or reject it.

5.3.2. eLabFTW

The electronic laboratory notebook eLabFTW (`https://www.elabftw.net/`) is another free alternative to an ELN, and this system is gaining popularity internationally. eLabFTW is constantly being developed (the current version in January 2026 is 5.3.11), with the developers deliberately seeking contact with users and welcoming suggestions for improvement from the community. If you would like commercial support for your eLabFTW instance, you can also obtain this from the small company Deltablot (`https://www.deltablot.com/`). In this manual, we only describe the most essential menus and workflows to ensure a successful start with this very powerful system. eLabFTW offers a wide range of options and possibilities for organising, conducting and automating experiments. The aim is to make work easier for its users and to ensure the secure handling of experimental data. Detailed instructions in English are provided in the eLabFTW online manual, which is always up to date and divided into sections with different information for users, administrators and system administrators (https://doc.elabftw.net/index.html). The level of detail available goes far beyond what we can and want to present in this best practice guide. We only want to provide readers of this guide with the practical information they need to get started as easily as possible.

Why use eLabFTW?

eLabFTW is open source and offers a number of advantages that make this system attractive to many users. It offers many design and customisation options to suit individual needs and can be configured for most laboratory environments with reasonable effort. The system is web-based, with no clients to install. A networked computer and a browser are sufficient. Thanks to its responsive design, eLabFTW can be used on networked devices with any screen size, from mobile phones to giant screens. Other important advantages are:

- Secure time stamps can be used for experiments (RFC 3161 as standard, or via a blockchain).
- The authenticity of an experiment can be additionally secured with a personal signature.
- The system can exchange data via interfaces (REST API).
- Various common file formats are available for importing and exporting data: PDF, ZIP, CSV, JSON, QR code.
- There is a comprehensive role and rights management system.
- For repetitive experiments, templates can be created and used individually.
- Databases for products/chemicals and protocols can be created (inventory management).
- To-do lists can be created.
- Laboratories and entire teams can be organised via a scheduler.
- There is a JSON and a molecule editor.
- eLabFTW has now been translated into 21 languages and each user can set her/his preferred language.

- You can run eLabFTW on your own closed network if you have special security requirements and want to handle the maintenance and operation of the ELN entirely yourself.
- A large number of different teams can work simultaneously with one installed instance without any overlap or conflicts (one of the authors of this manual currently works with more than 20 different teams at his institute using one instance).

Required infrastructure

The easiest way to install eLabFTW is with the help of a Docker container on a dedicated Linux server. Other options include installation in a cloud, on a NAS server, or on a Mac or Windows system. The authors of this manual prefer the first option, installation in a Docker container on a Linux server, as this option is quite simple and secure to operate and maintain, and can also be expanded if necessary (e.g. with additional storage space). For details on installation, please refer to the comprehensive eLabFTW online help (`https://doc.elabftw.net/`).

Role and rights management

eLabFTW has a sophisticated role and rights management system based on a strict hierarchical structure. The following diagram illustrates the different levels, with rights decreasing from top to bottom. The system administrator (Sysadmin for short) has full control over the installed instance and can simultaneously interact with the administrators (Admin for short) and users. They monitor ongoing operations and eliminate system malfunctions. However, the Admins and users are responsible for organising the practical work. The following diagram shows possible team divisions as they can be made by the Admins: When new users are added to the system, they must first be informed which team they will be assigned to. In a team, the first admin can appoint additional admins, who then have the same rights as the first admin. The users in a team can be further divided into groups, for example because they are to work together on a research project. It is also possible for an admin to invite users from other teams to collaborate in a group. The following table summarises the different rights of admins and users. We will discuss these differences in more detail below.

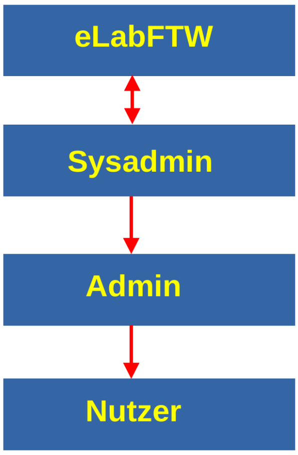


Figure 5.13.: Role management in eLabFTW (Nutzer -> User).

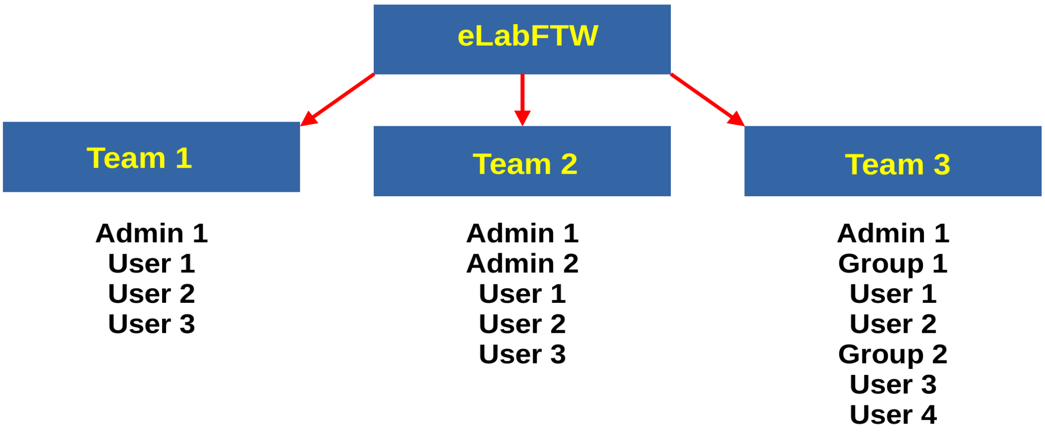
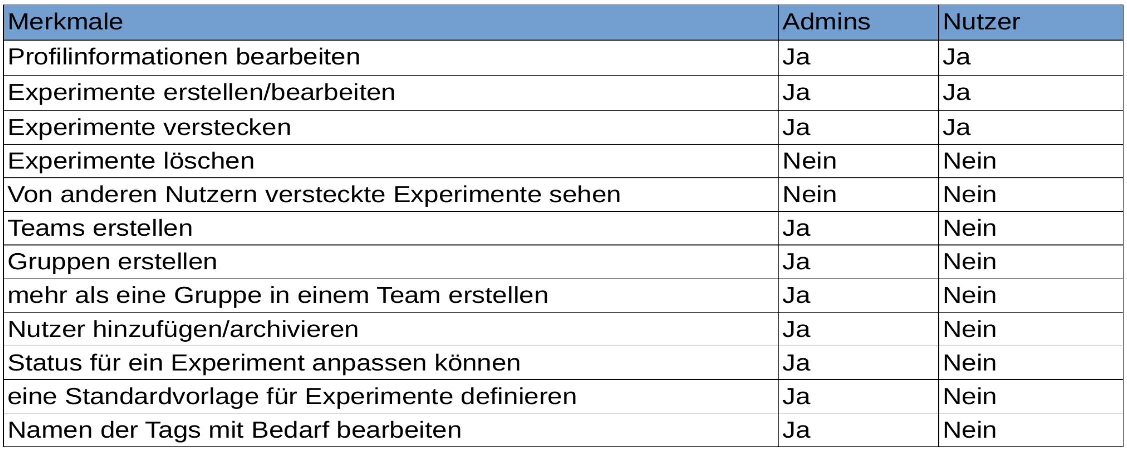


Figure 5.14.: Three examples of possible team divisions. The division of users into specific teams is carried out by the respective admins. Further variations are possible.

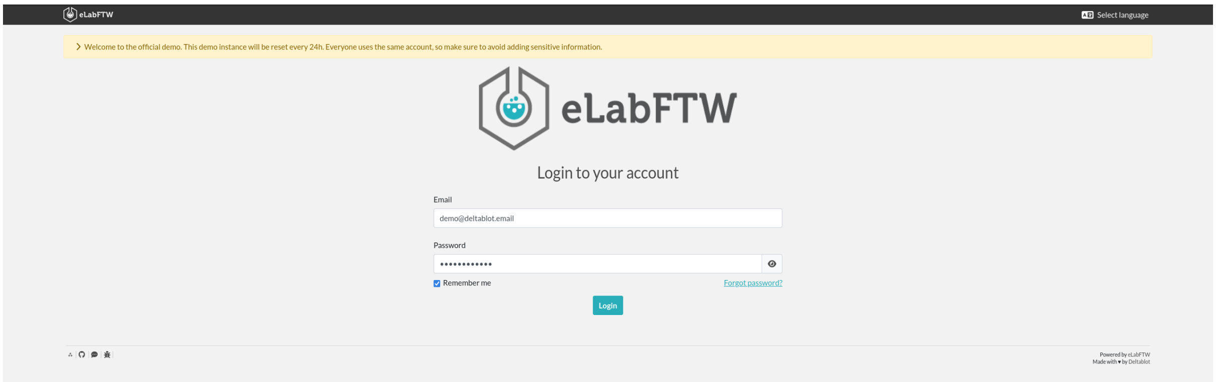


Merkmale	Admins	Nutzer
Profilinformationen bearbeiten	Ja	Ja
Experimente erstellen/bearbeiten	Ja	Ja
Experimente verstecken	Ja	Ja
Experimente löschen	Nein	Nein
Von anderen Nutzern versteckte Experimente sehen	Nein	Nein
Teams erstellen	Ja	Nein
Gruppen erstellen	Ja	Nein
mehr als eine Gruppe in einem Team erstellen	Ja	Nein
Nutzer hinzufügen/archivieren	Ja	Nein
Status für ein Experiment anpassen können	Ja	Nein
eine Standardvorlage für Experimente definieren	Ja	Nein
Namen der Tags mit Bedarf bearbeiten	Ja	Nein

Figure 5.15.: The different rights of administrators and users in eLabFTW.

Introduction to how eLabFTW works: Login and EXPERIMENTS

To access eLabFTW, users must first be registered in the system by their administrator and assigned to a team (see Figure 5.16). You can access the system with an initial password, which should be changed after the first login. In this example, the publicly accessible demo instance of eLabFTW is used (`https://demo.elabftw.net/`). We highly recommend starting with this demo version. It allows you to try out and familiarise yourself with most of the functions without any risk of damaging any data. You can also enjoy the convenience of trying out the functions and menus of the latest version in a simple manner. The screen with the EXPERIMENTS area appears (Figure 5.17). At the top of the screen, you will see a row of tabs (white text on a black background). Clicking on the eLabFTW logo or the house icon opens a dashboard that provides an overview of your current work. To the right of the EXPERIMENTS tab (where we are currently located in this menu), the next tab, TEMPLATES, appears. This shows the templates available for a team, i.e. macros for simplified and repeated execution of experiments. The experiments have been further developed into



Welcome to the official demo. This demo instance will be reset every 24h. Everyone uses the same account, so make sure to avoid adding sensitive information.

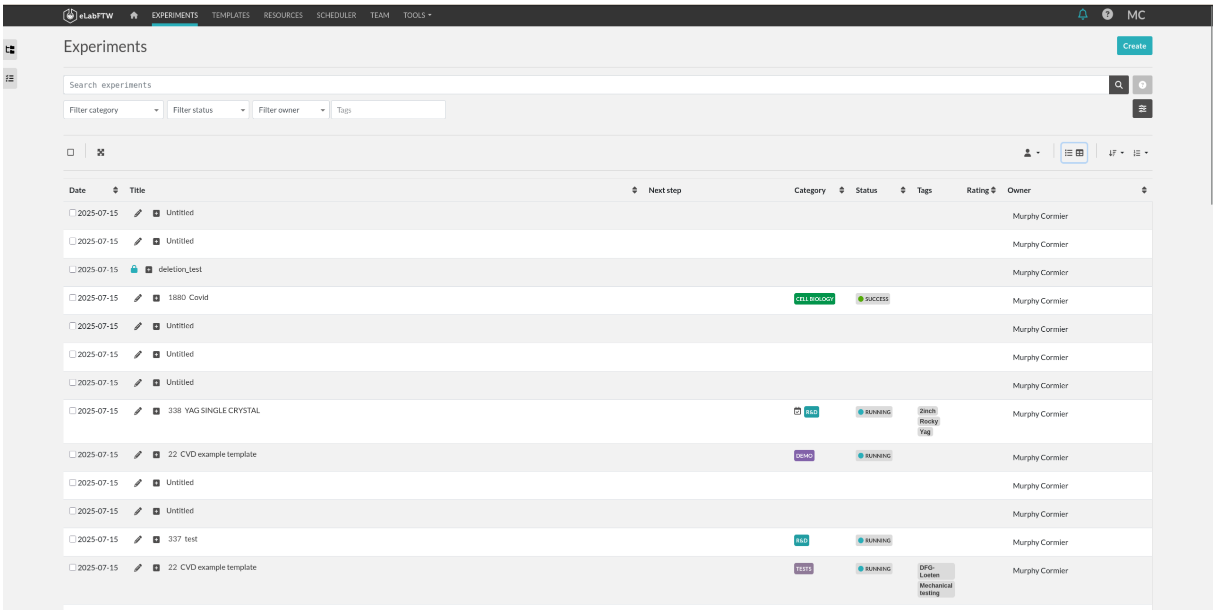
Email: demo@deltablot.email

Password: ••••••••••••

☑ Remember me

Forgot password?

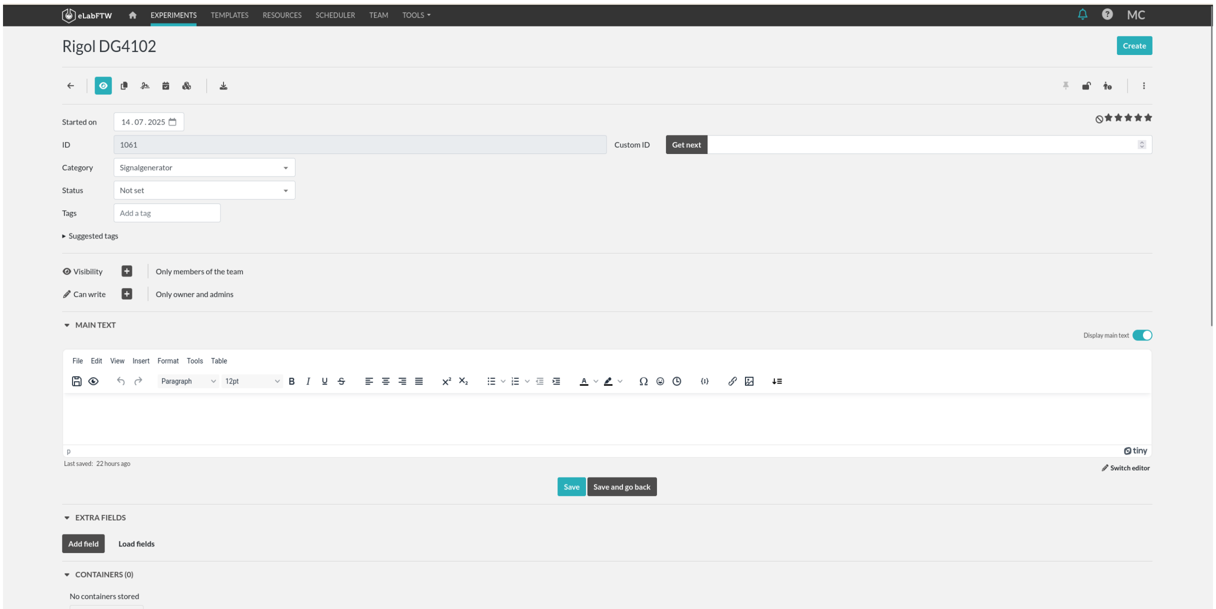
Figure 5.16.: eLabFTW browser interface: Login.



Date	Title	Next step	Category	Status	Tags	Rating	Owner
□ 2025-07-15	Untitled						Murphy Cormier
□ 2025-07-15	Untitled						Murphy Cormier
□ 2025-07-15	deletion_test						Murphy Cormier
□ 2025-07-15	1880 Covid		CELL BIOLOGY	SUCCESS			Murphy Cormier
□ 2025-07-15	Untitled						Murphy Cormier
□ 2025-07-15	Untitled						Murphy Cormier
□ 2025-07-15	Untitled						Murphy Cormier
□ 2025-07-15	338 YAG SINGLE CRYSTAL		R&D	RUNNING	2inch Rocky Yag		Murphy Cormier
□ 2025-07-15	22 CVD example template		DEMO	RUNNING			Murphy Cormier
□ 2025-07-15	Untitled						Murphy Cormier
□ 2025-07-15	Untitled						Murphy Cormier
□ 2025-07-15	337 test		R&D	RUNNING			Murphy Cormier
□ 2025-07-15	22 CVD example template		TESTS	RUNNING	DFG-Loeten Mechanical testing		Murphy Cormier

Figure 5.17.: Browser interface of eLabFTW: EXPERIMENTS.

what are known as RESOURCES. We will discuss the differences in more detail below. The next tab, SCHEDULER, is a time planner that teams can use to organise themselves and plan experiments, meetings, etc. The next tab, TEAM, lists all members of a team, including their email addresses. This facilitates communication within eLabFTW, for example. The TOOLS tab functions as a pull-down menu and provides useful additional programs. More on this later. When we select a specific experiment from the list of EXPERIMENTS by clicking on its name with the mouse, the interface for a single experiment appears with new menus and symbols, which we will explain individually in this chapter (Figure 5.18). An editor appears in the centre of this menu, which is available for each experiment and works similarly to a word processor. However, what distinguishes this editor from a normal word processor is the ability to work with the LaTeX typesetting system. For further details, please refer to the eLabFTW manual (`https://doc.elabftw.net/index.html`). In eLabFTW, there are two main types of objects: EXPERIMENTS and RESOURCES. In this chapter, we will focus on experiments. They are the property of the user who created or started an experiment. To simplify the execution of recurring experiments, templates can be created. Completed experiments can be time-stamped for legal reasons, for example for registering patents. It is also possible to add a personal signature to experiments. Both of these measures increase security with regard to a user's property rights. Both procedures will be explained in more detail later in this chapter. The following graphic (Figure 5.19) shows part of the browser interface of eLabFTW, where experiments are displayed. You can choose between the show mode (top) and the list layout (bottom). You can switch between these two modes by left-clicking on the button with the red border. Clicking on the word EXPERIMENTS in the top line of the browser window (see Figure 5.20) takes you to the corresponding menu. Here you can design and start a new experiment. In this example, Leo Riem is the owner of this experiment. The experiment was started on 07.08.2024 and is an X-ray diffraction measurement (title: Empyrean) using an X-ray diffractometer called Empyrean (hence the title). The window contains several symbols, which are explained individually below.



Rigol DG4102

Started on 14.07.2025

ID 1061

Custom ID

Category Signalgenerator

Status Not set

Tags Add a tag

Visibility Only members of the team

Can write Only owner and admins

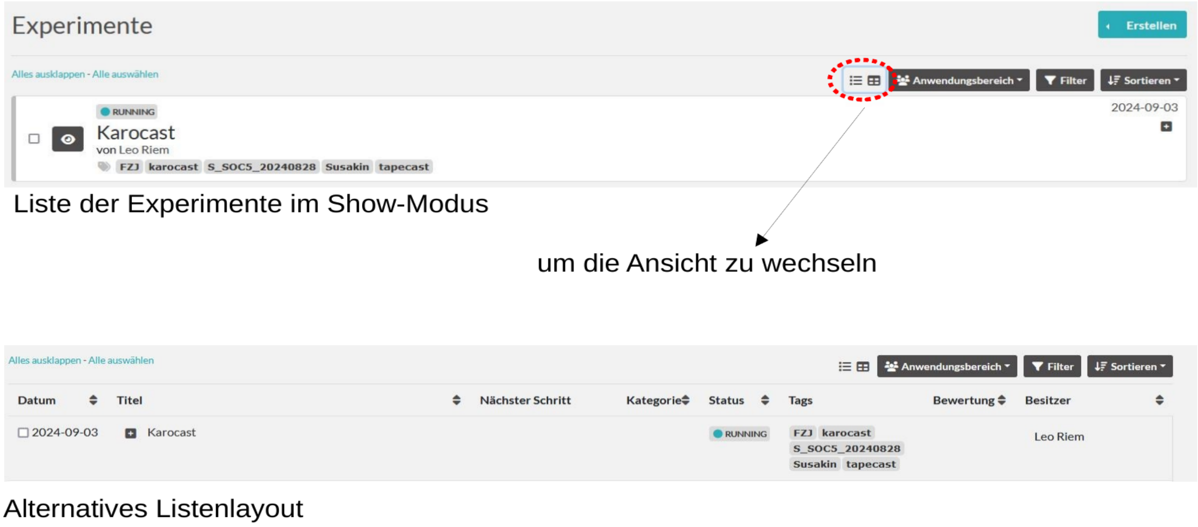
p

Last saved: 22 hours ago

No containers stored

Figure 5.18.: Browser interface of eLabFTW: A single experiment.

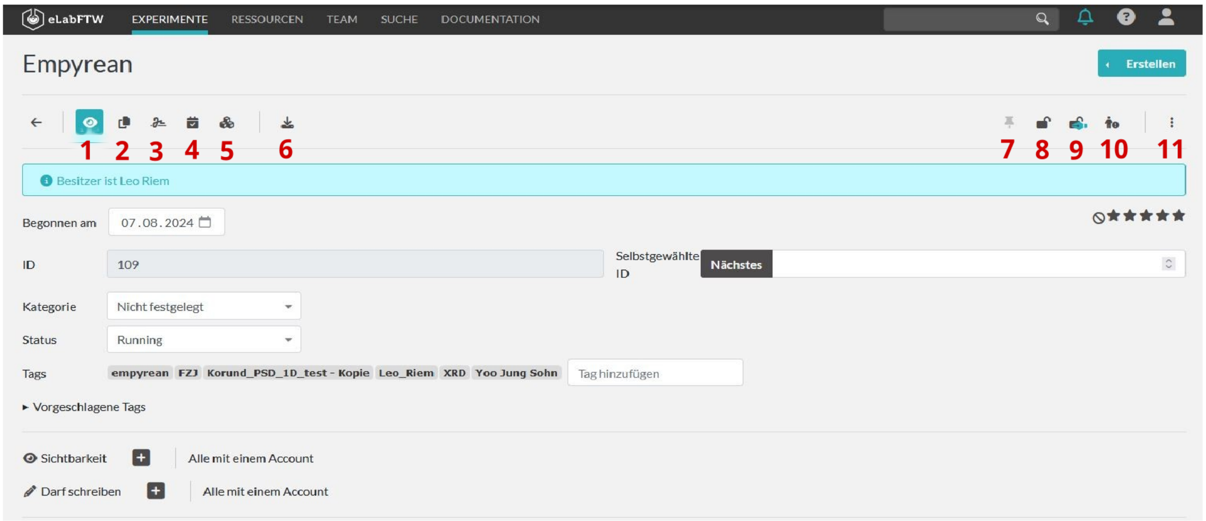
- 1 - Display mode.
- 2 - Duplicate experiments.
- 3 - Add signature.
- 4 - Timestamp.
- 5 - Timestamp using Blockchain.
- 6 - Export experiment to external format.
- 7 - Pin: The entry is placed at the top of the list of experiments.
- 8 - Lock/unlock an element.
- 9 - Activate/deactivate exclusive editing mode.
- 10 - Request action by other users.
- 11 - Ellipsis menu with additional possible functions.



Datum	Titel	Nächster Schritt	Kategorie	Status	Tags	Bewertung	Besitzer
□ 2024-09-03	Karocast			RUNNING	FZJ karocast S_SOC5_20240828 Susakin tapecast		Leo Riem

Figure 5.19.: Browser interface of eLabFTW: Show mode and list layout.

The icon 5 can be used to export experiments to various file formats for further processing if required. In the Ellipsis menu (icon 11), further settings can be made, such as transferring the ownership of an experiment or archiving and restoring an experiment. At this point, you can select another team member to enter as the new owner of an experiment. You can assign a category to each experiment (but you don't have to), but only the administrator can determine which categories are available to a team. These can be categories such as specific projects and their names, or only demo or test experiments, production processes, etc. The status can be used to determine the current phase of an experiment. An experiment may still be running, it may have been successfully completed, or it may need to be repeated. Tags (keywords) can be used to easily label and group experiments. There is no limit to the number of tags that can be used. This extends the metadata for an experiment, similar



Empyrean

Besitzer ist Leo Riem

Begonnen am 07.08.2024

ID 109

Selbstgewählte ID

Kategorie Nicht festgelegt

Status Running

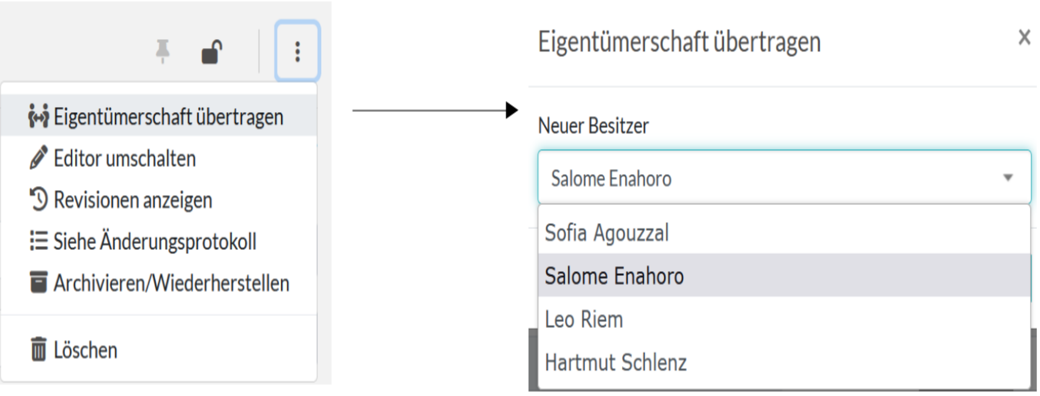
Tags empyrean FZJ Korund_PSD_1D_test - Kopie Leo_Riem XRD Yoo Jung Sohn

▸ Vorgeschlagene Tags

Sichtbarkeit Alle mit einem Account

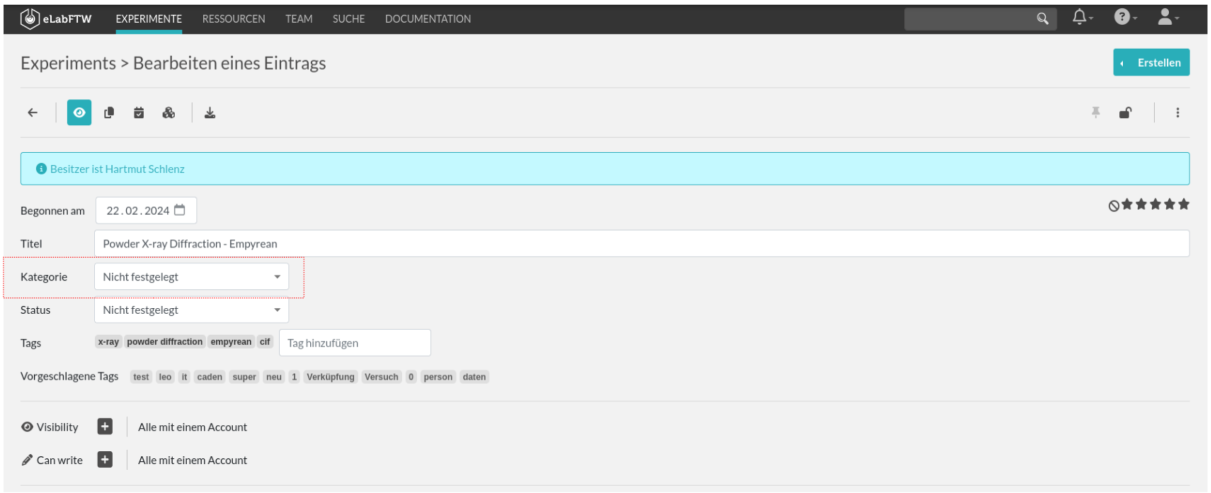
Darf schreiben Alle mit einem Account

Figure 5.20.: Browser interface of eLabFTW: The 11 most important symbols for an experiment.



Neuer Besitzer: Salome Enahoro

Figure 5.21.: Browser interface of eLabFTW: The ellipsis menu.



Experiments > Bearbeiten eines Eintrags

Besitzer ist Hartmut Schlenz

Begonnen am 22.02.2024

Titel Powder X-ray Diffraction - Empyrean

Kategorie Nicht festgelegt

Status Nicht festgelegt

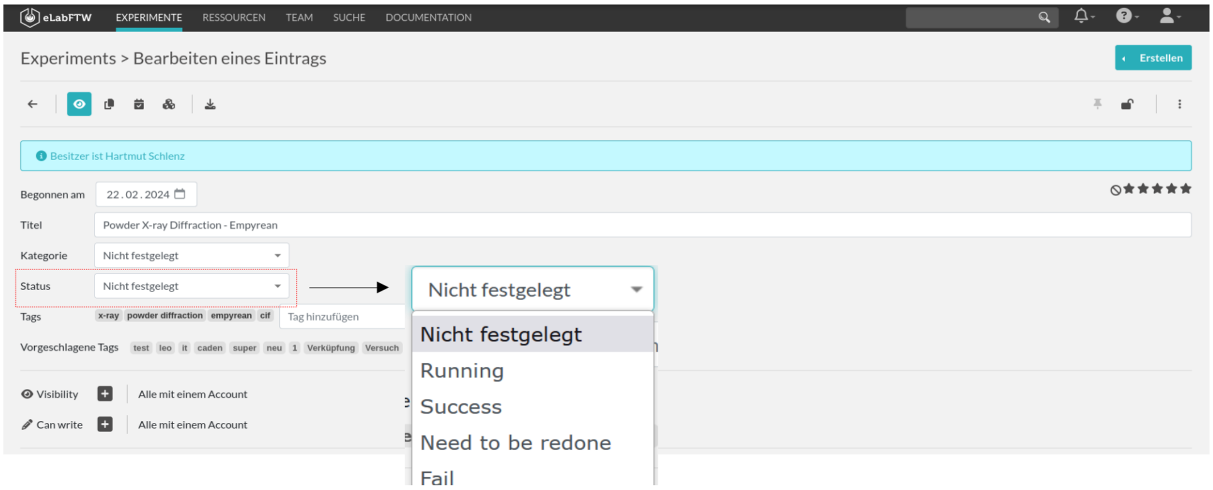
Tags x-ray powder diffraction empyrean cif

Vorgeschlagene Tags test leo it caden super neu 1 Verküpfung Versuch 0 person daten

Visibility Alle mit einem Account

Can write Alle mit einem Account

Figure 5.22.: eLabFTW browser interface: Specify category.



Experiments > Bearbeiten eines Eintrags

Besitzer ist Hartmut Schlenz

Begonnen am: 22.02.2024

Titel: Powder X-ray Diffraction - Empyrean

Kategorie: Nicht festgelegt

Status: Nicht festgelegt

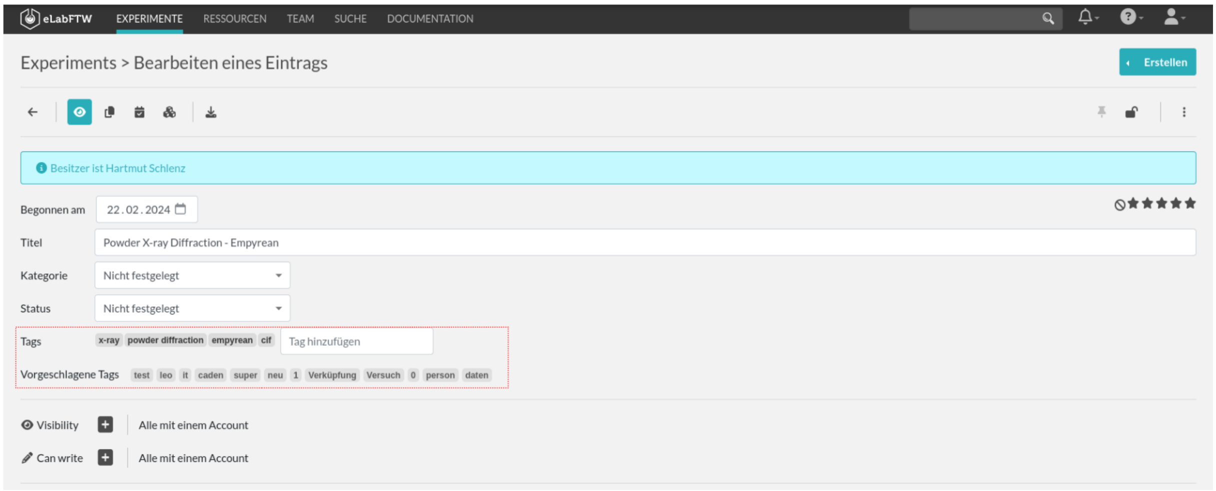
Tags: x-ray, powder diffraction, empyrean, cif — Tag hinzufügen

Vorgeschlagene Tags: test, leo, it, caden, super, neu, 1, Verküpfung, Versuch

Visibility: Alle mit einem Account

Can write: Alle mit einem Account

Figure 5.23.: Browser interface of eLabFTW: Set status.



Experiments > Bearbeiten eines Eintrags

Besitzer ist Hartmut Schlenz

Begonnen am: 22.02.2024

Titel: Powder X-ray Diffraction - Empyrean

Kategorie: Nicht festgelegt

Status: Nicht festgelegt

Tags: x-ray, powder diffraction, empyrean, cif — Tag hinzufügen

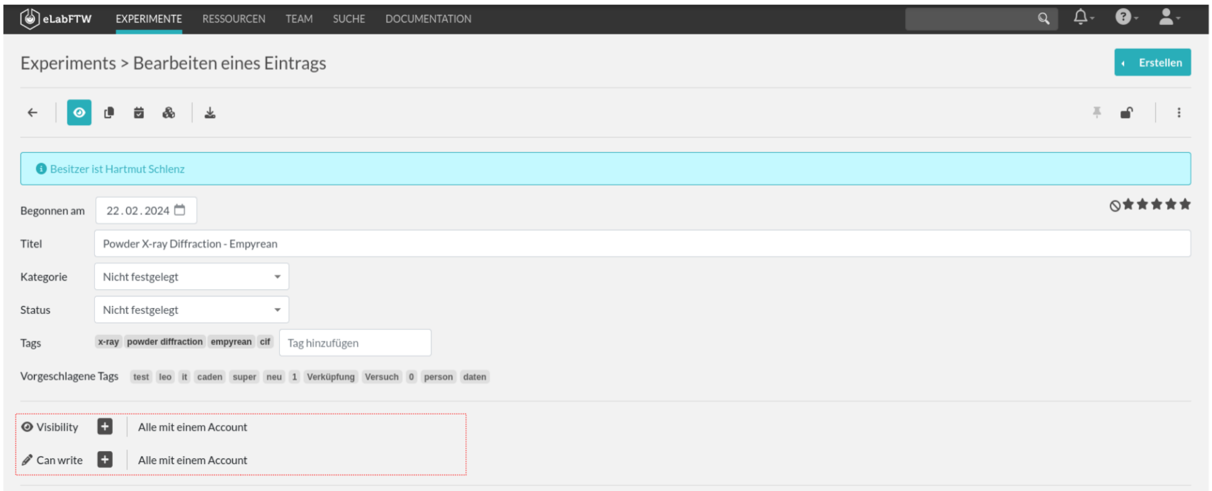
Vorgeschlagene Tags: test, leo, it, caden, super, neu, 1, Verküpfung, Versuch, 0, person, daten

Visibility: Alle mit einem Account

Can write: Alle mit einem Account

Figure 5.24.: Browser interface of eLabFTW: Select tags.

to photo editing software, where tags are also used to make it easier to find images. All experiments with the same tag can be accessed immediately by clicking on or searching for that tag. In edit mode, a single click on a tag is enough to remove it. A list of recently used tags is displayed in the menu. All tags are available to the entire team. The first item is already the settings for Roles and Permissions (more on this later). Under *Visibility* and *Can write*, each user can specify who can view and/or modify her/his experiments and in what form. If a user does not want anyone to be able to view their experiment at all, they can simply select the appropriate setting here. Even an administrator will then have no way of viewing or modifying an experiment. However, this approach does not make sense if you are working on an experiment as a team. In this case, the user who created and started the experiment can specify which team members can work on it.



Experiments > Bearbeiten eines Eintrags

Besitzer ist Hartmut Schlenz

Begonnen am 22.02.2024

Titel Powder X-ray Diffraction - Empyrean

Kategorie Nicht festgelegt

Status Nicht festgelegt

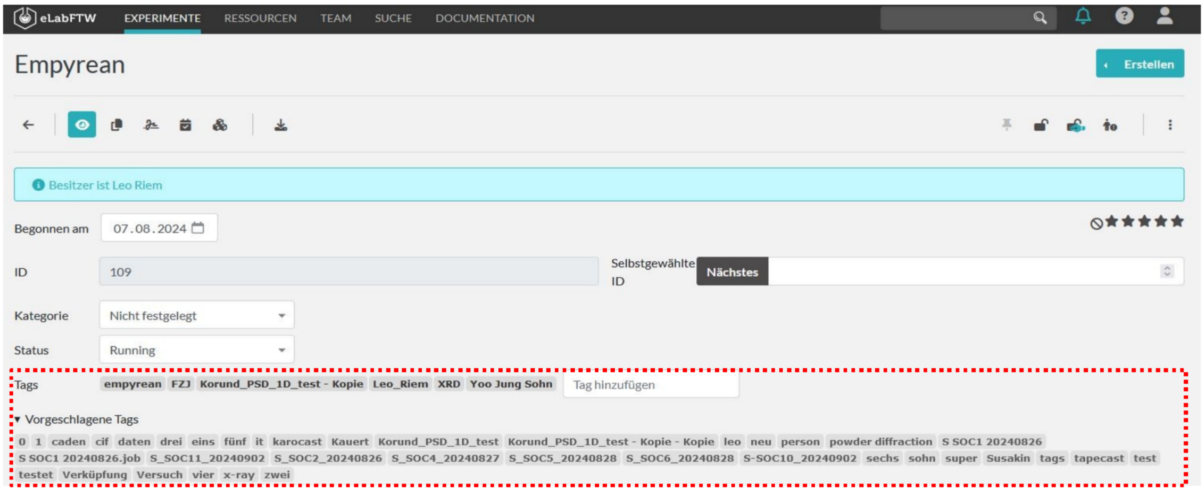
Tags x-ray powder diffraction empyrean cif Tag hinzufügen

Vorgeschlagene Tags test leo it caden super neu 1 Verküpfung Versuch 0 person daten

Visibility Alle mit einem Account

Can write Alle mit einem Account

Figure 5.25.: eLabFTW browser interface: Setting permissions.



Empyrean

Besitzer ist Leo Riem

Begonnen am 07.08.2024

ID 109

Selbstgewählte ID

Kategorie Nicht festgelegt

Status Running

Tags empyrean FZJ Korund_PSD_1D_test - Kopie Leo_Riem XRD Yoo Jung Sohn Tag hinzufügen

▾ Vorgeschlagene Tags

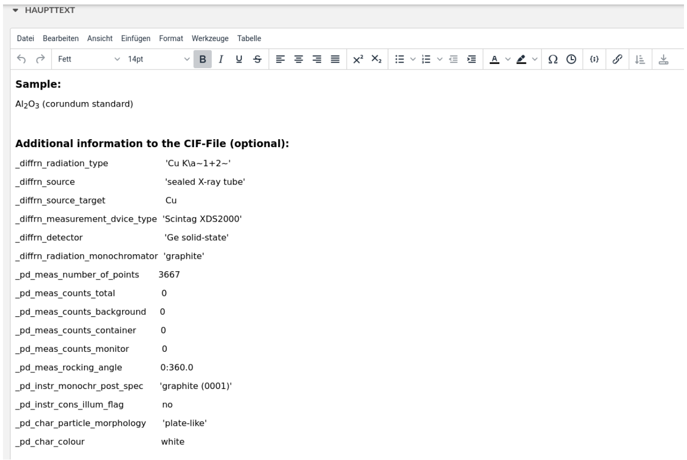
0 1 caden cif daten drei eins fünf it karocast Kauert Korund_PSD_1D_test Korund_PSD_1D_test - Kopie - Kopie leo neu person powder diffraction S SOC1 20240826 S SOC1 20240826.job S_SOC11_20240902 S_SOC2_20240826 S_SOC4_20240827 S_SOC5_20240828 S_SOC6_20240828 S-SOC10_20240902 sechs sohn super Susakin tags tapecast test testet Verküpfung Versuch vier x-ray zwei

Figure 5.26.: eLabFTW browser interface: Tag suggestions by eLabFTW.

We have now briefly explained the most important settings on the first page of a (new) experiment in eLabFTW, once you have created it via the EXPERIMENTS tab. The next section will deal with the creation and use of templates, which can be used to simplify and automate your work with eLabFTW.

TEMPLATES

Figure 5.27 shows the simplest form of a template, with reference to the example of a CIF file in section 2.1. Here, the editor available in each experiment is simply used, which works very similarly to a standard word processor, to write down the necessary information in a structured manner and to be able to change the parameters required for an experiment easily and conveniently. However, creating a template can also be made much more convenient, for example by generating pull-down menus. In Figure 5.28, various temperatures at which an experiment is to be carried out can be selected, among other things. The internal JSON editor is required to generate such templates. The eLabFTW online manual provides clear and detailed examples of how to use this editor to create your own templates, which can then be made available to all team members (`https://doc.elabftw.net/user-guide.html#templates`).



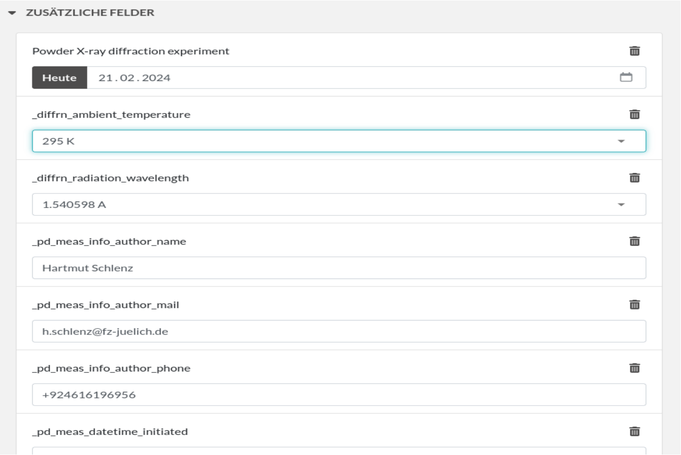
Sample:

Al_2O_3 (corundum standard)

Additional information to the CIF-File (optional):

```
_diffrn_radiation_type              'Cu K\a~1+2~'
_diffrn_source                      'sealed X-ray tube'
_diffrn_source_target               Cu
_diffrn_measurement_dvice_type  'Scintag XDS2000'
_diffrn_detector                    'Ge solid-state'
_diffrn_radiation_monochromator  'graphite'
_pd_meas_number_of_points       3667
_pd_meas_counts_total            0
_pd_meas_counts_background      0
_pd_meas_counts_container        0
_pd_meas_counts_monitor          0
_pd_meas_rocking_angle           0:360.0
_pd_instr_monochr_post_spec     'graphite (0001)'
_pd_instr_cons_illum_flag        no
_pd_char_particle_morphology     'plate-like'
_pd_char_colour                  white
```

Figure 5.27.: Browser interface of eLabFTW: Simple template in the editor.



Powder X-ray diffraction experiment: 21.02.2024

_diffrn_ambient_temperature: 295 K

_diffrn_radiation_wavelength: 1.540598 A

_pd_meas_info_author_name: Hartmut Schlenz

_pd_meas_info_author_mail: h.schlenz@fz-juelich.de

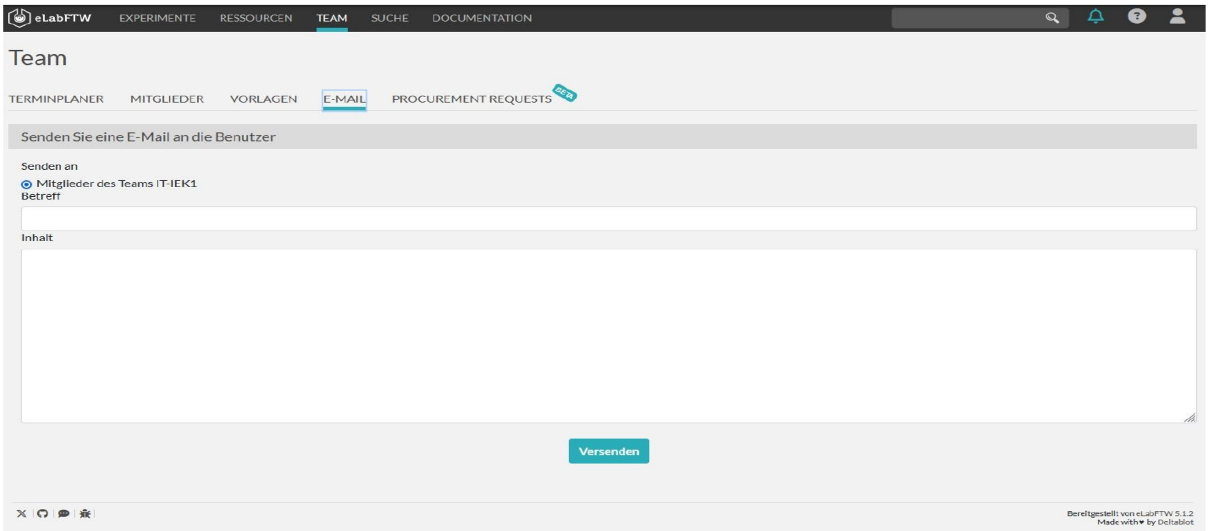
_pd_meas_info_author_phone: +924616196956

_pd_meas_datetime_initiated

Figure 5.28.: Browser interface of eLabFTW: Template with pull-down menus.

TEAM

Messages and information can be exchanged with members of your own team as well as with other users of an eLabFTW instance via the internal Email system, provided that a user is known in the system. Figure 5.29 shows the intuitive interface of the eLabFTW email function. In this example, the menu contains the names and email addresses of your own team members under the menu item MEMBERS. Simply clicking on the email address opens the necessary menu for writing. Alternatively, you can select the menu item EMAIL and compose a message directly. The templates available in a team can be found via the TEMPLATES menu item. To do this, click on the SETTINGS submenu item.



Team

Senden Sie eine E-Mail an die Benutzer

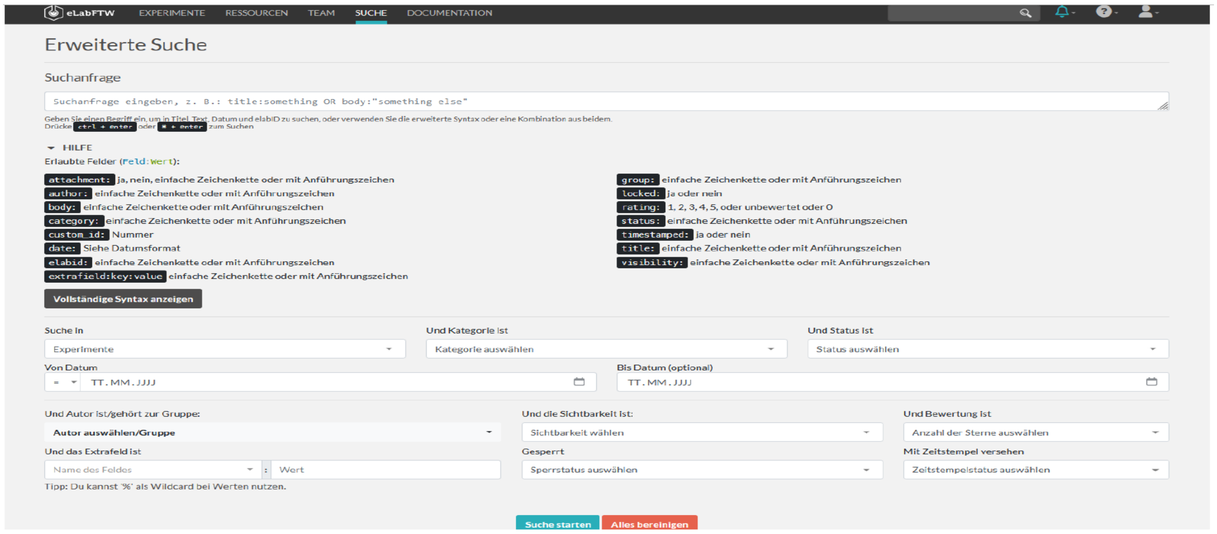
Senden an
● Mitglieder des Teams IT-IEK1
Betreff

Inhalt

Bereitgestellt von eLabFTW 5.1.2
Made with ♥ by Deltablot

Figure 5.29.: Browser interface of eLabFTW: Team communication (email function).

SEARCH



Erweiterte Suche

Suchanfrage

Suchanfrage eingeben, z. B.: title:something OR body:"something else"

Geben Sie einen Begriff ein, um in Titel, Text, Datum und elabID zu suchen, oder verwenden Sie die erweiterte Syntax oder eine Kombination aus beidem.
Drücke `ctrl + enter` oder `⌘ + enter` zum Suchen

HILFE
Erlaubte Felder (Feld:Wert):

- `attachment:` ja, nein, einfache Zeichenkette oder mit Anführungszeichen
- `author:` einfache Zeichenkette oder mit Anführungszeichen
- `body:` einfache Zeichenkette oder mit Anführungszeichen
- `category:` einfache Zeichenkette oder mit Anführungszeichen
- `custom_id:` Nummer
- `date:` Siehe Datumsformat
- `elabid:` einfache Zeichenkette oder mit Anführungszeichen
- `extrafield:key:value` einfache Zeichenkette oder mit Anführungszeichen
- `group:` einfache Zeichenkette oder mit Anführungszeichen
- `locked:` ja oder nein
- `rating:` 1, 2, 3, 4, 5, oder unbewertet oder 0
- `status:` einfache Zeichenkette oder mit Anführungszeichen
- `timestamped:` ja oder nein
- `title:` einfache Zeichenkette oder mit Anführungszeichen
- `visibility:` einfache Zeichenkette oder mit Anführungszeichen

Suche in: Experimente
Und Kategorie ist: Kategorie auswählen
Und Status ist: Status auswählen
Von Datum: = TT.MM.JJJJ
Bis Datum (optional): TT.MM.JJJJ
Und Autor ist/gehört zur Gruppe: Autor auswählen/Gruppe
Und die Sichtbarkeit ist: Sichtbarkeit wählen
Und Bewertung ist: Anzahl der Sterne auswählen
Und das Extrafeld ist: Name des Feldes : Wert
Gesperrt: Sperrstatus auswählen
Mit Zeitstempel versehen: Zeitstempelstatus auswählen
Tipp: Du kannst '%' als Wildcard bei Werten nutzen.

Figure 5.30.: Browser interface of eLabFTW: Detailed search function.

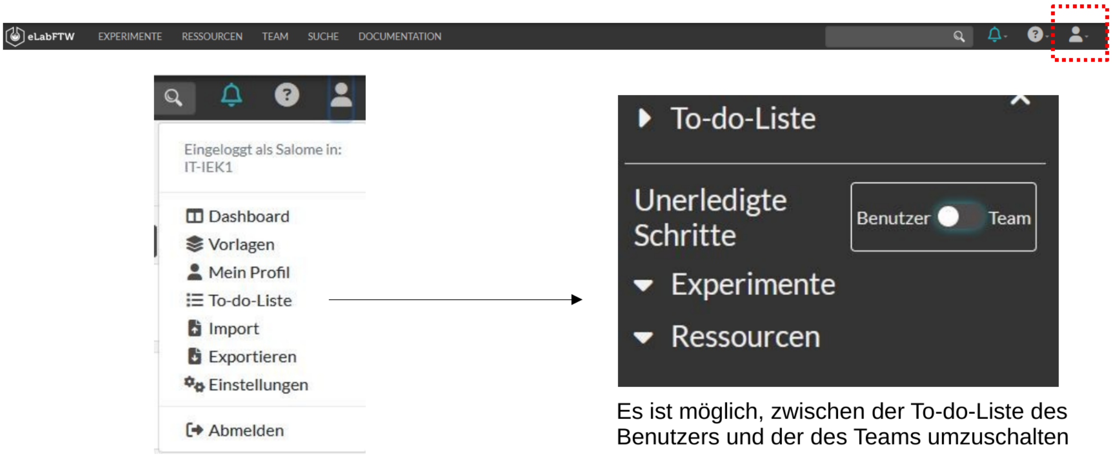
The search function in eLabFTW (see Figure 32) allows for a very detailed search for all possible terms or parameters. Since all experiments and their metadata are always stored in the eLabFTW database, the search function finds all information or data stored in an instance. This makes it possible, among other things, to search for the name of an experiment, for example, linked to a date or a chemical component. A simple search using tags is also possible, which is one of the most effective search methods.

RESOURCES

Resources are similar to experiments, but they serve a different purpose. Resources contain lists and the organisation of *items* that are necessary for an experiment and are grouped together in a package. These resources can be booked by users and are created in advance by the respective administrator. Normal users cannot create new resources themselves, but can only use existing ones. Any type of element can be stored in a resource: entire projects, chemicals, devices and measuring equipment, test benches, etc. In this way, for example, new employees can be provided with everything they need to start work successfully in a single package.

PERSONAL PROFILE AND SETTINGS

By clicking on the button for your own profile (red mark in the top right-hand corner of Figure 5.31), you can configure the system according to your own preferences. Here you can select the language (21 languages are available) or the display on the screen. You can define your own keyboard shortcuts to speed up operation or select output options for exporting PDF files. Figure 5.31 also shows that the personal profile can be used to view to-do lists, both for the respective user and for the entire team. This provides a quick overview of which steps in an experiment or resource are still pending.

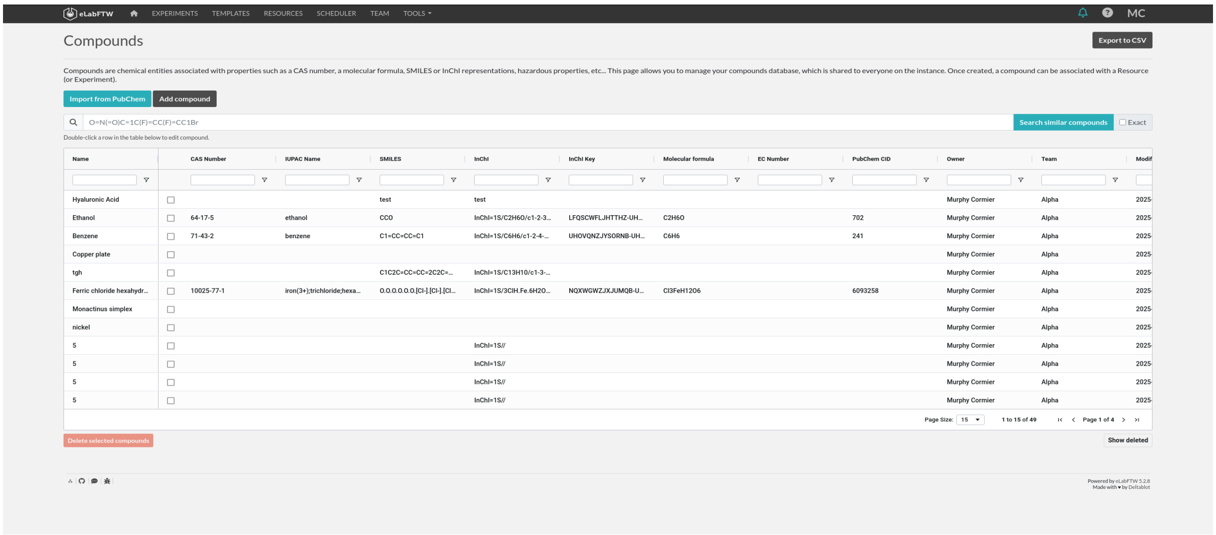


Es ist möglich, zwischen der To-do-Liste des Benutzers und der des Teams umzuschalten

Figure 5.31.: eLabFTW browser interface: Possible settings in the personal profile.

TOOLS (NEW in version 5.2.x of eLabFTW)

Finally, we would like to mention two useful additional programs that can be started via the TOOLS tab. The first is the option to inventory chemicals that are needed and used for experiments. The COMPOUNDS menu is intended for this purpose (see Figure 5.32). Depending on the user, the Molecule Editor (Figure 5.33) may also be useful, which in the current version of eLabFTW 5.3.11 (January 2026) offers extensive options, including the addition of molecules and compounds to the chemical component database (COMPOUNDS) and the option of searching for similar structures in the database.



Compounds

Compounds are chemical entities associated with properties such as a CAS number, a molecular formula, SMILES or InChI representations, hazardous properties, etc... This page allows you to manage your compounds database, which is shared to everyone on the instance. Once created, a compound can be associated with a Resource (or Experiment).

O=N(=O)C=1C(F)=CC(F)=CC1Br

Double-click a row in the table below to edit compound.

Name		CAS Number	IUPAC Name	SMILES	InChI	InChI Key	Molecular formula	EC Number	PubChem CID	Owner	Team	Modif
Hyaluronic Acid	□			test	test					Murphy Cormier	Alpha	2025-
Ethanol	□	64-17-5	ethanol	CCO	InChI=1S/C2H6O/c1-2-3...	LFQSCWFLJHTTHZ-UH...	C2H6O		702	Murphy Cormier	Alpha	2025-
Benzene	□	71-43-2	benzene	C1=CC=CC=C1	InChI=1S/C6H6/c1-2-4-...	UHOVQNZJYSORNB-UH...	C6H6		241	Murphy Cormier	Alpha	2025-
Copper plate	□									Murphy Cormier	Alpha	2025-
tgh	□			C1C2C=CC=CC=2C2C=...	InChI=1S/C13H10/c1-3-...					Murphy Cormier	Alpha	2025-
Ferric chloride hexahydr...	□	10025-77-1	iron(3+);trichloride;hexa...	O.O.O.O.O.O.[Cl-].[Cl-].[Cl...	InChI=1S/3ClH.Fe.6H2O...	NQXWGWZJXJUMQB-U...	Cl3FeH12O6		6093258	Murphy Cormier	Alpha	2025-
Monactinus simplex	□									Murphy Cormier	Alpha	2025-
nickel	□									Murphy Cormier	Alpha	2025-
5	□				InChI=1S//					Murphy Cormier	Alpha	2025-
5	□				InChI=1S//					Murphy Cormier	Alpha	2025-
5	□				InChI=1S//					Murphy Cormier	Alpha	2025-
5	□				InChI=1S//					Murphy Cormier	Alpha	2025-

Page Size: 15 ▾ 1 to 15 of 49 |< < Page 1 of 4 > >|

Powered by eLabFTW 5.2.8
Made with ♥ by Deltablot

Figure 5.32.: Browser interface of eLabFTW: Inventory of chemicals.

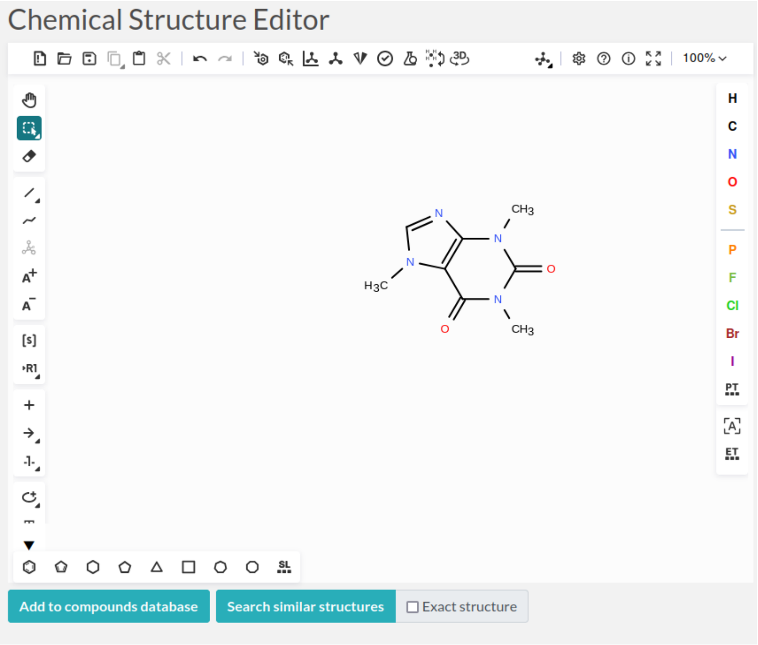

Chemical Structure Editor
100%
H
C
N
O
S
P
F
Cl
Br
I
PT
[S]
R1
N
N
N
N
CH3
CH3
H3C
O
O
3D
SL
Add to compounds database
Search similar structures
□ Exact structure

Figure 5.33.: Browser interface of eLabFTW: Molecule editor.

5.3.3. Kadi4Mat

Kadi4Mat is a generic and open-source virtual research environment. Originally developed in the context of materials science, Kadi4Mat can be used to manage any type of research data within various research disciplines and use cases. Its goal is to combine the ability to manage and exchange data, the repository component, with the ability to analyse, visualise and transform this data, the ELN component (Electronic Lab Notebook). The repository component focuses on warm data, i.e. unpublished data that is still to be analysed, while the ELN component focuses on the automated and documented execution of heterogeneous workflows via an application programming interface (API). This creates a customisable framework that facilitates good practices in research data management and collaboration between researchers. In this respect, Kadi4Mat goes beyond the capabilities of the ELNs JuliaBase and eLabFTW already presented and offers even more possibilities. In the following, we will briefly explain the main components and their application so that interested users can successfully start working with this system. Figure 5.34 illustrates the basic structure of Kadi4Mat. Further information can be found in the publications by Brandt et al. (2021), Griem et al. (2023) and Al-Salman et al. (2023) [**Brandt**, **Griem**, **Al-Salman**].

Similar to eLabFTW, there is also a demo version of Kadi4Mat (`https://demo-kadi4mat.iam.kit.edu/`). We highly recommend this demo to familiarise yourself with the system, as it makes many of the connections clear in an almost playful way and gives you a good feel for the various menus and how the different components interact.

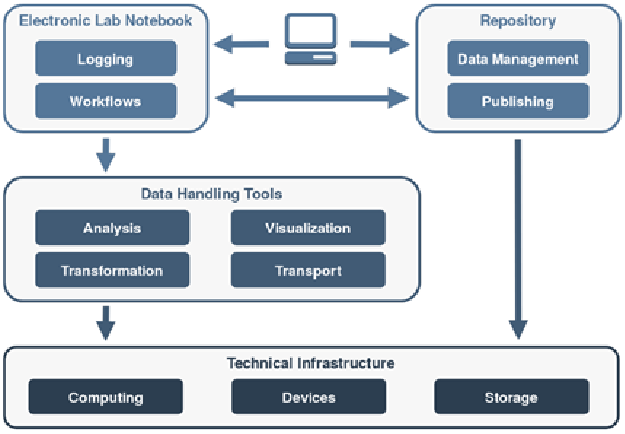


Figure 5.34.: The structure of Kadi4Mat.

Kadi4Mat actually represents an entire ecosystem of different programmes that build on each other and work together in an optimised way. Figure 5.35 shows the various (basic) components of Kadi4Mat, i.e. **Kadi-Web** (which you will mainly deal with as a new user), **Kadi-Studio**, **Kadi-AI** (components for artificial intelligence), **Kadi-FS** and **Kadi-APY**. To familiarise yourself with these components, we recommend using the demo version and the help files it contains, or applying for a user account directly.

Login

There are several ways to log in to Kadi4Mat. Each Kadi4Mat instance can have one or more authentication providers registered, which are briefly described below. **1. Credentials:** This authentication

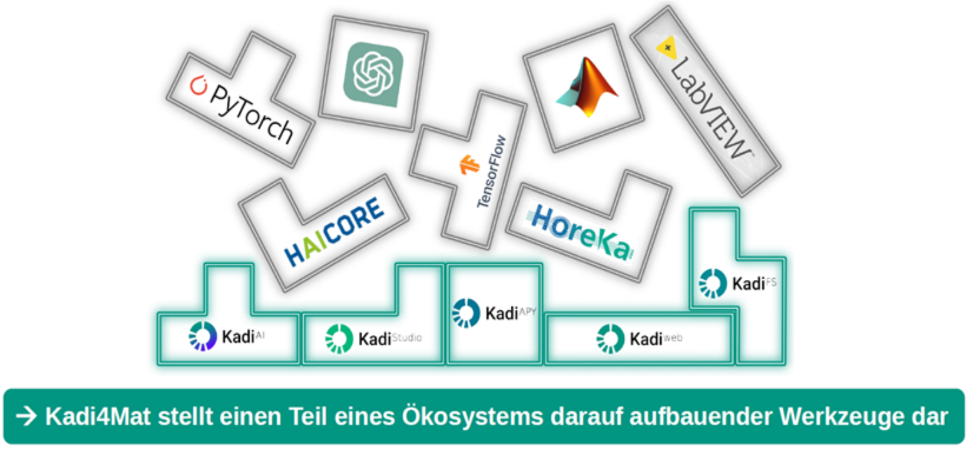
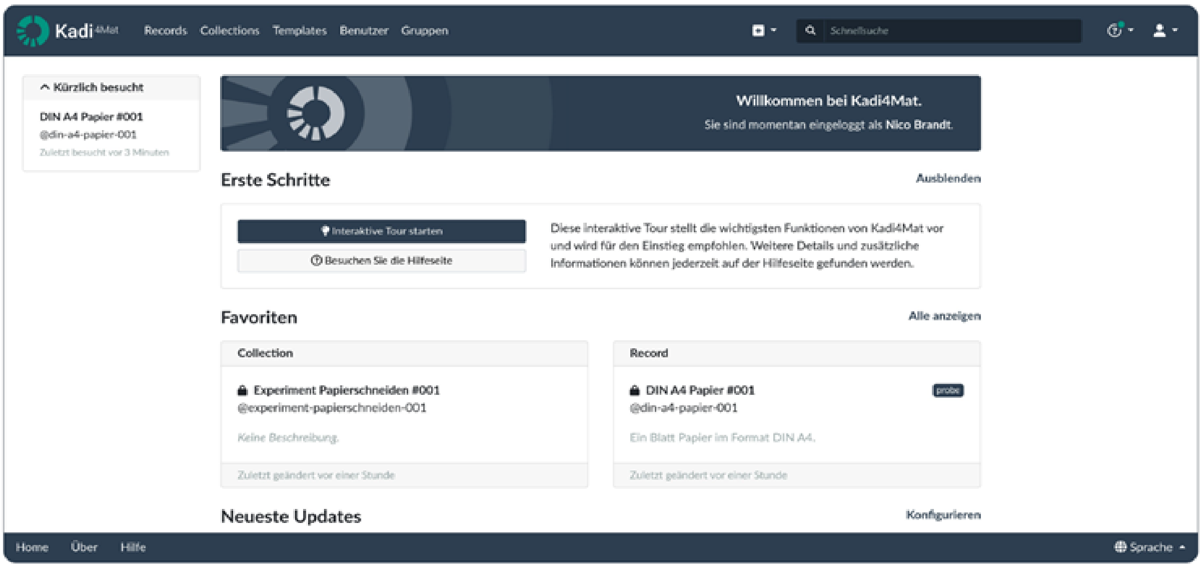

PyTorch
TensorFlow
LabVIEW
HAICORE
HoreKa
Kadi AI
Kadi Studio
Kadi APY
Kadi web
Kadi FS
→ Kadi4Mat stellt einen Teil eines Ökosystems darauf aufbauender Werkzeuge dar

Figure 5.35.: The Kadi4Mat ecosystem.

method uses separate accounts that are local to a specific Kadi4Mat instance. The Credentials of each user consist of a unique username and password. Depending on the configuration of an instance, the registration of new accounts may be possible by individual users or only by system administrators. **2. LDAP:** This authentication method allows users to log in using existing accounts managed by a specific LDAP (Lightweight Directory Access Protocol) installation. LDAP is a directory service commonly used for user authentication and the management of various types of information. Such a service can be used at various levels, e.g. for individual workgroups or entire institutions. **3. OpenID Connect:** This authentication method allows users to log in with existing accounts at one or more third-party web-based services via the OpenID Connect authentication protocol. Each service must be configured separately in Kadi4Mat by a system administrator. When authenticating with one of these services for the first time, users must agree to allow Kadi4Mat to access their user data. **4. Shibboleth:** This authentication method allows users to log in with existing accounts at their home institution via the Shibboleth authentication protocol. Each institution must be configured separately in Kadi4Mat by a system administrator. Please note that the identity providers of some institutions do not pass on all the required user attributes to Kadi4Mat by default. In this case, you may need to contact the administrator of the Shibboleth identity provider. Once you are in the system, you will first see the following browser interface (Figure 5.36): In the top bar with white text on a black background, there are various tabs (similar to eLabFTW) that contain corresponding menus. Figure 5.37 highlights these menus again for clarity. We will discuss the individual menus in detail below.

Navigation

After logging in, the menu at the top left of the navigation bar (Figure 5.37) allows you to access records, collections, templates, users and groups. Each navigation point leads to the corresponding main page of the respective resource, where existing resources can be searched and, if necessary, new resources can be created. Details on the individual resource types are described in the following sections. The first two items at the top right of the navigation bar also provide quick access to the creation of new resources and the search for existing resources. The two drop-down menus on the far right allow you to quickly navigate to various information pages and access the current user's profile page, including their created and deleted resources (see also Users). The latter also provides access to settings and logout. In addition to the navigation bar, there is a navigation bar at the bottom of all pages, regardless of whether you are logged in or not. This footer allows quick navigation to various



Kürzlich besucht

DIN A4 Papier #001
@din-a4-papier-001
Zuletzt besucht vor 3 Minuten

Willkommen bei Kadi4Mat.
Sie sind momentan eingeloggt als Nico Brandt.

Erste Schritte — Ausblenden

Interaktive Tour starten
Besuchen Sie die Hilfeseite

Diese interaktive Tour stellt die wichtigsten Funktionen von Kadi4Mat vor und wird für den Einstieg empfohlen. Weitere Details und zusätzliche Informationen können jederzeit auf der Hilfeseite gefunden werden.

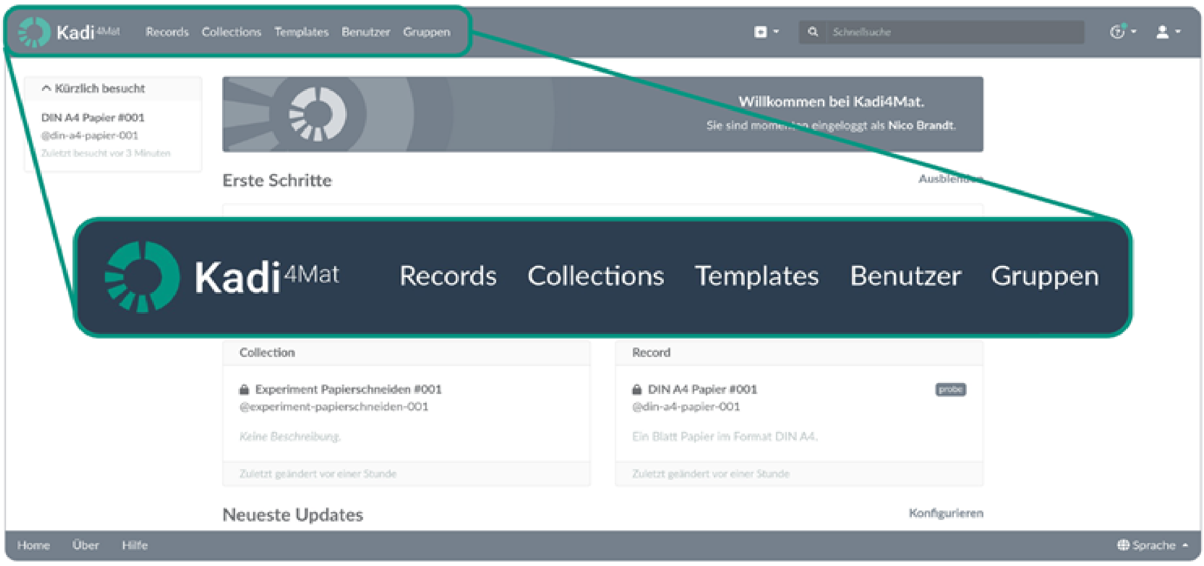
Favoriten — Alle anzeigen

Collection
Experiment Papierschneiden #001
@experiment-papierschneiden-001
Keine Beschreibung.
Zuletzt geändert vor einer Stunde

Record
DIN A4 Papier #001 — probe
@din-a4-papier-001
Ein Blatt Papier im Format DIN A4.
Zuletzt geändert vor einer Stunde

Neueste Updates — Konfigurieren

Figure 5.36.: The Kadi4Mat interface.



Kürzlich besucht
DIN A4 Papier #001
@din-a4-papier-001
Zuletzt besucht vor 3 Minuten
Willkommen bei Kadi4Mat.
Sie sind momentan eingeloggt als Nico Brandt.
Erste Schritte — Ausblenden
Collection
Experiment Papierschneiden #001
@experiment-papierschneiden-001
Keine Beschreibung.
Zuletzt geändert vor einer Stunde
Record
DIN A4 Papier #001 — probe
@din-a4-papier-001
Ein Blatt Papier im Format DIN A4.
Zuletzt geändert vor einer Stunde
Neueste Updates — Konfigurieren
Home Über Hilfe Sprache

Figure 5.37.: The menus in the Kadi4Mat interface.

information pages and also contains a selection for switching the current language.

Data sets

For the creation and further processing of data or data sets in Kadi4Mat, we would like to start this section with a simple example. Our device for conducting an experiment is to be a simple pair of scissors and our sample is a sheet of paper in DIN A4 format. We cut the paper and obtain two new samples, each half the size of a DIN A4 sheet (Figure 5.38). The corresponding graph in Kadi4Mat for this simple experiment in Figure 5.39 shows how the system structures such an experiment. We conduct an experiment called *cutting paper*, using a device called *scissors*, processing a sample (*A4 paper*), and producing two new samples as a result, namely two half sheets of paper. Figure 5.40 shows the general structure of how data is generated and processed in Kadi4Mat. The metadata for *paper cutting* is shown in Figure 5.41. As already explained several times in this manual, experimental (measurement) data and the corresponding metadata always belong together. In Figure 5.40, we can see how both are analysed, processed and, if necessary, published or reused by Kadi4Mat. Data sets are the basic components of Kadi4Mat and can represent any type of digital or digitised object,

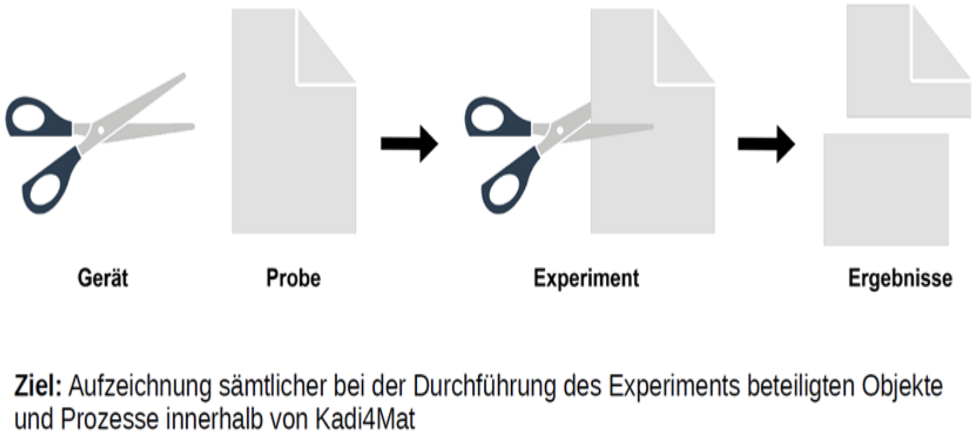

Gerät
Probe
Experiment
Ergebnisse
Ziel: Aufzeichnung sämtlicher bei der Durchführung des Experiments beteiligten Objekte und Prozesse innerhalb von Kadi4Mat

Figure 5.38.: A simple experiment.

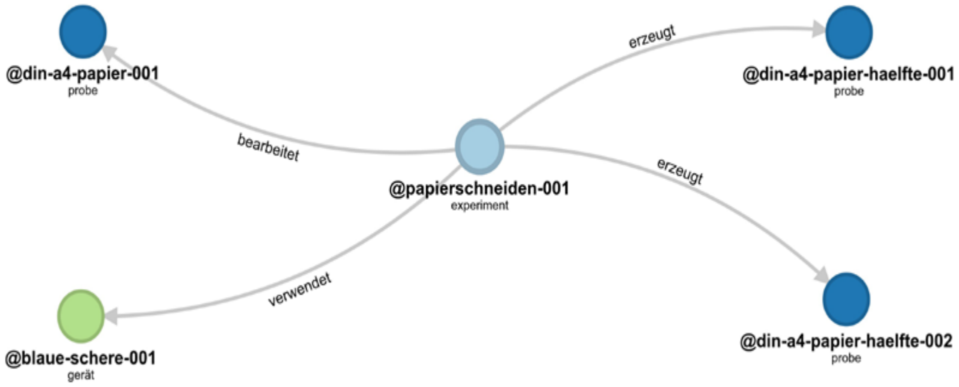


Figure 5.39.: The graph in Kadi4Mat for the simple experiment in Figure 5.38.

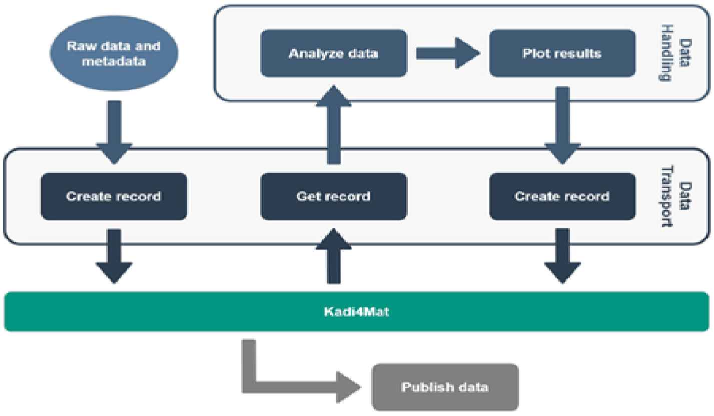
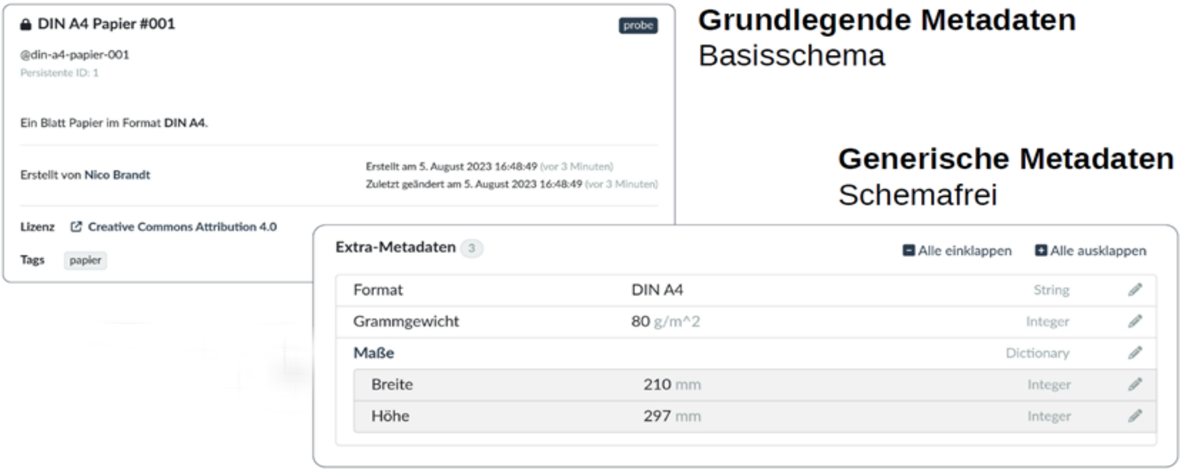


Figure 5.40.: The creation and processing of data sets in Kadi4Mat.



Grundlegende Metadaten
Basisschema

🔒 DIN A4 Papier #001 probe

@din-a4-papier-001
Persistente ID: 1

Ein Blatt Papier im Format **DIN A4**.

Erstellt von Nico Brandt

Erstellt am 5. August 2023 16:48:49 (vor 3 Minuten)
Zuletzt geändert am 5. August 2023 16:48:49 (vor 3 Minuten)

Lizenz Creative Commons Attribution 4.0

Tags papier

Generische Metadaten
Schemafrei

Extra-Metadaten 3 — Alle einklappen — Alle ausklappen

Format	DIN A4	String
Grammgewicht	80 g/m^2	Integer
Maße		Dictionary
Breite	210 mm	Integer
Höhe	297 mm	Integer

Figure 5.41.: The metadata of the simple experiment *Paper cutting*.

e.g. arbitrary research data, samples, experimental equipment or even individual processing steps. Data sets consist of metadata that either stand alone or can be linked to any number of corresponding data. They can also be grouped into collections or linked to other data sets, as described later. To create a new data set, the metadata must first be entered. This includes basic information such as title, (unique) identifier and description. In addition, generic additional metadata specific to the different types of data records can be specified. This metadata consists of extended key-value pairs, where each entry has at least one unique key, a type and a corresponding value. Optionally, a description, an additional term IRI (Internationalised Resource Identifier) and validation instructions can also be specified. It is also possible to create templates for the generic metadata, as described under Templates. The following value types can be used for this metadata: **String** - A single text value.
Integer - A single integer value. Limited to values between -($2^{53} - 1$) and ($2^{53} - 1$). Integer values can optionally have a unit that describes them further.
Float - A single double-precision (64-bit) floating-point value. Float values can optionally be accompanied by a unit that describes them in more detail.
Boolean - A single Boolean value that can be either true or false.
Date - A single date and time value.
Dictionary - A nested value that can be used to combine multiple metadata entries under a single key.
List - A nested value that works similarly to dictionaries, with the difference that none of the values in a list have a key.
Apart from the metadata, it is possible to set the visibility of a data record to either private or public, the latter allowing any logged-in user to search for the data record and view its contents without requiring explicit read permissions. Finally, a dataset can be linked directly to other resources, and its permissions can be set directly, both aspects of which can be managed even after the dataset has been created. Once the metadata for a dataset has been created, the actual data for the dataset can be added in a separate view, to which the application automatically redirects by default. This is only one of the various views available for managing datasets. The next section describes the purpose of the other views that can be selected after returning to the record overview page. For managing existing records, various views are available on the record overview page, each of which can be accessed via the corresponding tab in the navigation menu of a record. The individual tabs and their contents are briefly described below:

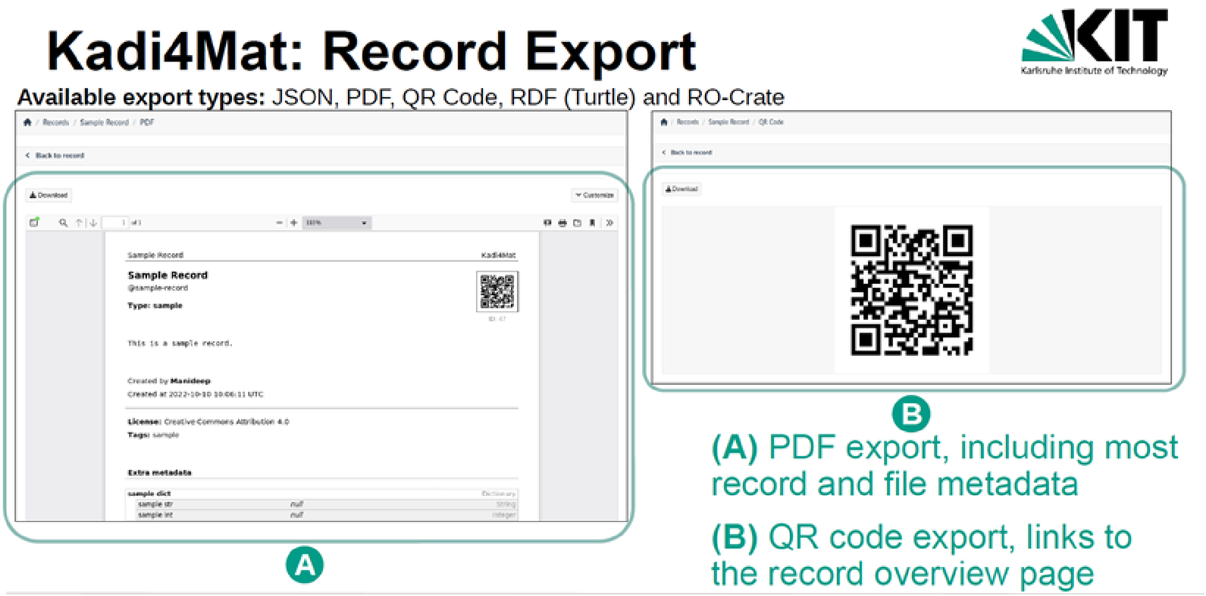
Overview: This tab provides an overview of a data record, mainly in terms of its metadata. Here, it is possible to edit or copy a data record if the corresponding permissions are fulfilled, whereby editing a data record also allows it to be deleted. Note that the data record is first moved to the recycle bin; see also User. Data records can also be exported in various formats, published or added to favourites. Note that the publish function is only available if at least one publication provider is registered with the application.

Files: This tab provides an overview of the files associated with a data record. With the appropriate permissions, new files can be added, which is usually done by uploading locally stored files. However, certain file types can also be created directly via the web interface. Existing files can be downloaded either as a whole or individually via the quick navigation of the respective file. Depending on your permissions, this navigation also displays additional actions for quick file management. Clicking on a file takes you to a separate overview page for the corresponding file, which displays all additional metadata for the file. In addition, many file types have an integrated preview function. Here, it is also possible to edit the metadata or the content of a file, whereby editing the metadata also allows the file to be deleted. For some file types, direct editing of the actual file content is possible; otherwise, the regular upload functionality can be used. **Links:** This tab provides an overview of the resources to which a data set is linked, including other data sets and collections. Collections represent logical groupings of multiple data sets, while links between data sets specify their relationship and may also contain additional metadata. In addition, record links can be visualised in an interactive diagram. Clicking on a record link takes you to a separate view that provides a more detailed overview of the link and the associated records. Linking resources requires linking permission in both resources that are to be linked. Note that users will still not be able to view linked resources unless they have explicit permission to do so. However, a limited amount of information about record links (the link ID, the linked record, the name and the term IRI of the link) is always displayed as part of the record revisions.

Permissions: This tab provides an overview of the access permissions granted to individual users or groups of multiple users for a specific record. New permissions can be granted if the corresponding permissions are met, which currently works via predefined roles. Details about the specific permissions and actions that each role provides can be found by clicking on the Roles popover. Note that group roles for users who can manage permissions are always displayed, even if the group would not normally be visible. This allows existing group roles to be changed and/or removed at any time. Such group roles contain only very limited information about the group itself (its ID, title, identifier, and visibility).

Revisions: This tab provides an overview of changes to the metadata of a record, to the metadata of a file, and to links to other records. Clicking on Show revision of a revision entry opens a separate view that provides a more detailed overview of the respective revision and the corresponding changes.

Similar to other ELNs, Kadi4Mat also allows data records to be exported in various file formats (Figure 5.42). For example, in JSON, PDF, QR code, RDF (Turtle) and RO Crate formats. For detailed questions, please refer to the corresponding help functions in Kadi4Mat.



Kadi4Mat: Record Export

Available export types: JSON, PDF, QR Code, RDF (Turtle) and RO-Crate

Sample Record

Sample Record
@sample-record

Type: sample

This is a sample record.

Created by **Manideep**
Created at 2022-10-10 10:06:11 UTC

License: Creative Commons Attribution 4.0
Tags: sample

Extra metadata

sample dict		Dictionary
sample str	null	String
sample int	null	Integer

Records / Sample Record / PDF
Back to record
Download
Customize
Kadi4Mat
Records / Sample Record / QR Code
Back to record
Download
A
B
Karlsruhe Institute of Technology

(A) PDF export, including most record and file metadata

(B) QR code export, links to the record overview page

Figure 5.42.: Export of data records.

Collections

Collections represent logical groupings (e.g., projects, simulation studies, or experiments) of multiple data sets or other collections. Creating a new collection is similar to creating new data sets, with the difference that only some basic metadata is required. In addition, it is possible to specify a data set template that will be used as the default when adding new data sets to a collection. Note that a limited subset of information about such templates is always displayed as part of the collection revisions (the ID of the template) and when editing the collection (the ID and identifier of the template).
Similar to records, collections have their own navigation menu that provides various tabs for viewing and managing collections. Since most of the content is similar to that of records, only the most important differences are listed in the following sections:

Overview: This tab provides an overview of a collection, mainly in terms of its metadata. The records that are part of a collection are also listed directly in this overview.

Links: In addition to linking collections to records, collections can also be linked to other collections, which is displayed in this tab. This feature can be used to create simple hierarchies of parent and child collections to improve the structuring of multiple resources, e.g. by representing projects and corresponding sub-projects. Note that each collection can only have one parent collection, while the permissions required to link collections are handled similarly to other resource links.

Permissions: In addition to managing the permissions of collections themselves, it is also possible to manage the roles of users and groups of all linked records in a collection on the corresponding tab, as collection permissions are not currently inherited by linked resources. This applies in particular to all linked records for which the current user can manage permissions. If you select an empty role, all existing permissions for the corresponding user or group are removed instead.

Templates

Templates allow you to create drafts for various resources. There are several types of templates that define the actual content that a template can contain. Currently, the following types of templates are available:

Record: Record templates can contain all metadata that can be specified for a record, including its general additional metadata, linked collections, record links, and permissions. Note that for linked resources or groups, their IDs are visible to all users who can access the template. Record templates can be selected in most places where new records can be created, as well as for creating general additional metadata.

Extras: Extras templates are focused on the general additional metadata of a record. This type of template can be selected and combined wherever such metadata can be specified, including when creating other templates.
Creating a new template is similar to creating other types of resources. For each template, at least a title and an identifier for the template itself must be specified, while the actual template data depends on the type of template. Similar to other resources, templates have their own navigation menu with various tabs for displaying and managing templates.

Users

This view displays all registered users of the current Kadi4Mat instance. Clicking on a user takes you to a separate page containing another navigation menu that provides access to various subpages. The respective contents of these pages are briefly described below:

Profile: This page displays basic information about a user, such as their user name and account type. Users can control some of the information displayed on this page via the settings.

Resources: This page displays all accessible resources created by a specific user. When viewing other users, resources that have been (explicitly) shared with a user, either directly or via groups, including shared groups, are also displayed.

Recycle bin: This page is only visible to the current user. It contains deleted resources, namely deleted records, collections, templates or groups. The resources can either be restored or permanently deleted. The latter also happens automatically after 1 week. Until then, the identifier of the deleted resources cannot be reused for newly created resources. Note that currently only the creator of a resource can restore or permanently delete it.

Groups

Groups can be used to combine multiple users, which makes access management easier. Creating a new group is similar to creating other resources. In addition to the basic metadata, a group image can also be uploaded, which is displayed in the group overview and on the search results pages. Existing groups are managed in a similar way to other resources. In addition to the usual content, there are additional tabs that allow you to display resources that have been shared with a group:

Overview: This tab provides an overview of a group, mainly in terms of its metadata. The members of a group are also displayed in this overview. Managing the members of a group is similar to managing the access permissions of other resources, as group membership is linked to a group's access permissions. As long as a user has any role in a group, they are also a member of that group and, of course, have the corresponding rights that their role within the group provides. In addition to directly managing members, it is also possible to automate the assignment of roles within a group by defining rules with various conditions. Each rule is applied when a new user is registered, but can also be applied retroactively to all existing users.

Settings

In the settings, you can manage everything that is not directly related to the actual creation and management of resources. The settings navigation menu provides access to various subpages:

Profile: In this menu, you can change the basic user information that is displayed in a user's profile. Note that some options may be disabled depending on the type of user account.

Password: In this menu, users can change their Password. Depending on the type of user account, this menu item may be hidden.

Preferences: In this menu, users can change their settings regarding the behaviour or appearance of Kadi4Mat.

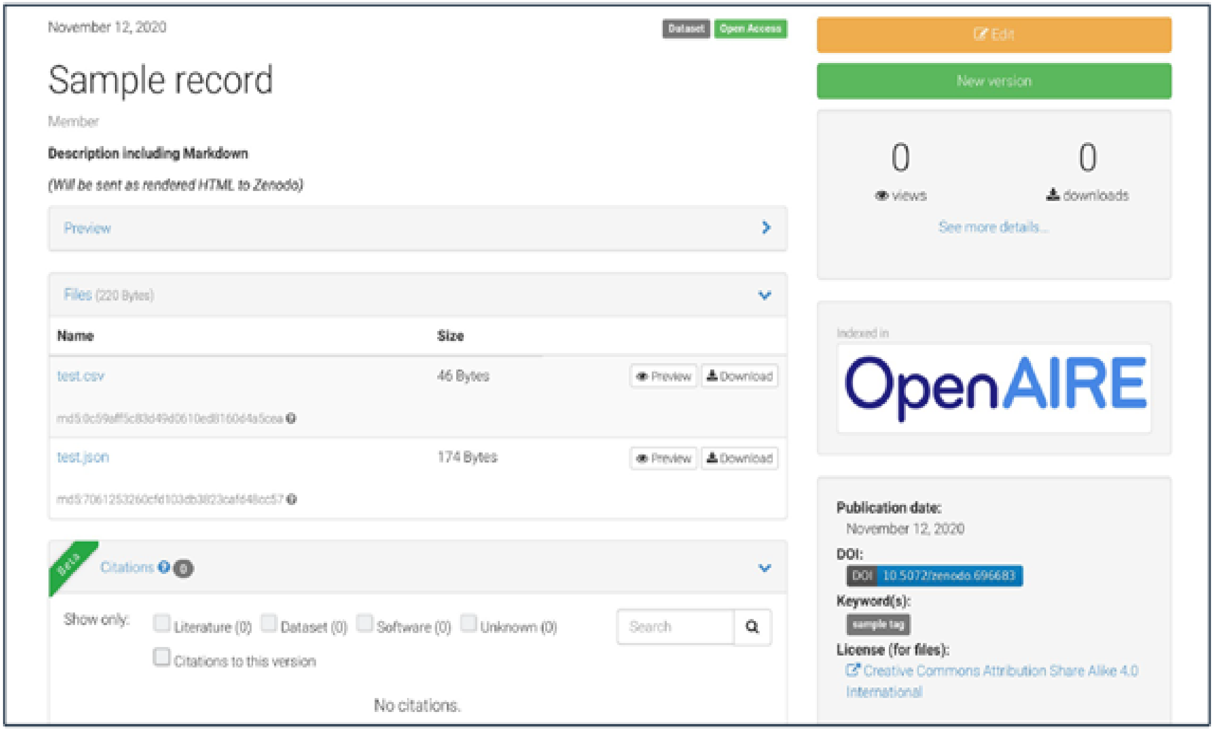
Access tokens: In this menu, personal access tokens (PATs) can be created and managed. Creating a PAT allows direct interaction with the HTTP API provided by Kadi4Mat. Detailed information about the API and PATs can be found in the Kadi4Mat documentation. Some parts of the API can also be tested directly via the web browser by navigating to `https://demo-kadi4mat.iam.kit.edu/api/v1` while a OpenAPI specification of the API can be accessed via `https://demo-kadi4mat.iam.kit.edu/openapi.json?v=v1`. Please note that you must be logged in for both links.

Applications: In this menu, you can register and manage registered or authorised applications. An application can be used to integrate another service with the HTTP API provided by Kadi4Mat. OAuth2 tokens are used for authorisation, which an application can request by implementing the corresponding OAuth2 authorisation flow. Detailed information about the API and OAuth2 tokens can be found in the Kadi4Mat documentation.

Connected services: In this menu, users can manage their connections to various third-party services. Each service must be registered as a plugin in the application and can be used for various tasks, e.g. for publishing resources. If no services are available in the current Kadi4Mat instance, this menu item is hidden.

Publishing data sets

Kadi4Mat offers a particularly convenient way to publish datasets. Datasets can be uploaded and published directly from the system to Zenodo (`https://zenodo.org`), where they are assigned their own DOI, which can be referenced later. As soon as a user account in Kadi4Mat is linked to a corresponding user account at Zenodo, an upload can be made (Figure 5.43). More information on this can be found in the chapter *Data publication*.



November 12, 2020

Sample record

Member

Description including Markdown

(Will be sent as rendered HTML to Zenodo)

Files (220 Bytes)

Name	Size
test.csv md5:0c59aff5c83d49d0610ed8160d4a5cea	46 Bytes
test.json md5:7061253260cfd103db3823caf648cc57	174 Bytes

Citations 0

Show only: □ Literature (0) □ Dataset (0) □ Software (0) □ Unknown (0)
□ Citations to this version

No citations.

0 views 0 downloads

Indexed in OpenAIRE

Publication date: November 12, 2020

DOI: 10.5072/zenodo.696683

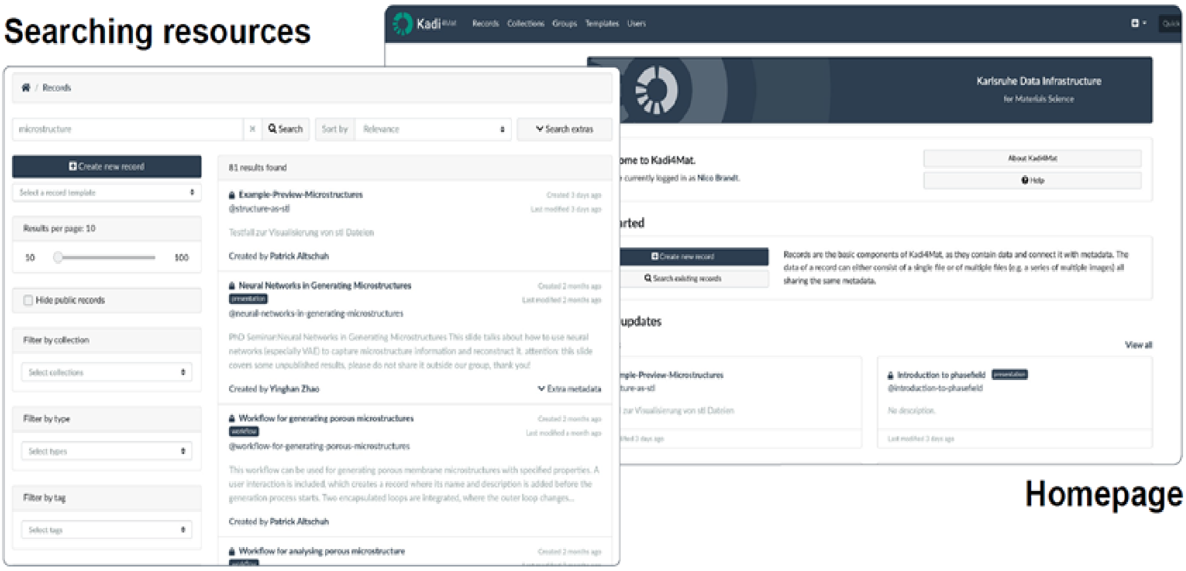
Keyword(s): sample tag

License (for files): Creative Commons Attribution Share Alike 4.0 International

Figure 5.43.: Publication of data sets on Zenodo.

The search functions in Kadi4Mat

Similar to the search functions in eLabFTW, for example, Kadi4Mat also offers extensive options for searching for information and data. The search can be optimised and accelerated by using filters. Figure 5.44 provides a visual impression of how a search can be designed and started directly from the browser.



microstructure

Relevance

Results per page: 10

50 100

81 results found

Example-Preview-Microstructures
@structure-as-stl
Created 3 days ago
Last modified 3 days ago
Testfall zur Visualisierung von stl Dateien
Created by Patrick Altschah

Neural Networks in Generating Microstructures
@neural-networks-in-generating-microstructures
Created 2 months ago
Last modified 2 months ago
PhD Seminar:Neural Networks in Generating Microstructures This slide talks about how to use neural networks (especially VAE) to capture microstructure information and reconstruct it. attention: this slide covers some unpublished results, please do not share it outside our group, thank you!
Created by Yinghan Zhao

Workflow for generating porous microstructures
@workflow-for-generating-porous-microstructures
Created 2 months ago
Last modified a month ago
This workflow can be used for generating porous membrane microstructures with specified properties. A user interaction is included, which creates a record where its name and description is added before the generation process starts. Two encapsulated loops are integrated, where the outer loop changes...
Created by Patrick Altschah

Workflow for analysing porous microstructure
Created 2 months ago

ome to Kadi4Mat.
currently logged in as Nico Brandt.

rted

Records are the basic components of Kadi4Mat, as they contain data and connect it with metadata. The data of a record can either consist of a single file or of multiple files (e.g. a series of multiple images) all sharing the same metadata.

updates

mple-Preview-Microstructures
ture-as-stl
zur Visualisierung von stl Dateien
ified 3 days ago

Introduction to phasefield
@introduction-to-phasefield
No description.
Last modified 3 days ago

Figure 5.44.: The search functions of Kadi4Mat.

6. Data quality

In the previous chapters, we demonstrated how research data can be systematically collected. Now we need to ensure data quality. This point is very important for the further use of research data, whether for the development of follow-up products or for analysis using AI. In general, "*garbage in, garbage out*" applies, i.e. the quality of the data input determines the quality of the result.
When analysing data, we must distinguish between systematic and statistical errors in measurement data or research data. Systematic errors can arise, for example, from inaccuracies in a measuring device or from operating errors. Ideally, such systematic errors can be identified by so-called outliers (Outlier) in the overall data set. The authors of this handbook are currently working as part of the NFDI4ING consortium on an AI-supported, intrinsic data analysis that runs automatically and directly in an electronic laboratory notebook without the need for user intervention. The aim is to automatically detect outliers in measurement data, whereby the user is informed about these outliers and can decide how to proceed with such data. As this work is currently still in development, we will only be able to include a corresponding chapter in a subsequent edition of this manual.

Data analysis can be very extensive and includes, among other things, the mathematical fields of linear algebra, statistics and probability theory. Even though there are large and established software packages (commercial or open source) on the market today that can do a lot of the work for you, it is still important to understand how such analyses work and how the results are obtained. For those who wish to delve deeper into the subject matter, we recommend the following two textbooks by Thomas Nield [**Nield**] and the standard work by Lothar Papula [**Papula**] as comprehensible introductions for self-study.

Commercial software for the analysis of research data includes, for example, the established programme *Origin*® (`https://www.originlab.com/`), or the much cheaper alternative with similar features, the program *QtiPlot*© (`https://www.qtiplot.com/`). The latter has the advantage that it is not only available for Windows, but also for other operating systems such as Linux. These two programmes are professional alternatives to the established MS Excel, with significantly more options for data analysis. However, there are also excellent free programs for data analysis. We would like to mention just two examples here, namely the two statistics packages *R* (`https://www.r-project.org/`) and *Gretl* (`https://gretl.sourceforge.net/`). We will use *Gretl* to demonstrate a simple example of data analysis below and also recommend this programme for those who are just starting out in statistical data analysis. *R* offers even more extensive options, including additional modules.

6.1. Measurement data and its errors

For a simple example, let us assume an arbitrary, normally distribution of data points $x_1, x_2, x_3, ..., x_n$, i.e. we consider n independent measurements with the same accuracy. This assumes that all data points have been determined using the same measurement method, the same measuring instrument and by the same observer. The mean value $\overline{x}$ of such a measurement series is calculated as follows:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} \tag{6.1}$$

The standard deviation of a single measurement is then calculated as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}, (n \geq 2) \tag{6.2}$$

This gives the standard deviation of the mean value:

$$s_{\overline{x}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n(n-1)}}, (n \geq 2) \tag{6.3}$$

A measurement result x must therefore be formulated correctly as follows:

$$x = \overline{x} \pm \Delta x = \overline{x} \pm t \, \frac{s}{\sqrt{n}} \tag{6.4}$$

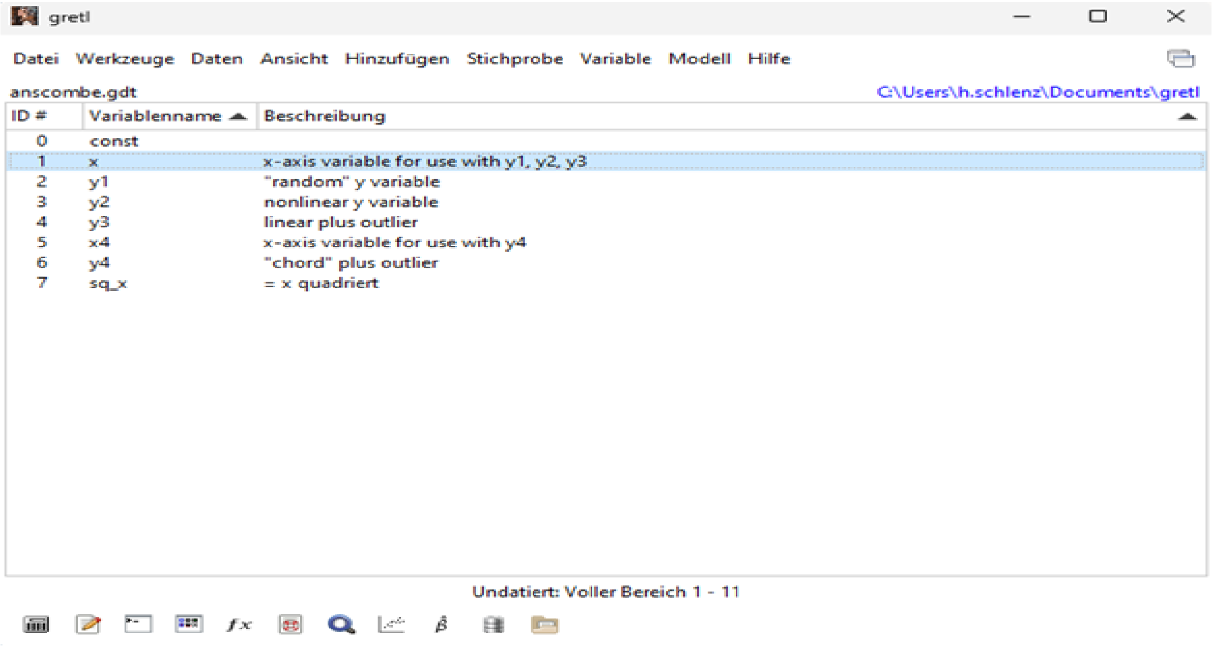
The number factor t depends on the selected confidence level γ (e.g. $\gamma = 95\%$) and the number n of individual measurements. The following table contains some values for t. More detailed tables can be found in the literature cited above:

Table 6.1.: Values for t, depending on the number n of measurements and the selected confidence level γ.

n	$\gamma = 68.3\%$	$\gamma = 90\%$	$\gamma = 95\%$	$\gamma = 99\%$
2	1.84	6.31	12.71	63.66
10	1.06	1.83	2.26	3.25
50	1.01	1.68	2.01	2.68
100	1.00	1.66	1.98	2.63

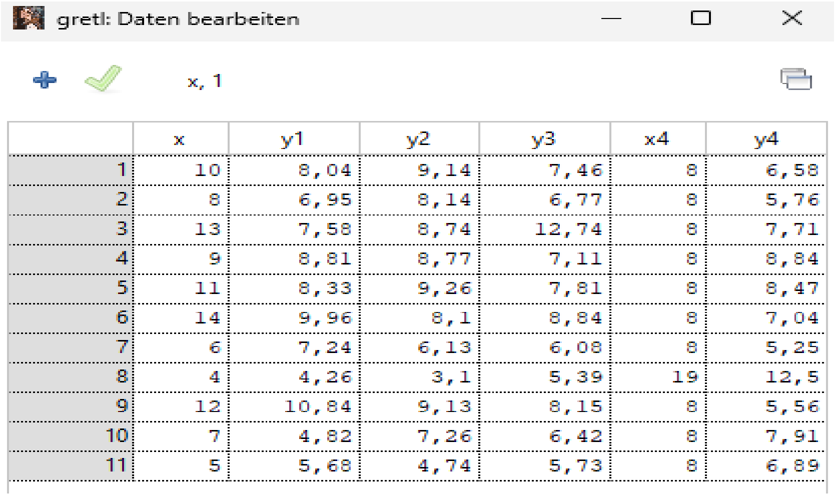
6.2. Data analysis and data visualisation

In this section, we would like to demonstrate how a simple, artificially generated data set (anscombe.gdt) can be analysed quickly using the statistical software *Gretl* (`https://gretl.sourceforge.net/`) mentioned above. This data set is available as an example data set in the current version of *Gretl 2025b* (July 2025), and we would like to encourage all readers to reproduce our analysis with this data set and to perform more refined and complex analyses to familiarise themselves with the capabilities of the programme. This will also highlight how closely linked numerical and graphical data analysis are and how they should complement each other. The human eye and brain can usually detect graphical inconsistencies very quickly, which can greatly speed up data analysis. Outliers can often be detected visually. Figure 6.1 shows the start screen of *Gretl*. In this menu, we have already selected our sample data set *anscombe.gdt* and thus get a first impression of the data structure. In the following figure 6.2, we can explicitly see the data set with all values. Different y values (y_1 to y_3)


gretl
Datei Werkzeuge Daten Ansicht Hinzufügen Stichprobe Variable Modell Hilfe
anscombe.gdt
C:\Users\h.schlenz\Documents\gretl
Undatiert: Voller Bereich 1 - 11

ID #	Variablenname	Beschreibung
0	const	
1	x	x-axis variable for use with y1, y2, y3
2	y1	"random" y variable
3	y2	nonlinear y variable
4	y3	linear plus outlier
5	x4	x-axis variable for use with y4
6	y4	"chord" plus outlier
7	sq_x	= x quadriert

Figure 6.1.: The interface of the *Gretl* software.



	x	y1	y2	y3	x4	y4
1	10	8,04	9,14	7,46	8	6,58
2	8	6,95	8,14	6,77	8	5,76
3	13	7,58	8,74	12,74	8	7,71
4	9	8,81	8,77	7,11	8	8,84
5	11	8,33	9,26	7,81	8	8,47
6	14	9,96	8,1	8,84	8	7,04
7	6	7,24	6,13	6,08	8	5,25
8	4	4,26	3,1	5,39	19	12,5
9	12	10,84	9,13	8,15	8	5,56
10	7	4,82	7,26	6,42	8	7,91
11	5	5,68	4,74	5,73	8	6,89

Figure 6.2.: The data set named anscombe.gdt.

can be assigned to the variable x. In addition, there is a column with squared x values (x_4) and a corresponding y_4 value. For our example analysis, we select the value pairs (x,y_1). To analyse the data, we want to calculate a smoothing curve. The simplest form of a *equalisation curve* is a Regression line. A straight line with the formula $y = a\,x + b$ fits perfectly to measurement points $P_i = (x_i, y_i)$, with $i = 1, 2, ..., n$ and $n \geq 3$. The slope a of this straight line (the regression coefficient) and the intercept b can be calculated as follows:

$$a = \frac{n \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\Delta} \tag{6.5}$$

$$b = \frac{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{\Delta} \tag{6.6}$$

$$\Delta = n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 \tag{6.7}$$

In order to assess the accuracy of the fit, the *correlation coefficient* r is usually calculated. The n measurement points always lie almost on a straight line if r differs only slightly from -1 or +1. In the case of |r| = 1, the measurement points lie exactly on a straight line. However, in linear regression, it is not the correlation coefficient r that is often used, but its square, the so-called *coefficient of determination* R^2 (which is then mistakenly referred to as the correlation coefficient). R^2 indicates how well a linear model fits the observed data. In other words, in our simple example, the regression line $y = a\,x + b$ should be our first mathematical model for describing the data, whereby the slope a and the intercept b must also be determined numerically for the sake of completeness.

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2)(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2)}}, (-1 \leq r \leq 1) \tag{6.8}$$

The following figure 6.3 shows the linear fit with *Gretl*. By clicking on the eighth icon from the left in the lower toolbar of *Gretl* (Figure 6.1), you can access the graphics menu and simply select the variables x and y_1 for the display. The linear regression is preset and you will immediately see the graph shown. For the slope a, we obtain the value 0.5, and for the y-intercept b, the value is 3.0. Using the graphics menu, we can also obtain the value for R^2 of 0.6665. From this value and also from visual observation, it is clear that the fit with this simple model is not optimal. Therefore, we will next try a quadratic fit $y = a + bx + cx^2$. We see the result in Figure 6.4. Now we obtain a value of $R^2 = 0.6873$, which is a slight improvement towards 1, with $a = 0.755$, $b = 1.07$ and $c = -0.0316$. This fit is also not yet optimal. Try it yourself and see if a cubic fit or another function provides a better model for describing the data set. Finally, it is also possible to use *Gretl* to generate a correlation matrix (also known as a heat map) for the entire data set in order to obtain an overview of all existing positive and negative correlations. In Figure 6.5, we see a strong, positive correlation between x and y_1 with a value of 0.8, i.e. when the value of x increases, we also see a corresponding positive increase in the value of y_1, which can be easily understood using Figure 6.3, for example.

With this simple example of data analysis, we would like to illustrate the basic principles. Of course, the analysis of large data sets can be much more complex and extensive, but the principle always remains the same. The models required to describe large and complex data sets require methods of probability theory, descriptive and inferential statistics, linear algebra, logistic regression and classification, or even the use of AI methods such as neural networks. However, a description of these methods would go far beyond the scope of this handbook, and we therefore refer you to the literature

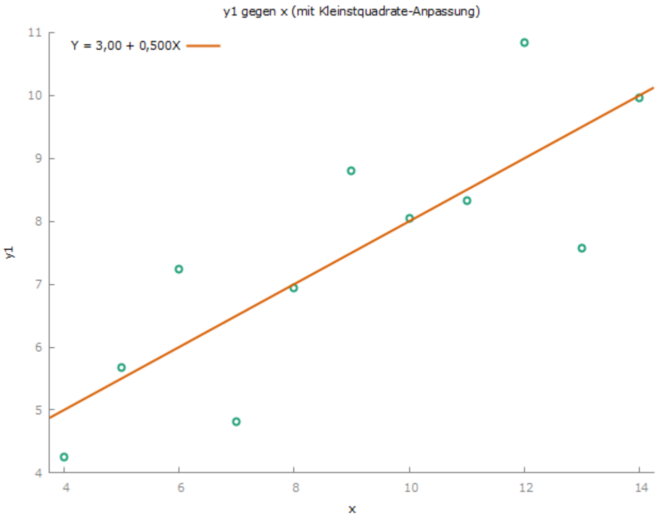


Figure 6.3.: Linear regression for the data points P(x, y_1).

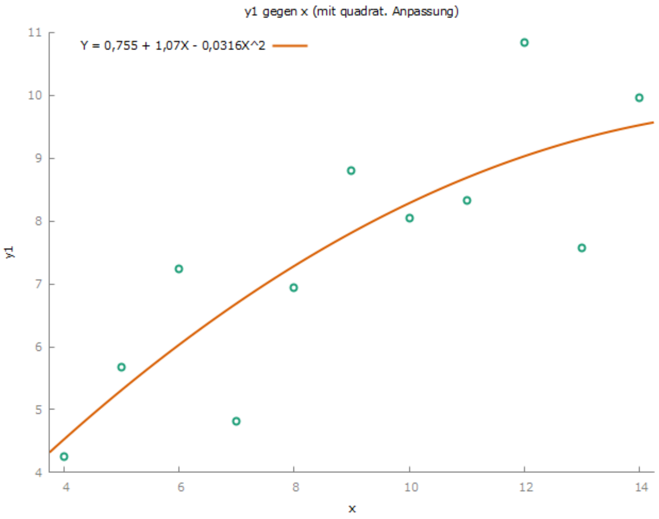


Figure 6.4.: Quadratic regression for the data points P(x, y_1).

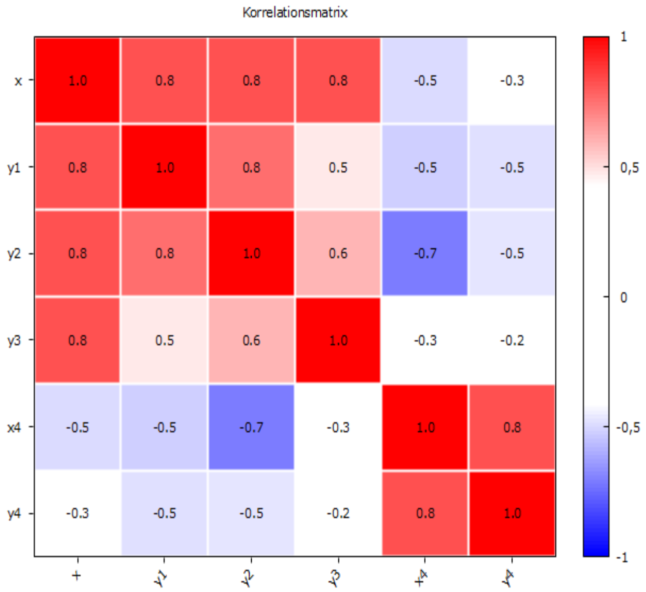


Figure 6.5.: Correlation matrix (heat map) for the entire data set.

listed in the appendix. We will only discuss them in more detail in the chapter on *Machine learning (AI)*.

There are a variety of commercial and free programs available for the visualisation of research data. *Gretl* uses the free program *Gnuplot* (`http://www.gnuplot.info/`) internally, which is also available as a standalone program for almost every operating system. Other free alternatives include *R* (`https://www.r-project.org/`), the very user-friendly *Veusz* (`https://veusz.github.io/`), *Labplot* (`https://labplot.org/`), or, for users with programming skills, *Python* libraries such as *Matplotlib*. There is also no shortage of commercial systems on the market, and the two representatives *Origin*® (`https://www.originlab.com/`) and *QtiPlot*© (`https://www.qtiplot.com/`) have already been mentioned above.

Computer algebra systems like *Mathematica* (`https://www.wolfram.com/mathematica/`), *Matlab* (`https://de.mathworks.com/?s_tid=gn_logo`), as well as *Maple* (`https://www.maplesoft.com/products/maple/index.aspx`) and others can also be used for data analysis and visualisation, depending on your budget and personal preferences. However, regardless of the software used, it is always important that you get a picture (in the truest sense of the word) of your data in order to be able to evaluate it in the best possible way.

7. Data exchange and data tracking

This chapter is aimed at advanced users who want to delve deeper into research data management. The content of the previous chapters is required for understanding. Here we would like to describe how research data can be exchanged between different electronic laboratory notebooks, even between different ELNs around the world. Another focus will be on data tracking (also known as provenance tracking), i.e. the ideally seamless tracking of individual process steps, for example in materials research or process engineering. However, other areas of application are also conceivable here. Data tracking can take place within an instance of an ELN, but also between different ELNs that can communicate with each other.

7.1. Data exchange between electronic laboratory notebooks with SciMesh

We have chosen the relatively new system SciMesh (`https://scimesh.org`) for two reasons to illustrate the processes of data exchange and data tracking: 1. The authors of this manual are (in part) also the developers of SciMesh; 2. SciMesh seems to us to be an ideal solution for effectively solving the above-mentioned tasks. However, for the sake of completeness, we will also briefly discuss alternative solutions in this chapter.

Introduction to SciMesh

SciMesh is a set of specifications that define the representation of scientific results in the form of a knowledge graph. While many such data formats focus on simulation data, SciMesh explicitly includes physical instances as first-class citizens. In this way, the origin of both data and instances can be documented. The immediate purpose of SciMesh is to represent the content of electronic laboratory notebooks (ELNs) for the exchange of scientific results between ELN instances. This allows collaboration partners who use different ELNs (even different ELN software) to obtain a unified view of their shared results without media discontinuity.

SciMesh represents scientific knowledge as a knowledge graph. In its current implementation, it focuses on sample-based workflows in which samples (physical samples or data artefacts) undergo a sequence of processing and measurement steps. However, it is not limited to this. In general, it represents scientific knowledge by explaining a cause-and-effect relationship. In other words, if certain conditions are met, certain observations are made. Figure 7.1 illustrates this. Starting from an initial state *zero* (*nil*), the processes change this state. From left to right, a monotonic increase over time can be observed. Consequently, this time axis is also contained in a chain of processes. A process can be: *take silicon substrate from the shelf*, *heat sample*, *mix two substances together* or *wait for solar eclipse*. It can be kept this general. Each process has one or more causes. If there is only one, it can be the initial state. The initial state is completely empty. It contains no information whatsoever, which is why the chain of processes must define the intermediate states as completely as necessary in order

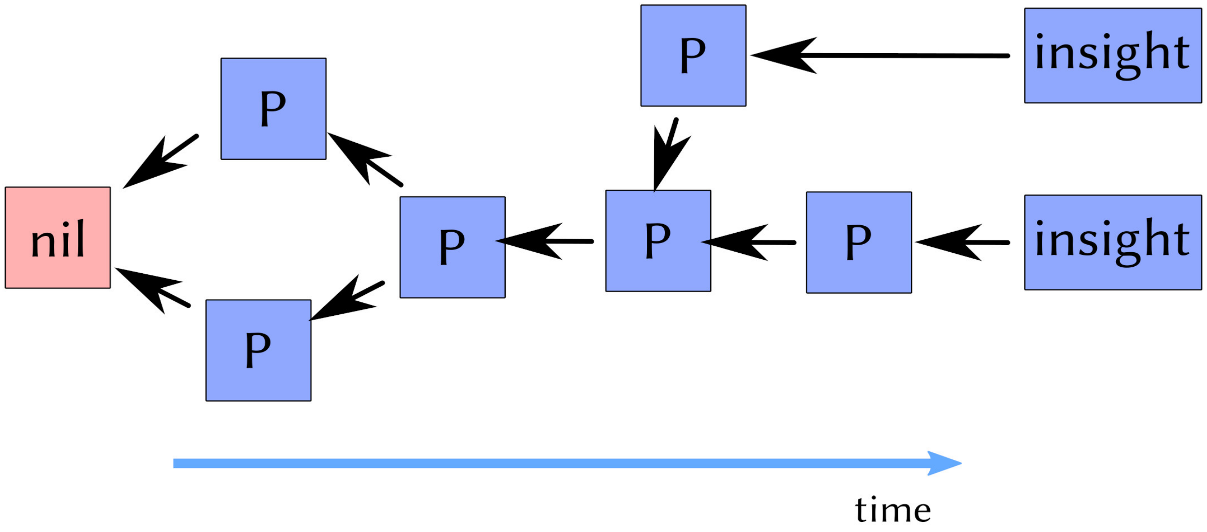


Figure 7.1.: Example graph of a process in SciMesh.

to be useful for scientific conclusions. Conversely, one process can be the cause of several others. In this way, chains can be branched off, possibly by different scientists years after working on the main trunk. Figure 7.2 shows how the essence of scientific work, the relationship between effect and cause, is represented in the graph: The processes up to a certain point are the cause, and the observations at that point are the effect. It is therefore valid to refer to this point as "knowledge," which represents a separate node in the graph.

Two things are important here. First, the process to which the observation data is linked is a measurement, and often a measurement does not change the state. Nevertheless, it is in all respects a process in the true sense of the word. We do not think it is necessary to distinguish between measurements and processes that actually change something. Furthermore, a measurement changes the state, even if the change may not be significant. And second, a particular graph of processes can contain many insights that point to it. In particular, the processes that follow a measurement can lead to a second measurement with new results, i.e., to a new insight. Speaking of insights: Figure 7.3 shows the

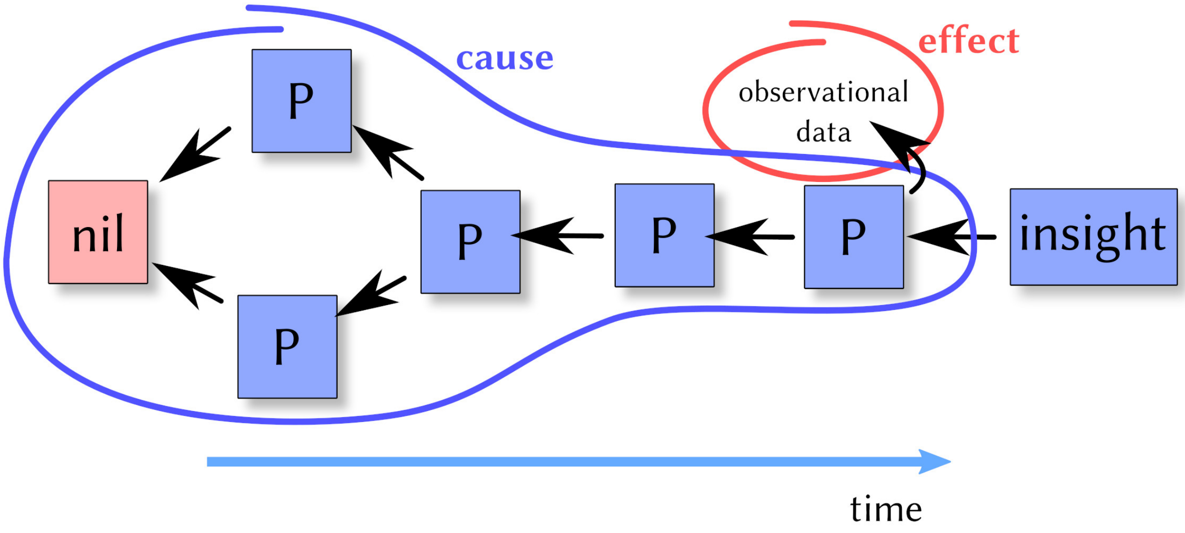


Figure 7.2.: Example graph of a process showing cause and effect.

relationships between things that can indicate a certain state (also called a process) in a diagram. All of these things are insights, but if they are a chain of concrete processes at specific points in time, this should be called an experiment. If all of this happened with the same sample, the experiment can be identified with this sample. If the processes are not concrete processes but general designs of processes, the experiment is actually a recipe for experiments. Or it is a scientific hypothesis that brings cause and effect together. The following two examples illustrate experiment and hypothesis:

1. On 21 October 2019, John threw a ball from the Eiffel Tower, which moved downwards.
2. A body is released in a gravitational field and moves according to the force field.

The currently specified SciMesh RDF data model is more narrowly defined than the very general concept outlined above. The reason for this is very simple: our work is at an early stage and we need to focus on specific areas of research in order to make the best use of our resources. Since we are working in the *Task Area Caden* of NFDI4ING, the research area of our choice is the sample-based workflow. We hope that in the near future we will also be able to provide guidelines for extending SciMesh to other areas.

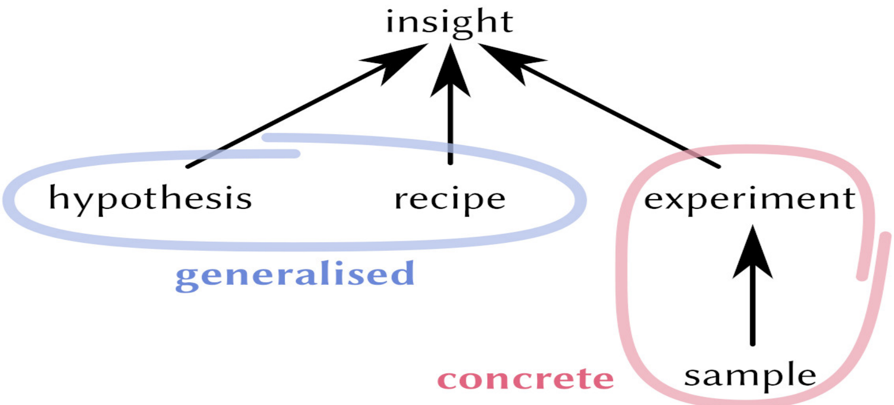


Figure 7.3.: Is-a relationships of insight-like entity classes.

The data model

The RDF data model is based on two concepts: patterns and processes. Both are to be understood in the broadest sense. A sample can represent both the state of a physical instance and a data record. A process can create a new sample state (i.e., change the sample, e.g., an etching process) or generate new data (e.g., a measurement on a sample) or both. Figure 7.4 provides an overview of the anatomy of a knowledge graph in SciMesh. It is a very simple graph, but it contains most of the basic concepts. At its core, it consists of a sequence of processes (below) that work on the sample. The first process, *substrate*, creates the sample (the sample begins its life as a blank substrate). Its RDF properties determine the basic physical properties of the sample (e.g., material and size). Subsequently, further layers of material are applied to the substrate in a deposition process. Common properties of most processes are name, method, timestamp, operator, and comments. Please note that three different namespace domains are involved: 1. The domain of the ELN instance: the processes, the process types, the sample and its intermediate instances. This is shown in blue. 2. The domain of the ELN software: the sample type. This is shown in green. 3. External domains such as BFO/OBO and RDF. This area is shown in red. A more detailed description of the data model can be found at `https://scimesh.org`.

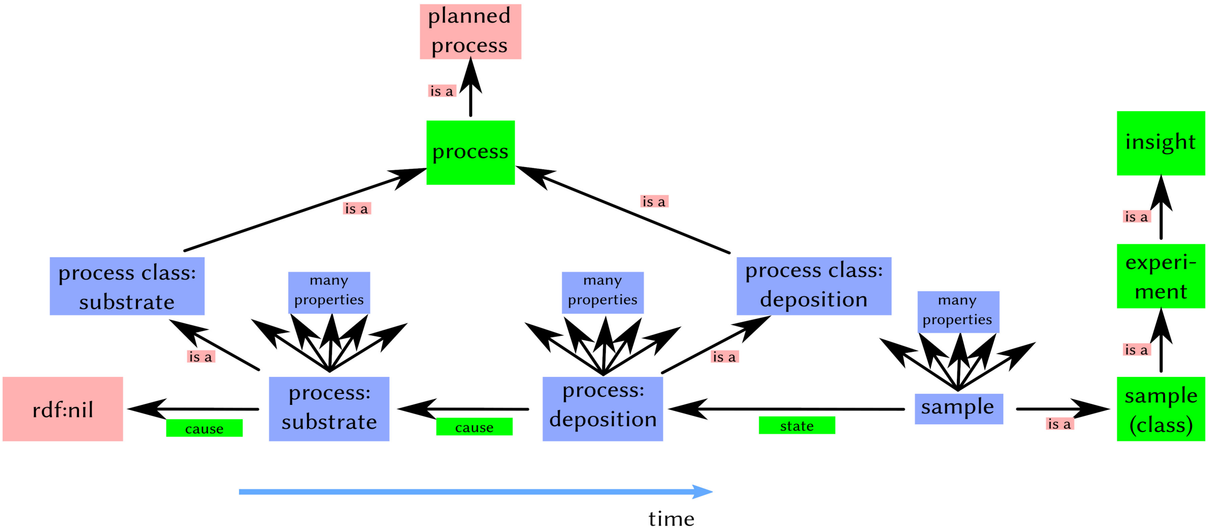


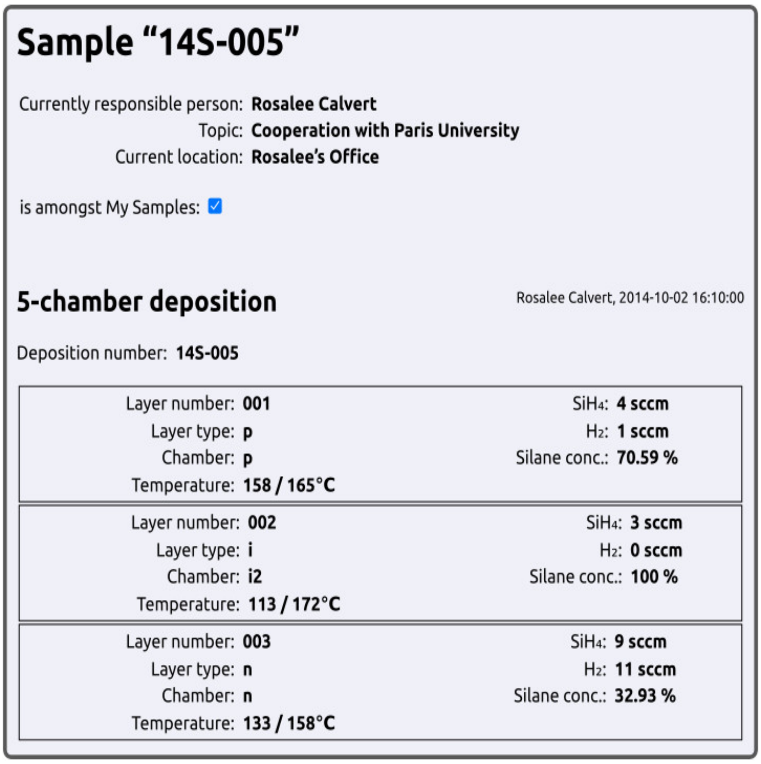
Figure 7.4.: Simplified example topology of a SciMesh knowledge graph. Colours denote namespaces: blue is the ELN namespace, green is the SciMesh namespace, and red are external namespaces (RDF, OWL, OBO, etc.).

An example: JuliaBase as a prototype

We have already discussed the ELN JuliaBase in more detail in Chapter 5.3. In the following, the application of SciMesh will be demonstrated using an example from JuliaBase. JuliaBase is a Python/Django framework for creating ELNs or similar databases with a high degree of customisability. Since it implements a process- or sample-based workflow in a highly structured manner, it is a good candidate for prototyping SciMesh. Figure 7.5 shows a simple sample data sheet. In chronological order, you can see what was done with the sample. In this case, only one thing: the *5-chamber coating* is the only experiment here. It consists of three silicon layers deposited on the substrate, each with its own configuration (temperature, gas flow rates). Although this sample data sheet is so simple, its RDF representation is quite complex, see Figure 7.6. This RDF representation is in Turtle format, which is human-readable (at least with some experience). After the namespace prefixes (the lines beginning with @prefix), which are only used to abbreviate common prefixes with very short names, you see the example with its properties living in the jb-s namespace. The property cause refers to the last process that was performed with this example. This process is the only one at the same time (sm:cause: ()). It is the deposition process. It ends with the empty line. The instance-specific entities (i.e. the things that are specific to the respective institute, such as the test methods) are located in the namespace ns1. What follows is the first layer with its data. It is linked to its repository via the property jb:is Subprocess. The data of the second layer and the complete third layer are omitted here for clarity. The prototype implementation is synchronised with the SciMesh website. It is created against the JuliaBase software in its graph branch. There is also a short guide on how to obtain RDF data from a JuliaBase test instance.

7.2. Data tracking

Data processes convert input data into output data. They can include simulations, but also data conversion, evaluation, aggregation and visualisation. They should be atomic, i.e. not consist of sub-



Sample "14S-005"

Currently responsible person: Rosalee Calvert
Topic: Cooperation with Paris University
Current location: Rosalee's Office

is amongst My Samples: ☑

5-chamber deposition

Rosalee Calvert, 2014-10-02 16:10:00

Deposition number: 14S-005

Layer number: 001
Layer type: p
Chamber: p
Temperature: 158 / 165°C
SiH_4: 4 sccm
H_2: 1 sccm
Silane conc.: 70.59 %

Layer number: 002
Layer type: i
Chamber: i2
Temperature: 113 / 172°C
SiH_4: 3 sccm
H_2: 0 sccm
Silane conc.: 100 %

Layer number: 003
Layer type: n
Chamber: n
Temperature: 133 / 158°C
SiH_4: 9 sccm
H_2: 11 sccm
Silane conc.: 32.93 %

Figure 7.5.: Data sheet for sample *14S-005* as displayed by the browser in a JuliaBase instance..

processes, but this is not a prerequisite. In SciMesh, data processes are of the class *Process*, just like experimental processes. They can be chained with *cause* relations, i.e. a data process has as potential input all data produced by its predecessors.

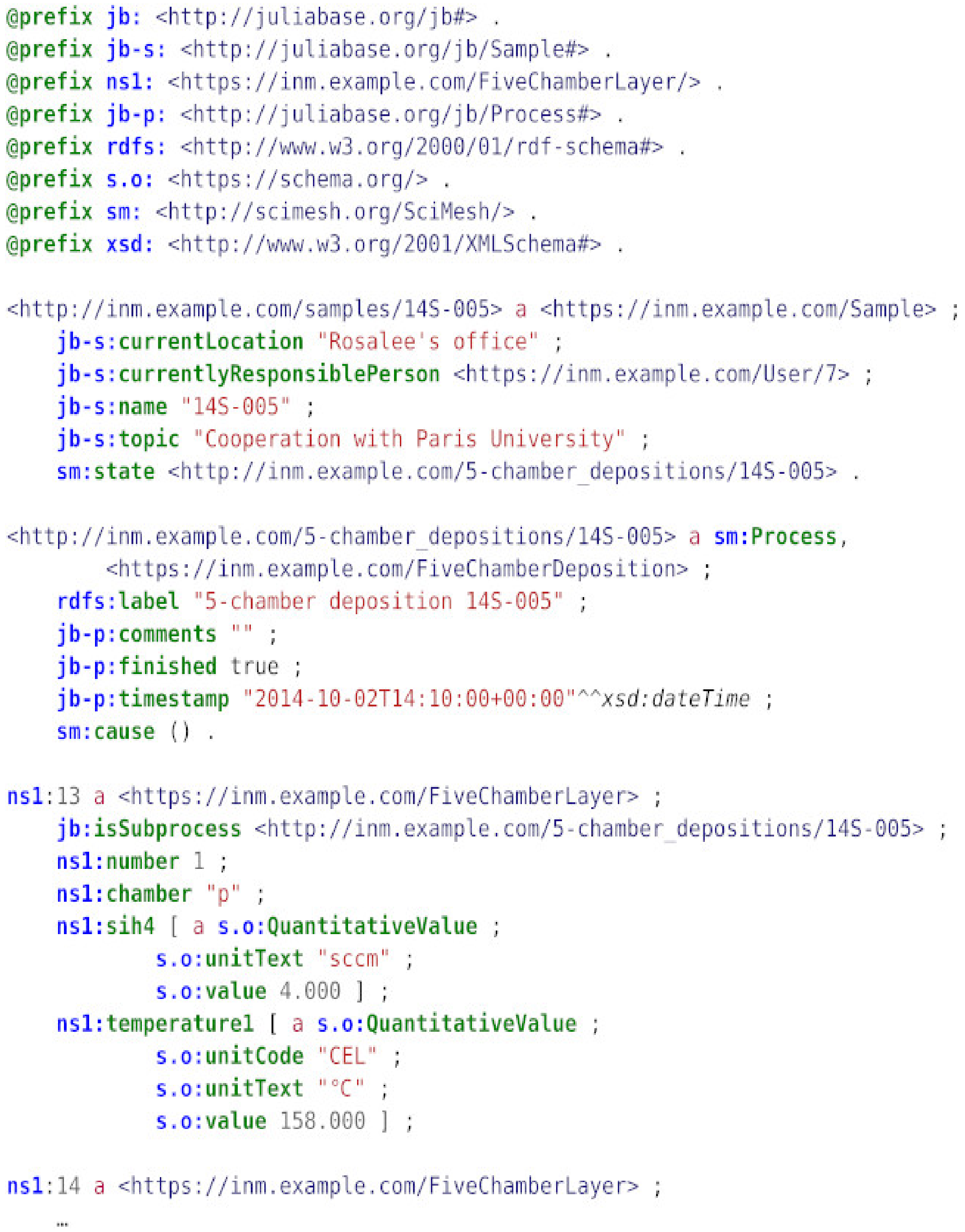
Mass data: By *mass data*, we mean an opaque octet stream of data under a specific URL. In order to be referenced in a SciMesh graph, the web server's response must contain a correct content type. In addition, the URL must contain the checksum of the data. If the protocol scheme itself does not provide for this (e.g., IPFS URLs), the URL fragment (the part after the #) must contain a hash using multiformats. The format is as follows:

<base>base(<version><multihash>)

In other words, the binary <multihash> is encoded by the function *base()* (e.g. base32), and the character <base> that denotes this function (*b* in the case of base32) is prefixed. <version> is always the byte 0x01.

Data input: To see exactly which data is used, you need to delve deeper into the process (e.g. by looking at the input data in the processing programme). In SciMesh, the URLs to bulk input data are not explicit. (Of course, you can make them explicit with your own vocabulary.) Analogous to physical samples, the input is the entire graph of processes (and in particular their data outputs) that led to this process. Technically, the programme that performs the data processing can download all input data, but a valid SciMesh graph ensures that all this data was output by a previous process. Violating this rule means that not all parameters that influence the sample are included in a physical process. In some cases, this may mean that you have to create a previous process just to connect it to mass output URLs. Just do it, it's fine.

Data output: All output data is represented by URIs, which are resolved to retrievable URLs con-



```
@prefix jb: <http://juliabase.org/jb#> .
@prefix jb-s: <http://juliabase.org/jb/Sample#> .
@prefix ns1: <https://inm.example.com/FiveChamberLayer/> .
@prefix jb-p: <http://juliabase.org/jb/Process#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix s.o: <https://schema.org/> .
@prefix sm: <http://scimesh.org/SciMesh/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://inm.example.com/samples/14S-005> a <https://inm.example.com/Sample> ;
    jb-s:currentLocation "Rosalee's office" ;
    jb-s:currentlyResponsiblePerson <https://inm.example.com/User/7> ;
    jb-s:name "14S-005" ;
    jb-s:topic "Cooperation with Paris University" ;
    sm:state <http://inm.example.com/5-chamber_depositions/14S-005> .

<http://inm.example.com/5-chamber_depositions/14S-005> a sm:Process,
        <https://inm.example.com/FiveChamberDeposition> ;
    rdfs:label "5-chamber deposition 14S-005" ;
    jb-p:comments "" ;
    jb-p:finished true ;
    jb-p:timestamp "2014-10-02T14:10:00+00:00"^^xsd:dateTime ;
    sm:cause () .

ns1:13 a <https://inm.example.com/FiveChamberLayer> ;
    jb:isSubprocess <http://inm.example.com/5-chamber_depositions/14S-005> ;
    ns1:number 1 ;
    ns1:chamber "p" ;
    ns1:sih4 [ a s.o:QuantitativeValue ;
            s.o:unitText "sccm" ;
            s.o:value 4.000 ] ;
    ns1:temperature1 [ a s.o:QuantitativeValue ;
            s.o:unitCode "CEL" ;
            s.o:unitText "°C" ;
            s.o:value 158.000 ] ;

ns1:14 a <https://inm.example.com/FiveChamberLayer> ;
    ...
```

Figure 7.6.: Turtle representation of sample *14S-005*.

This uses an image by Vectorportal.com, licensed under CC BY

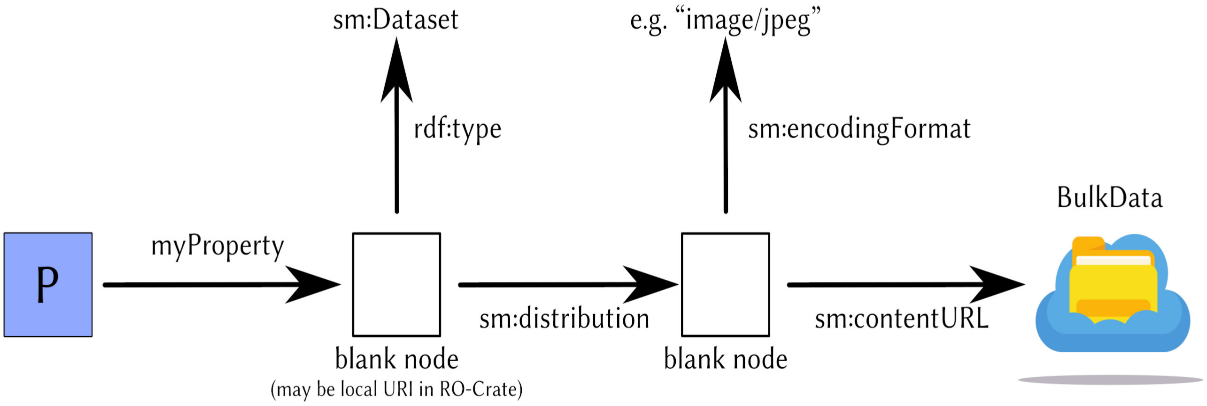


Figure 7.7.: Representation of mass output data in SciMesh. Here, *sm* is the namespace *http://schema.org/*.

taining this data, which are linked to the process using a user-defined vocabulary (as is the case with measurement data for experimental processes). The process must be the subject of such triples (see Figure 7.7).

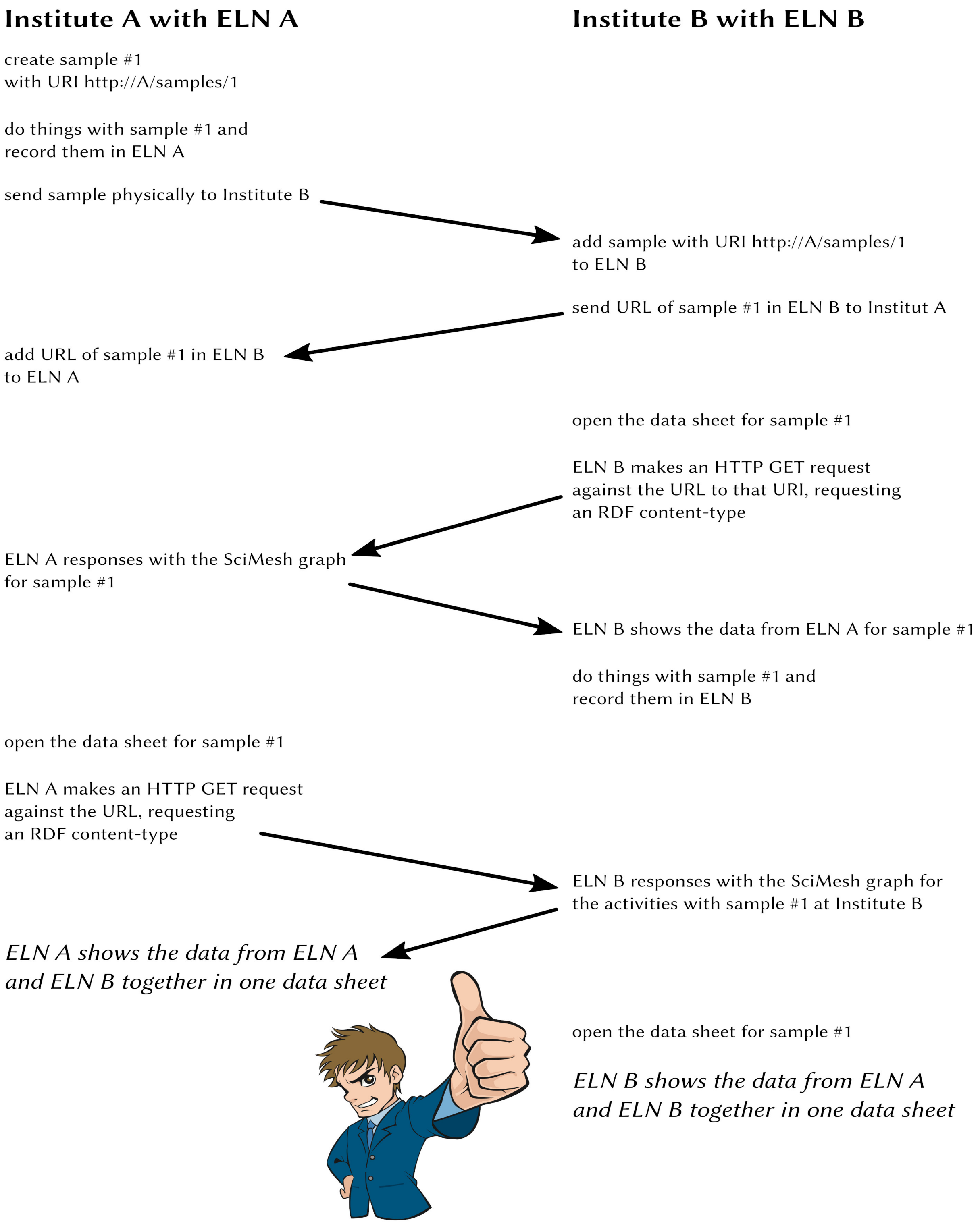
7.3. Implementation of SciMesh

The following describes best practices and normative requirements for implementing SciMesh in an institution, particularly in an ELN or sample database.

The workflow: Figure 7.8 shows the bouncing of samples and data between two institutes that are working together. It concerns the experimental activities with sample No. 1, which is created in institute A but also examined in institute B. It is necessary to be able to view all data in both institutes. The text-heavy figure contains most of the details. We will therefore only make a few comments below. Both institutes have their own ELN. These two ELNs are not only different instances on different computers, but may also be different software. It is important to note that Institute A is the *actual* home of the sample, as the URI of the sample is actually a URL in the domain of Institute A. However, there is a URL (not URI) for sample No. 1 at Institute B. Anyone who wants to collect all data for sample No. 1 must query both URLs. It is possible that instances may provide data that was also found at other instances, but this is not guaranteed. Institute A maintains a list of all URLs that also have data for sample No. 1 and exports it as part of the SciMesh graph for sample No. 1. Two aspects are not covered at all in this diagram: caching and permissions. While the former is an optional but important optimisation, the latter is essential.

7.4. Obtaining the graph

A major challenge in SciMesh is the fact that the data from a sample or an insight is generally scattered across many instances, possibly in different institutions and countries. The two most important things to keep in mind here are:



Institute A with ELN A

Institute B with ELN B

create sample #1
with URI http://A/samples/1

do things with sample #1 and
record them in ELN A

send sample physically to Institute B

add sample with URI http://A/samples/1
to ELN B

send URL of sample #1 in ELN B to Institut A

add URL of sample #1 in ELN B
to ELN A

open the data sheet for sample #1

ELN B makes an HTTP GET request
against the URL to that URI, requesting
an RDF content-type

ELN A responses with the SciMesh graph
for sample #1

ELN B shows the data from ELN A for sample #1

do things with sample #1 and
record them in ELN B

open the data sheet for sample #1

ELN A makes an HTTP GET request
against the URL, requesting
an RDF content-type

ELN B responses with the SciMesh graph for
the activities with sample #1 at Institute B

*ELN A shows the data from ELN A
and ELN B together in one data sheet*

open the data sheet for sample #1

*ELN B shows the data from ELN A
and ELN B together in one data sheet*

Figure 7.8.: Possible workflow for two institutes working with the same samples.

1. All URIs of samples and processes are constant. They never change. 2. All of these URIs are also URLs at the same time. So, if an ELN wants to display all data for a specific sample, it first performs an HTTP GET with the sample URI. This determines the entity of the sample and possibly some process entities. The ELN then traverses the process graph backwards in time. Whenever it encounters a process with a missing cause, it performs an HTTP GET against its process URI. This results in a new graph, which is merged with the existing one. The traversal is then continued. Eventually, there are no more causes to search for (all remaining cause fields contain rdf:nil), or their URLs cannot be retrieved (because the servers do not respond or we do not have the necessary permissions). The diagram is then displayed to the user.

Requirements for ELNs: Participating databases or ELNs must implement the following: 1. Each process and its process history (i.e., the graph back in time) must be the response to an HTTP GET to that process URI. External processes (i.e., with URIs under the control of other systems) do not need to be included. 2. An HTTP GET to the sample URI must return the sample entity, the processes it references in the *state* properties, and the entire process graph. Again, external processes do not need to be included. 3. An HTTP POST to the sample URI with a JSON payload of the form:

state: ["http://example.com/processes/1","http://example.com/processes/2"]

adds the process URIs to the example, i.e. *status* properties with these URIs as objects are added. Note that all these requests – including the POST – can be answered with HTTP 30x codes and must be repeated with the new URL.
For further information on the visualisation of processes and graphs, good URIs and authentication, please refer to the SciMesh website.

7.5. MetaData4Ing

Under certain circumstances, MetaData4Ing can be an alternative to the SciMesh system mentioned above, at least as far as data tracking is concerned. MetaData4Ing is an ontology for describing the creation of research data in the context of scientific activity. The target audience for MetaData4Ing (m4i) is not so much users in research data management, but rather IT experts such as application developers and software engineers. For the sake of completeness, we would therefore like to provide only a brief overview of the project at this point. Further information can be found at `https://nfdi4ing.pages.rwth-aachen.de/metadata4ing/metadata4ing/`.

The Ontology m4i provides a framework for the semantic description of research data and the entire data generation process, which includes the object of investigation, all sample and data processing methods and tools, the data sets themselves, and the Roles of persons and institutions. The structure and application of the ontology are based on the principles of modularity and inheritance. m4i is a collection of terms (classes and properties) that can be used to enrich a research data set with semantic and machine-readable metadata in order to annotate it or integrate it into a database or a larger knowledge graph. The metadata can be serialised in formats such as JSON-LD, YAML-LD or Turtle.

Why use MetaData4Ing to describe research data?

Metadata contains structured information for a contextual description of data and is, so to speak, data about data. Metadata is needed to find, manage and use data, not only when publishing data, but also in everyday research. In this context, it is important that all information necessary to find and understand the data is expressed in a common and consistent language consisting of unique, well-documented terms. This approach is a prerequisite for FAIRe (meta)data, especially for its interoperability. m4i offers a general process-based model that enables a flexible description of research activities and their results, with a focus on the provenance of both data and material objects. m4i provides a selection of general concepts such as processing steps, inputs and outputs, methods and tools used, which enable information about research processes and results to be modelled in a structured, consistent and machine-processable manner. One of the main advantages of using m4i is that the resulting description is highly interoperable and allows data from very different scientific disciplines to be integrated into a single knowledge graph. Furthermore, m4i makes extensive use of concepts from well-known general or top-level ontologies, such as Basic Formal Ontology (BFO), Data Catalog Vocabulary (DCAT) or PROV Ontology (PROV-O), allowing the information modelled in m4i to be seamlessly embedded in larger contexts. By documenting your research data with m4i, you not only meet the requirements of good scientific practice, but can also use consistent metadata when searching, analysing or otherwise using your data, and benefit from collaboration. You can store RDF metadata as JSON-LD. This format provides semantically enriched information that is understandable to both humans and machines. The availability of machine-readable documentation of your data also facilitates the publication or archiving of your data in data repositories in a citable form.

The general process model

One of the main goals of the Metadata4Ing-Ontology is to enable researchers to document the provenance of data and material objects that are created or modified during research processes. Metadata4Ing achieves this with the help of a generalised process model centred around the Class processing step. The data and material objects mentioned above are described as the output of the processing step. Other relevant information, such as the methods or tools used in a research process, is described in separate classes that can be linked to the processing step. A series of processing steps can be used to map complex research processes. Metadata4Ing can therefore be understood as a system of building blocks that refer to processing steps and, taken together, provide a complete description of the origin of a data set or a material object. Figure 7.9 illustrates the general process model.

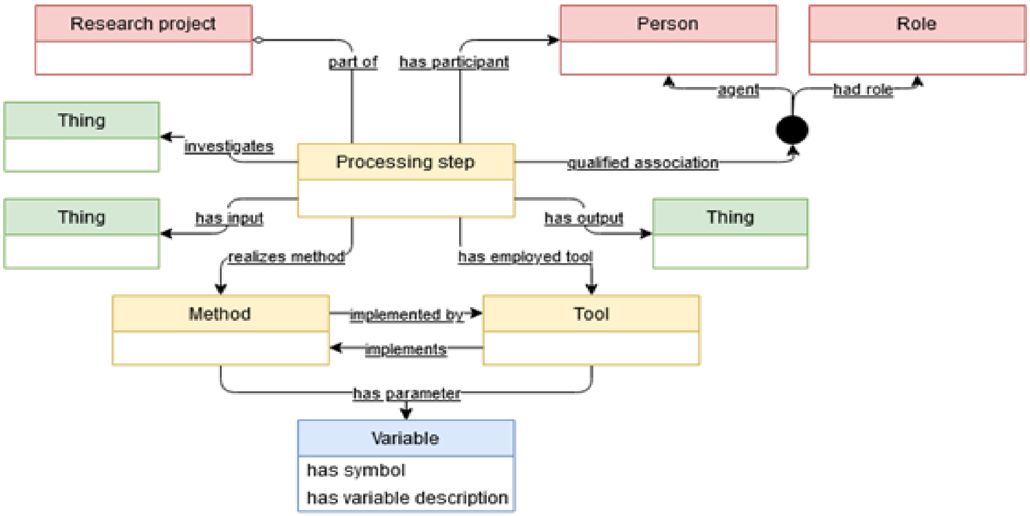


Figure 7.9.: The general process model of MetaData4Ing.

8. Data publication

Publishing data as a supplement to an article has been common practice for a long time. However, new publication channels have emerged in recent years. Data can be published via repositories as standalone data publications. Research data is also frequently exchanged informally among colleagues. It can also be published independently, provided that there are no arguments against this, such as data protection or other usage concepts (see, among other things, the legal framework in Chapter 3). The official publication of data has advantages for both the researchers who make it available and the users. These include:

- No extra effort is required if the research data is requested in the future.
- Data can be cited clearly, just like text publications, which enhances reputation.
- The publication of data increases the transparency and visibility of one's own research.
- The creation of the data is recognised as independent scientific work.
- Publication prevents duplicate data collection and thus unnecessary expenditure of time and money.
- Requirements, e.g. from research funding bodies, publishers and guidelines, are met.

8.1. Publishing data sets

In principle, data sets can be published in data repositories or in data journals. However, even if research data is to be published in a data journal, it is usually also archived in a data repository. Data repositories are online services where digital objects can be archived, documented and published. They fulfill the following functions:

- Securely archive data for the long term.
- Keep data and metadata together.
- Share data (either publicly or only with a limited group of users).
- Enable users to search for data (query and search options).
- Incorporate data from other authors into your own work.
- Data can be referenced permanently using persistent identifiers such as DOIs.

In general, a data repository is an Internet service where data (i.e. digital objects) are uploaded and assigned a permanent identifier (PID) and metadata. The metadata describe the content of the data, how they were created, the software and methods used, legal aspects and terms of use. The data sets can usually also be linked to an associated text publication. Search functions allow the data to be found, viewed and, with the appropriate authorisation, downloaded. The Registry of Research Data Repositories (*re3data*; `https://www.re3data.org/`) offers a very comprehensive and easily filterable overview of data repositories worldwide. *re3data* is very helpful for gaining an overview of which repositories are suitable for your discipline or working group. *re3data* references over 2500 data repositories worldwide and includes both discipline-specific and generic repositories. It offers good filtering options, e.g. by subject area, data types, terms of use or subject-specific metadata standards, and allows users to make a rough assessment of the quality of the individual repositories.

Subject-specific repositories e.g. are the *TIB Hannover* (media-specific data repository; `https://www.tib.eu/de/`), *NoMaD* (subject-specific data repository; novel materials discovery; free data repository for material data; `https://nomad-lab.eu/nomad-lab/`) and *Zenodo* (generic data repository; repository hosted at CERN (`https://home.cern/`) for data sets up to 50 GB in volume; `https://zenodo.org/`) and *JülichDATA* (institutional data repository; Forschungszentrum Jülich; `https://data.fz-juelich.de/`).

Subject-specific data repositories are the central location where researchers can search for data from a specific subject area, which means that these data are highly visible within the professional community. The repositories usually contain subject-specific metadata standards for describing the data and are often equipped with special services, such as tools for searching, analysis and visualisation. If there is such a repository for your subject area, it is advisable to use it. Generic data repositories are open to all research areas and all types of research data. Their metadata schemas are usually universally applicable. Institutional data repositories are usually open to different subjects, similar to general data repositories. JülichDATA is the data repository of Forschungszentrum Jülich and serves as a (bibliographic) reference system for data output. The metadata is assigned a globally unique ID and can be searched and downloaded. The data sets do not have to be published, but can be made available to the public so that they are easy to cite and have a DOI. In addition to **re3data**, other sources are available for searching repositories and research data. With **RIsources** (RI = Research Infrastructure), the DFG offers an information portal for searching research infrastructures. You can search for repositories and other infrastructure services in the catalogue (`https://risources.dfg.de/home_de.html`). **Dataset Search**, a search engine from Google that helps researchers find freely accessible data (`https://datasetsearch.research.google.com/?hl=de`). It complements Google Scholar, the search service for academic studies and reports. **Mendeley Data**, a search engine for research data from the Elsevier publishing house (`https://data.mendeley.com/`). DataCite Commons (**DataCite**), a global search of all publications that have been assigned a DOI. The DataCite consortium is an international association of mostly public institutions that promote access to research data and want to make it available (`https://commons.datacite.org/`). **B2FIND** (EUDAT), a search engine for research data provided by the European Union as part of the European network **EUDAT** (`https://eudat.eu/service-catalogue/b2find`).

8.2. Example: Publishing a dataset on Zenodo

To publish a research dataset on Zenodo, you must first register and log in or log in with your GitHub account. You can then upload your data, add metadata such as title, authors and description, and publish the dataset. Zenodo automatically assigns a DOI (Digital Object Identifier) to your dataset, which makes it permanently identifiable.

1. Registration and login: Visit the Zenodo website and register with your email address or log in with your GitHub account. If possible, connect your ORCID account to Zenodo to confirm your identity and improve the discoverability of your research.

2. Create a new upload: Click on the plus sign in the header and select *New Upload*. Upload your research data either by dragging and dropping or by selecting the files in your file directory. Zenodo supports uploading files, not folders. If you have a folder structure, compress it into a ZIP file and upload the ZIP file.

3. Add metadata: Provide basic information such as title, authors, description, and resource type. Include any relevant information that is important for the discoverability and reuse of your data. Select the appropriate resource type (e.g., Dataset, Publication, Image). If necessary, add an existing DOI or let Zenodo assign a new one.

4. Publication: Carefully review your data and metadata. You can set the visibility of the dataset (public or restricted) and set an embargo if necessary. Click the green *Publish* button to publish your dataset. Please note that a published dataset cannot be changed or deleted.

5. Community: You can add your dataset to one or more Zenodo communities to associate it with a specific discipline or research project.

6. DOI: Zenodo assigns a DOI (Digital Object Identifier) to each uploaded dataset, which is a unique and permanent identifier. This DOI allows your data to be easily cited and referenced in other publications.

7. Testing: You can use the test platform (`https://sandbox.zenodo.org`) to familiarise yourself with Zenodo before uploading your actual research data. Further helpful instructions for publishing research data can be found directly at Zenodo, for example the publication by F. Schmitt (`https://doi.org/10.5281/zenodo.10868941`).

8.3. The reuse of research data

Research data can be reused in various ways by making it publicly available and archiving it in appropriate repositories. Good documentation of the data and its origin is crucial for reuse by others. Assigning unique identifiers (e.g. DOIs) and describing the data with metadata ensures that it is easy to find and cite. It is essential to take into account both the FAIR principles and legal aspects (see chapters 2 and 3). It is important to clarify the copyrights and rights of use for the research data. Licence conditions regulate how the data may be reused. It is also advisable to contact a research data centre or repository at an early stage to clarify any open questions. The reuse of research data can support the scientific work of other researchers and lead to new insights. The publication and archiving of data promotes the transparency and reproducibility of research results (keyword: **good scientific practice**). Reuse can also help to ensure that data does not remain unused and that its full potential can be exploited. When considering reuse, it is important to differentiate it from other terms. By **reuse**, we mean the use of research data/publications by parties other than the original creators, often for new purposes. **Subsequent reuse** means the repeated use of one's own data within the scope of one's own projects. **Recycling** is defined as the use of old data in a modified or processed form. **Secondary use** refers to the use of existing data by other research teams, e.g. in meta-studies or for verification and replication. In the following, we will focus only on reuse in the narrow sense.

The end of a project is an important time to carry out data management activities to prepare for future reuse of data. This is because you still remember all the important details about your data and can make good decisions about how to prepare the data for the future. Data is often stored in a file type that can only be opened with specific, expensive software – this is known as a *proprietary file type*. You can tell that your data is stored in a proprietary file type if you lose access to the data when you lose access to the software. If data is in a proprietary file type, it is always a good idea to copy the data to a more common, open file type as a backup; you may lose some functionality, but it is better to have a backup than no data at all.
To save yourself time in the future searching through all your research files, store the most important files in a separate *Archive* folder. Do this at the end of the project, while you still know which files are important and where they are located. The *Archive* folder should only contain a small subset of the most important documents that are likely to be reused. You will still have to search through all your files, but in most cases you will save time by simply finding what you need in the *Archive* folder. If you use an electronic lab notebook, as described several times in this manual, you can save yourself this effort because you already have all your data structured and can export your data to a wide variety of file formats (see Chapter 5).

Researchers regularly leave research institutions to take up new positions elsewhere. As this happens frequently, it represents a critical transition during which data can be lost. The use of an electronic lab notebook, including linked databases, prevents such data loss (see chapters 5 and 9).

If you now want to use the data of another scientist, you face a number of challenges. First, the legal issues must be clarified (Chapter 3). Then, the quality of the data and its preparation for meaningful reuse must be clarified [**Briney**]. Major challenges often arise from a lack of documentation and incorrect data. Adequate documentation is one of the most important aspects for reuse, as you need to know and understand the details of a data set in order to be able to use it. If, for example, the names of variables are unknown, meaningful use is not possible. If an article has been published in parallel, you can try to obtain the necessary information with its help. The last resort may be to contact the author or creator of the dataset directly. Errors in the data set can also be a serious problem (Chapter 6). These can be inconsistencies, invalid values, missing or incorrect values. Even if no errors can be detected during an initial review of the data, a few simple tests should be carried out before using the data productively. A graphical analysis as shown in Chapter 6.2 can provide quick results here. At the same time, you will develop a better understanding of the data set. Therefore, we recommend that you start reusing research data from other sources by to "play around" with the data to make sure that you can and want to use it.

In summary, good preparation and transparent handling of research data greatly facilitate reuse and thus promote scientific progress.

8.4. Research data for machine learning (AI)

A very interesting form of re-use of research data, whose importance is growing exponentially, is processing by means of machine learning (ML) or, more generally, through the use of artificial intelligence (AI). The existing literature on artificial intelligence and machine learning now fills entire libraries, and the topic has become ubiquitous due to the increased use of large language models (LLM) such as ChatGPT, Grok, Gemini, Mistral, DeepSeek, Claude, and many more. For a comprehensible introduction, we recommend the two books by J. Frochte (Machine Learning (2021); [**Frochte**]) and O. Zeigermann & C.N. Nguyen (Machine Learning kurz & gut (2024); [**Zeigermann**]). Therefore, we will not go into detail about the various methods and ML or AI algorithms here, but will instead describe the essential characteristics that research data must have in order to be used (automatically) by AI.

The simplest file format is tables in the form of **CSV files**, i.e. simple tables of numerical values separated by commas, where the first row consists of short descriptions of the respective columns. A simple example can be seen in Figure 6.2. In this case, we refer to **structured data**. Data in **JSON format** is also suitable, as this format is generally directly machine-readable. The electronic laboratory notebooks JuliaBase, eLabFTW and Kadi4Mat (Chapter 5), for example, can export data directly to JSON format. JSON (JavaScript Object Notation) is a lightweight data interchange format that is easy for humans to read and write and easy for machines to analyse and generate. It is based on a subset of the JavaScript programming language and is commonly used for transferring data between a server and web applications. The following code shows a simple example of describing a person (`https://en.wikipedia.org/wiki/JSON`):

```
{
"first_name": "John",
"last_name": "Smith",
"is_alive": true,
"age": 27,
"address": {
"street_address": "21 2nd Street",
"city": "New York",
"state": "NY",
"postal_code": "10021-3100"
},
"phone_numbers": [
{
"type": "home",
"number": "212 555-1234"
},
{
"type": "office",
"number": "646 555-4567"
}
],
"children": [
"Catherine",
"Thomas",
"Trevor"
],
"spouse": null
}
```

Another category is **unstructured data**. Unstructured data is information that is not stored in a predefined, relational format. It often comes in the form of text, images, videos, audio files or other data sets that do not have a clear structure. Unlike structured data, which is stored in database tables with fixed columns and rows, unstructured data can exist in various formats and without a defined arrangement.
Another possibility is so-called **Big Data**. This refers to data sets that pose a problem for conventional data processing or data analysis due to their volume, complexity, weak structure and/or fast pace [**Frochte**]. Big data is characterised by a large data volume, the high speed at which new data is generated, and a wide range of data types and sources. Large volumes of data are generally not a problem, but high speed and bandwidth are. Most AI algorithms rely on the characteristics of the data remaining constant. If, for example, a new data source is introduced, a new sensor for measuring temperature or humidity, and this sensor provides data that was not previously available, then you are initially faced with a challenge. Dealing with such unstructured data, which must also be considered separately for each individual case, requires expert knowledge, and an appropriate presentation would go far beyond the scope of this manual. At this point, we would therefore simply like to raise awareness of the problem. For beginners, we recommend starting with the supposedly much simpler handling of structured data.

9. Permanent data storage

To avoid unwanted Data loss, good and established strategies are required for the secure storage, backup, transfer and disposal of data. In collaborative projects, additional challenges arise with regard to the joint storage of and access to data. Therefore, the basic structures and strategies should be defined in the data management plan at the beginning of a new project (Chapter 4). Answering the following fundamental questions will help with planning [**Corti**]:

- How much storage space is required for the project?
- Who needs access to the data?
- What security measures must be taken against data loss?
- Will personal data be processed or stored?

The following principles should be observed for the long-term storage of research data:

1. Data should be stored in uncompressed and non-commercial (proprietary) formats or in open standard file formats to ensure long-term readability.
2. Copy or migrate files to new media every two to five years, as both optical and magnetic storage media age.
3. Implement a storage strategy, even for shorter projects, using at least two different storage formats, such as hard drives and optical storage media (CD, DVD), in parallel.
4. Regularly check the data integrity of stored data, for example by means of checksums.
5. Organise and label stored data clearly using a predefined naming scheme so that data can be retrieved (more easily) later.
6. Ensure that the rooms in which the storage media are located meet valid security requirements, i.e. are also protected against fire and flooding. Prevent unauthorised access.
7. Create digital copies of paper data in PDF/A format for long-term backup.

If you use an electronic laboratory notebook for your work and have written or have access to a DMP, then most of the tasks mentioned above are already done automatically for you. Most ELNs use their own database, and if your ELN is maintained by a system administrator, then regular backups of all data should also be made to additional storage media. This is exactly the case with the three ELNs described in Chapter 5, especially since these three ELNs are operated by the authors of this manual themselves.

Encryption

Another option for the long-term secure handling of research data is the encryption of data, for example for backups and data transfer. Not only individual files can be encrypted, but also entire storage media or containers for many files, the latter for example with the open source software *VeraCrypt* (`https://veracrypt.io/en/Downloads.html`). Encryption software uses special algorithms to encrypt data, and a key in the form of a password or passphrase is required for subsequent decryption. The larger the key size (e.g. 256 bits), the more difficult it is to gain unauthorised access to the data. However, quantum cryptographic methods will require new strategies in the future to ensure secure encryption. At present, however, this is still largely a dream of the future. There are a variety of encryption programs available on the market. One standard is Pretty Good Privacy (PGP), which is freely available as an open-source version in the form of GnuPG. To use it, a public key and a private key must be generated, as well as a passphrase. These components are used for a digital signature, which allows the recipient of a data record to verify the identity of the sender. The recipient's public key must be available to the sender so that they are authorised to encrypt data for that specific recipient. Here is a quick guide to encrypting data with PGP:

1. The following steps only need to be performed once:

- Install the encryption software, e.g. *GnuPG*.
- Generate a key pair with a public and a private key and a passphrase.
- Download the public key of an institution with which you wish to exchange data.
- Import this public key into your PGP software.

2. The following steps must then be carried out for each encryption process, for example when sending encrypted data to a specific recipient:

- Select the files to be encrypted.
- Select the appropriate public key for the recipient.
- Sign the files to be encrypted with your private key and passphrase.
- Then encrypt the files with the recipient's public key.
- Send the data to the recipient via a secure file transfer protocol (e.g. FTPS, HTTPS, SFPT, etc.) or by post on an external data carrier.

Summary:

Before we go into the differences between databases, repositories and their uses, here is a brief summary of the most important points to consider for secure and long-term data storage. Researchers should develop a well-thought-out strategy for storing, backing up, transferring and disposing of research data for their projects. This will protect research data from external attacks and prevent loss. Such a strategy also supports the handling of research data in accordance with the FAIR principles (see Chapter 2) and should achieve the following:

- Identify and select the best location for your data storage.
- Be aware of the risks and benefits of storing data in a cloud and weigh them carefully.

- Develop a personal strategy for encrypting your data.
- Consider how often and where backups of your data should be stored.
- Also consider how effectively and securely data that is no longer needed can be deleted or disposed of in various locations once you have saved all data in a database or repository at the end of a project.

9.1. Databases

A database is an organised collection of data that can be easily accessed, managed and kept up to date. Even if the use of a database is not explicitly planned at the start of a new project (which a good DMP should prevent!), your data will inevitably come into contact with one or more databases during processing. If you use an ELN, this will definitely happen. Databases offer many advantages. For one thing, you can use scripts to exchange data between different applications, and database languages such as SQL help you organise data and answer questions about it. In addition, many database systems allow the integration and execution of your own program code (e.g. Python) to improve the performance and modularity of an application.

There are two categories of databases: **relational and non-relational databases** (NoSQL). Relational databases have fixed schemas for how data is stored. This approach is intended to ensure data integrity, consistency and accuracy. However, the major disadvantage of relational databases is that they are not easily scalable as data volumes increase. In contrast, NoSQL databases have no restrictions on data structures, allowing for greater flexibility, adaptability and scalability. If you want to delve deeper into database techniques and are not afraid of developing SQL scripts and the Python programming language, the book by Y. Vasiliev [**Vasiliev**] is a good and easy introduction. Currently, **relational databases** are the most common and provide a structured way to store data. To work with such a database, a data schema must first be defined, for example, for books, the fields book title, author, publisher, year, etc. The data must be stored in the database in this predefined schema. When working with a relational database, you start by defining such a schema. You define a collection of tables, each consisting of a set of fields or columns, and you specify what type of data the different fields should store. You also establish relationships between the tables. This allows you to store data in a relational database, retrieve data from it, or update data. Well-known relational database management systems are **MySQL**, **MariaDB** and **PostgreSQL**.

NoSQL databases, on the other hand, do not require a predefined organisational scheme for the data to be stored and do not support standard database operations such as *join*. An SQL JOIN is an operation in relational databases that enables queries across multiple database tables. JOINs merge data stored in different tables and output it in filtered form in a result table. Instead, NoSQL databases allow large volumes of data to be stored in a flexible form, simplifying the handling of large to very large amounts of data. The storage of so-called key values enables data to be stored and output as pairs of key values. This eliminates the need to store different information about a specific object in the form of multiple tables. Among other things, documents can be stored and processed in JSON format. These are then also machine-readable in principle (see Chapter 8.4). NoSQL databases are particularly suitable for **real-time applications** and **big data** projects, which is why Google uses them for its email service Gmail and the business platform LinkedIn. **MongoDB** is a NoSQL database management system well known for such purposes.

At the beginning of a project, you should therefore consider which type of database is suitable for your project and record this in the DMP. For practical use and programming of specific databases, we would like to refer you to the widely available literature or, if in doubt, recommend that you contact appropriate specialists at your scientific institution.

9.2. Repositories

We have already discussed repositories in Chapter 8 in connection with the publication of data sets. In this chapter, we would like to supplement this information with the possible use of repositories for data storage, in particular for the storage of metadata (Chapter 2.1). A repository is a storage location for digital objects that makes them available to the public or a restricted group of users. It is a type of database or electronic archive that is used to store, manage and provide access to data, publications or other digital resources. A repository offers functions for organising, classifying and managing the stored data and enables authorised persons or the public to access and use the stored resources. In the scientific field, repositories are often used to store research data, publications (e.g. dissertations, articles) or Open Educational Resources (OER). Although repositories are often compared to archives, there is a difference. While archives are primarily used for the long-term preservation of historical documents, repositories can also be used for the short- or medium-term storage and use of current data, taking into account different requirements and terms of use.

We have already encountered a repository that is popular in the scientific community several times in this handbook: Zenodo. Research data repositories are specialised archives that permanently store, organise and make research data accessible. They serve to secure and reuse research data and can be subject-specific, institutional or generic. The selection of a suitable repository should be based on the conventions of the respective discipline or the requirements of funding institutions. Examples of such repositories are **Zenodo** - A generic repository operated by CERN (see above), **DRYAD** - A repository for research data from the life sciences, **Figshare** - Another generic repository for research data, **DaKS** - The data repository of the University of Kassel, **ResearchData** - the repository of Heinrich Heine University Düsseldorf, **GFZ Data Services** - A data repository for the geosciences, and **PANGAEA** - Another data repository for the geosciences. The website `https://open-access.network/informieren/publizieren/repositorien` now lists more than 5700 repositories, some of which store data in addition to the examples mentioned above.

In Chapter 8.2, we demonstrated in principle how data can be published and stored in such a repository using the example of Zenodo. Research data can also be stored internally in a repository without being published. Many repositories offer the option of placing data under an embargo so that it can only be made publicly available at a later date. There are also data journals that publish research data using a peer review process, similar to traditional scientific journals.

9.3. Coscine

Coscine (**Co**llaborative **Sc**ientific **In**tegration **E**nvironment) is neither a pure database nor a pure repository, but rather sees itself as a **platform for research data management**. The platform is hosted at the RWTH Aachen University (RWTH Aachen) (https://www.itc.rwth-aachen.de/cms/it-center/services/forschung/ smhwy/coscine/) and offers storage space (access to free storage space on the Research Data Storage), integration (access to project-related data sources, e.g. research data storage, linked files, archived data), collaboration (access for all project members), metadata (automatic linking to project data), individuality (creation of project-specific metadata as application profiles) and archiving (archive research data on site).

The motivation behind Coscine is to provide researchers with a secure and supportive research data management platform for collaborative research work during the active project phase, while complying with FAIR principles. Coscine does not seek to replace established services (such as S3 storage, GitLab, or cloud services), but rather adds a FAIR layer (including metadata annotation and PID assignment) around the services (known as resources). This enables project-wide access and facilitates collaboration. The aim is to integrate Coscine well into the existing processes of researchers and to deliver added value for research data management. The following figure 9.1 illustrates the basic workflow when working with Coscine.

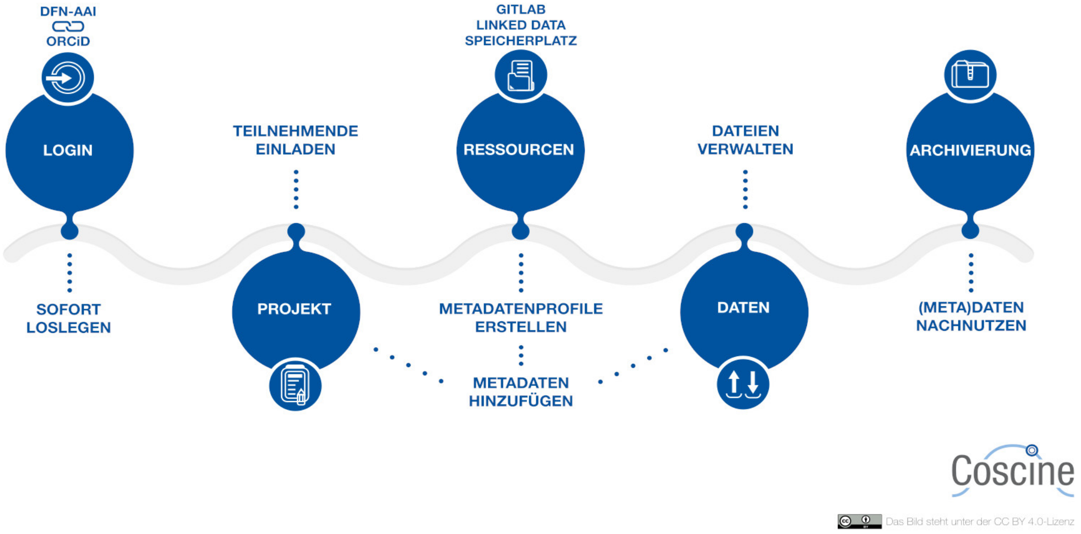


Figure 9.1.: Typical workflow of Coscine.

For researchers who cannot or do not want to use an ELN, Coscine may be an alternative under certain circumstances. However, it should always be kept in mind that this is a cloud-based solution with all the advantages and disadvantages that this entails, particularly in terms of data security. Compared to an ELN that is available on site, there are certain limitations, especially when connecting experiments that are located in an institute in an IT-secured area. Coscine is open source and is developed on GitLab (https://git.rwth-aachen.de/coscine).

A. Appendix

We expressly accept no responsibility for the accuracy and security of the web links (URLs) listed in this manual or for the content of the linked websites.

A.1. Research data organisations in Germany

- **Alliance of German Science Organisations** (`https://www.allianz-der-wissenschaftsorganisationen.de/`)
- **Data Competence Centres** at universities and universities of applied sciences (`https://www.bmbf.de/DE/Forschung/Wissenschaftssystem/Forschungsdaten/DatenkompetenzenInDerWissenschaft/datenkompetenzeninderwissenschaft_node.html`)
- German Research Foundation **DFG** (`https://www.dfg.de/de`)
- State initiative **fdm.nrw** for research data management in North Rhine-Westphalia (`https://www.fdm.nrw/`), as an example of a state initiative
- Information portal **forschungsdaten.info** (`https://forschungsdaten.info/`)
- **NFDI** National Research Data Infrastructure (`https://www.nfdi.de`)

A.2. Further information on the Internet

B2FIND `https://eudat.eu/service-catalogue/b2find`

CERN `https://www.home.cern/`

COD `http://www.crystallography.net/cod/`

Coscine `https://www.itc.rwth-aachen.de/cms/it-center/services/forschung/~smhwy/coscine/`

DaKS `https://daks.uni-kassel.de/home`

DataCite `https://commons.datacite.org/`

Dataset Search `https://datasetsearch.research.google.com/?hl=de`

DFG `https://www.dfg.de/antragstellung/forschungsdaten`

DOI `https://www.doi.org/`

DRYAD https://datadryad.org

eLabFTW https://www.elabftw.net/

EUDAT https://eudat.eu

figshare https://figshare.com

research data.info https://forschungsdaten.info/

GNU https://www.gnu.org

GnuPG https://www.gnupg.org

HMC https://helmholtz-metadaten.de/de

IPFS https://docs.ipfs.tech

JuliaBase https://www.juliabase.org/

JülichDATA https://data.fz-juelich.de/

Kadi4Mat https://kadi.iam.kit.edu/

Mendeley Data https://data.mendeley.com/

Metadata4Ing https://nfdi4ing.pages.rwth-aachen.de/metadata4ing/metadata4ing

NFDI https://www.nfdi.de/

NFDI4Ing https://nfdi4ing.de/

ODC https://opendatacommons.org/

PANGAEA https://www.pangaea.de

PROV-O https://www.w3.org/TR/prov-o/

RDF https://www.w3.org/RDF/

RDMO https://rdmorganiser.github.io/

re3data https://www.re3data.org/

RIsources https://risources.dfg.de

SciMesh https://scimesh.org/about/

VerbundFDB https://www.forschungsdaten-bildung.de/

Zenodo https://zenodo.org/

B. List of abbreviations

API Application Programming Interfaces

ASCII ASCII format for text files (American Standard Code for Information Interchange)

B2FIND Search engine for research data, provided by the European Union

BDSG Federal Data Protection Act

BFO Basic Formal Ontology

BMFTE Federal Ministry of Research, Technology and Space

BMWE Federal Ministry for Economic Affairs and Energy

BPersVG Federal Personnel Representation Act

BVerwG Federal Administrative Court

CC-BY Creative Commons licence model

CCO Creative Commons licence model

CERN European Organisation for Nuclear Research (Conseil Européen pour la Recherche Nucléaire)

CIF Crystallographic Information File (standard text file format for representing crystallographic information)

COD Crystallography Open Database (freely accessible database containing crystal structures from scientific publications)

Coscine Collaborative Scientific Integration Environment

CSV Comma-separated values (file format)

DaKS Data repository of the University of Kassel

DataCite An international consortium that aims to provide easy access to scientific research data.

DCAT Data Catalog Vocabulary

DFG German Research Foundation

DMP Data management plan

DOI Digital Object Identifier (A unique and permanently valid identifier for publications, research data, videos and other scientific resources on the Internet, similar to an ISBN or ISSN for books or journals.)

DRYAD Open Data Publishing Platform

DSGVO General Data Protection Regulation (engl. GDPR)

ELN Electronic Lab Notebook

EU European Union

EXIF Exchangeable Image File Format

EUDAT European Data Initiative

FAIR Findability, Accessibility, Interoperability, and Reusability

FDM Research data management

figshare Provider of Open Research Repository Infrastructure

GG Basic Law of the Federal Republic of Germany

GNU GNU’s Not Unix (alternative, free operating system)

GnuPG GNU Privacy Guard

HGF Helmholtz Association of German Research Centres

HMC Helmholtz Metadata Collaboration

HTTP Hypertext Transfer Protocol

IPFS InterPlanetary File System

JSON JavaScript Object Notation

AI Artificial Intelligence

LLM Large Language Model

m4i Metadata4Ing

ML Machine learning

NFDI National Research Data Infrastructure

NFDI4Ing The NFDI4Ing consortium for engineers, part of NFDI

NMR Nuclear Magnetic Resonance (Nuclear Magnetic Resonance Spectroscopy)

ODC Open Data Commons - licence model

OER Open Educational Resources

PANGAEA Repository for geosciences

PGP Pretty Good Privacy

PID Persistent and unique identifier (permanent digital identifier)

PROV-O The PROV Ontology of W3C

RDF Resource Description Framework

RDMO Research Data Management Organiser

re3data Directory of research data repositories

RIsources Portal for research infrastructures

UrhG Copyright Act

URI Uniform Resource Identifier (identifier consisting of a string of characters used to identify an abstract or physical resource)

URL Uniform Resource Locator (identifies and locates a resource, e.g. a website)

Bibliography

[Allemang] Allemang, D.; Hendler, J.; Gandon, F. *Semantic Web for the Working Ontologist*; Morgan & Claypool Publishers: ACM Books #33, Association for Computing Machinery, Kentfield CA, U.S.A., 2020.

[Al-Salman] Al-Salman, R.; Aguiar Teixeira, C.; Zschumme, P.; Lee, S.; Griem, L.; Aghassi-Hagmann, J.; Kirschlechner, C.; Selzer, M. *KadiStudio Use-Case Workflow: Automation of Data Processing for in Situ Micropillar Compression Tests*; Data Science Journal, 22:21, 1-11, 2023.

[Baumann] Baumann, P. *Legal Issues in Decisions on the Use and Storage of Research Data, especially in Inter-institutional Research Projects*; Presentation at the NFDI4ing Congress, Germany, 2023.

[Brandt] Brandt, N.; Griem, L.; Herrmann, C.; Schoof, E.; Tosato, G.; Zhao, Y.; Zschumme, P.; Selzer, M. *Kadi4Mat: A Research Data Infrastructure for Materials Science*; Data Science Journal, 20:8, 1-14, 2021.

[Brehm] Brehm, E. *Guidelines for Text and Data Mining for Research Purposes in Germany*; NFDI4ing and TIB - Leibniz Information Centre for Science and Technology, University Library: Hanover, Germany, 2022.

[Bremecker] Bremecker, D., *Mitbestimmung/Mitwirkung / 2.4.17 Introduction and application of technical control devices*; Haufe TVöD Office Professional für die Verwaltung: Haufe-Lexware GmbH und Co. KG, Freiburg, Germany, 2023.

[Briney] Briney, K. Data Management for Researchers; Pelagic Publishing: Exeter, UK, 2015.

[Corti] Corti, L.; Van den Eynden, V.; Bishop, L.; Woollard, M. *Managing and Sharing Research Data*; SAGE Publications Ltd.: London, UK, 2020.

[DFG] German Research Foundation *Recommendations for Handling Research Data*; DFG: Bonn, Germany, 2023.

[EU] European Commission, *H2O2O Programme AGA - Annotated Model Grant Agreement Version 5.2*, EU: Brussels, Belgium, 2019.

[Frochte] Frochte, J. *Machine Learning*; 3rd edition; Hanser-Verlag, Munich, Germany, 2021.

[Griem] Griem, L.; Zschumme, P.; Laqua, M.; Brandt, N.; Schoof, E.; Altschuh, P.; Selzer, M. *KadiStudio: FAIR Modelling of Scientific Research Processes*; Data Science Journal, 21:16, 1-17, 2022.

[Jalali] Jalali, M.; Luo, Y.; Caulfield, L.; Sauter, E.; Nefedov, A.; Wöll, C. *Large language models in electronic laboratory notebooks: Transforming materials science research workflows*; Materials Today Communications, 40, 109801, 2024.

[Johannes] Johannes, P.C. *The researcher's right to data protection*; Springer-Verlag, Data Protection and Data Security - DuD, 11, 817–822, 2012.

[Lauber] Lauber-Rönsberg, A. *Legal aspects of research data management*; In: *Practical handbook for research data management*; Putnings, M.; Neuroth, H.; Neumann, J. (Eds.); De Gruyter: Berlin/Boston, 2023.

[Nield] Nield, T. *Math Basics for Data Scientists*; O'Reilly, Heidelberg, 2024.

[Papula] Papula, L. *Mathematics for Engineers and Scientists, Volume 3*; Springer Vieweg, Wiesbaden, 8th edition, 2024.

[Putnings] Putnings, M.; Neuroth, H.; Neumann, J. (Eds.) *Practical Handbook of Research Data Management*; De Gruyter, Berlin/Boston, 2023.

[Stamile] Stamile, C.; Marzullo, A.; Deusebio, E. *Graph Machine Learning*; Packt Publishing: Birmingham, UK, 2021.

[Vasiliev] Vasiliev, Y. Python for Data Science; No Starch Press, San Francisco, USA, 2022.

[ZB] ZB-MED (ed.) *ELN Guide: Electronic Laboratory Notebooks in the Context of Research Data Management and Good Scientific Practice - A Guide for the Life Sciences*; 2nd edition; Publisso: Cologne, Germany, 2020.

[Zeigermann] Zeigermann, O.; Nguyen, C.N. *Machine Learning kurz & gut*; 3rd edition; O'Reilly, Heidelberg, Germany, 2024.

Index