Forschungsdatenmanagement
Eine praxisorientierte Einführung

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

Creative-Commons-Lizenz*

*Dieser Text steht unter der Creative-Commons-Lizenz (Namensnennung) - Weitergabe unter gleichen Bedingungen) 4.0 International. Um eine Kopie dieser Lizenz zu sehen, besuchen Sie creativecommons.org/licenses/by-sa/4.0/.

Inhaltsverzeichnis

1. Einführung **5**

2. Der Lebenszyklus von Forschungsdaten **7**
2.1. Daten und Metadaten 8
2.2. Die Erzeugung FAIRer Forschungsdaten 9
2.2.1. Auffindbarkeit 10
2.2.2. Zugänglichkeit 10
2.2.3. Verarbeitbarkeit 10
2.2.4. Nach-Nutzbarkeit 11

3. Rechtliche Aspekte **13**
3.1. Urheberrechte bei der Nutzung fremder Daten 13
3.1.1. Text und Data Mining 13
3.2. Verarbeitung personenbezogener Daten 14
3.3. Rechtliche Rahmenbedingungen für die Weitergabe von Daten 14

4. Der Datenmanagementplan **17**
4.1. Datenbeschreibung 18
4.2. Dokumentation und Datenqualität 18
4.3. Speicherung und technische Sicherung während des Projektverlaufs 19
4.4. Rechtliche Verpflichtungen und Rahmenbedingungen 19
4.5. Datenaustausch und dauerhafte Zugänglichkeit der Daten 19
4.6. Verantwortlichkeit und Ressourcen 20

5. Die Datenerfassung, Datenspeicherung und Dokumentation **21**
5.1. Elektronische Laborbücher 21
5.2. Populäre elektronische Laborbücher 22
5.3. Elektronische Laborbücher in der Praxis 23
5.3.1. JuliaBase 23
5.3.2. eLabFTW 35
5.3.3. Kadi4Mat 48

6. Datenqualität **61**
6.1. Messdaten und ihre Fehler 62
6.2. Datenanalyse und die Visualisierung von Daten 62

7. Datenaustausch und Datennachverfolgung **67**
7.1. Datenaustausch zwischen elektronischen Laborbüchern mit SciMesh 67
7.2. Datennachverfolgung 71
7.3. Implementierung von SciMesh 74

7.4. Den Graphen erhalten 74
7.5. MetaData4Ing 76

8. Datenpublikation **79**
8.1. Die Veröffentlichung von Datensätzen 79
8.2. Beispiel: Publikation eines Datensatzes bei Zenodo 81
8.3. Die Nachnutzung von Forschungsdaten 82
8.4. Forschungsdaten für Maschinelles Lernen (KI) 83

9. Dauerhafte Datenspeicherung **87**
9.1. Datenbanken 89
9.2. Repositorien 90
9.3. Coscine 91

A. Anhang **93**
A.1. Forschungsdatenorganisationen in Deutschland 93
A.2. Weiterführende Informationen im Internet 93

B. Abkürzungsverzeichnis **95**

Literatur **99**

Index **101**

1. Einführung

Das professionelle Management von Forschungsdaten gewinnt in der Wissenschaft immer mehr an Bedeutung und wird inzwischen von vielen Drittmittelgebern (DFG, BMFTR, BMWE, EU, u.a.) für die Bewilligung von Fördergeldern für neue Forschungsprojekte dringend empfohlen bzw. vorausgesetzt. Die zentrale Motivation dafür ist eine umfangreiche Sammlung von Forschungsdaten, deren Erzeugung vorwiegend aus Steuermitteln finanziert wird. Dabei steht zum einen der Aspekt guter wissenschaftlicher Praxis und die Nachvollziehbarkeit von Forschungsergebnissen im Fokus. Zum anderen geht es um die ermöglichte und sinnvolle Nachnutzung, beispielsweise als Trainingsdaten im Bereich des maschinellen Lernens (ML) oder allgemein bei Anwendungen mit künstlicher Intelligenz (KI).

Ein weiterer Aspekt ist die Schonung von Ressourcen, sei es finanzieller Art oder zum Schutz der Umwelt und des Klimas. Die Durchführung von Experimenten und Computer-Simulationen benötigt ggf. viel Energie und trägt damit auch zu einem verstärktem CO_2-Ausstoss in die Atmosphäre bei. Hier kann die Wissenschaft ihren Beitrag leisten, indem die wiederholte Generierung von bereits vorhandenen Daten vermieden wird.

Angestrebt wird die Erzeugung FAIRer Daten, d.h. von Forschungsdaten die auffindbar, zugänglich, verarbeitbar und nachnutzbar sind (FAIR: Findability, Accessibility, Interoperability, and Reusability). Damit dieses Ziel erreicht werden kann, ist der Aufbau einer umfangreichen Infrastruktur erforderlich (Hardware, Software, u.a.). Die Bundesregierung fördert diesen Aufbau zum Beispiel durch die Finanzierung der Nationalen Forschungsdateninfrastruktur (NFDI) und den darin zusammengefassten dreißig Konsortien, die sich jeweils an den Bedarfen der verschiedenen wissenschaftlichen und technischen Fachrichtungen orientieren.

Mit dieser Einführung in das Forschungsdatenmanagement (FDM) möchten wir allen, die Forschungsdaten erzeugen und/oder diese nutzen wollen, unabhängig von der Art der Daten, praktische Hilfestellungen für alle FDM-Stadien an die Hand geben. Nach der Lektüre dieses Best-Practice-Ratgebers sollten Sie in der Lage sein, wesentliche FDM-Strukturen zu verstehen und sicher mit Ihren Daten umgehen zu können, von der Erzeugung, über die Analyse, bis zur dauerhaften Speicherung Ihrer Daten.

Forschungsdaten sind eine wertvolle Ressource und es lohnt sich, sorgfältig und nachhaltig mit ihnen umzugehen. In diesem Sinne wünschen wir Ihnen viel Vergnügen bei der Lektüre dieses Ratgebers.

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

Jülich und Karlsruhe, im Juli 2025

2. Der Lebenszyklus von Forschungsdaten

Als Forschungsdaten bezeichnet man sämtliche Informationen, die während wissenschaftlicher Arbeiten entstehen und verarbeitet werden. Die Registrierung und Dokumentation kann analog erfolgen, d.h. traditionell mit Papier und Bleistift, oder in digitaler Form mit Hilfe von Computern. Dann werden Informationen meistens als Dateien in unterschiedlichsten Formaten gespeichert. Bei Experimenten und Messungen können zum Beispiel Forschungsdaten als einfache Textdateien (ASCII-Format), als komplexere Tabellen in proprietären Formaten (z.B. MS-Excel), als Bilder und Grafiken (z.B. im JPG-Format), als Audio- oder Videodateien (z.B. im MP3- oder MP4-Format) sowie als Quellcode von Software erzeugt werden (z.B. Python- oder C-Code). Zu unterscheiden sind dabei die eigentlichen Daten und die sie beschreibenden Metadaten. Auf diesen wichtigen Unterschied werden wir im folgenden Abschnitt näher eingehen. Die folgende Abbildung 2.1 soll einen typischen Lebenszyklus von Forschungsdaten veranschaulichen. Im Lebenszyklus von Forschungsdaten spielen elektronische

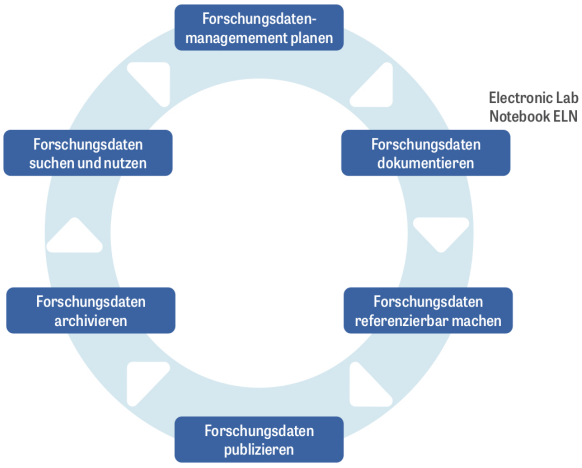


Abbildung 2.1.: Der Lebenszyklus von Forschungsdaten [**ZB**].

Laborbücher ELN eine zentrale Rolle [**ZB**], von der Erfassung der Daten und Metadaten, der Prozessierung und Datenanalyse, bis hin zur Publikation, Archivierung und dem Teilen von Daten. Zu den Best-Practices eines modernen und umfassenden Forschungsdatenmanagements FDM gehört obligatorisch der Einsatz eines ELN [**ZB**, **Briney**, **Corti**, **Putnings**]. Wir werden daher in diesem Ratgeber in den folgenden Kapiteln noch intensiv und wiederholt darauf eingehen.

Zusammenfassung:

Forschungsdaten sind sämtliche analoge und digitale Informationen, die während wissenschaftlicher Arbeiten entstehen, verarbeitet und gespeichert werden.

2.1. Daten und Metadaten

Zunächst stellt sich die Frage, was eigentlich genau Forschungsdaten sind? Diese Frage ist nicht ganz einfach zu beantworten, da wissenschaftliche Forschungsfelder sehr divers und heterogen sein können und dementsprechend auch die entstehenden Daten.
Praktisch handelt es sich bereits bei einem einzelnen Messwert (eine Zahl mit einer zugehörigen Dimension; z.B. ein Druck von 1 bar) bereits um Forschungsdaten, auch wenn in diesem Beispiel nur ein einzelner Datenpunkt vorliegt, der aber ggf. mit großem Aufwand bestimmt worden ist. Große Forschungsdatensätze können in der Medizin entstehen, zum Beispiel bei dem Scannen von menschlichen Gehirnen mittels Kernspinresonanzspektroskopie NMR. Möglich sind auch riesige Datensätze, die nach ihrer Entstehung nur an ihrem Entstehungsort gespeichert und weiter verarbeitet werden können und allein wegen ihrer Größe nicht transferierbar sind. Ein Beispiel sind Messdaten von Teilchenkollisionen am CERN (Europäische Organisation für Kernforschung) in der Nähe von Genf, wo der Aufbau von Materie erforscht wird.
Generell unterscheiden wir primäre und sekundäre Forschungsdaten. Primärdaten sind die zuerst entstandenen Rohdaten und bei den Sekundärdaten handelt es sich um weiter verarbeitete Primärdaten. Bei den Primärdaten kann eine weitere Unterteilung vorgenommen werden. Die folgende Tabelle veranschaulicht diese Einteilung:

Tabelle 2.1.: Primäre und sekundäre Daten.

Primärdaten	Sekundärdaten
Beobachtungsdaten	Verarbeitete Rohdaten
Experimentelle Daten	Gesammelte Daten
Simulierte Daten	

Beobachtungsdaten können als Beispiel Wetterdaten sein, wie die kontinuierliche Aufzeichnung von Temperatur und Luftdruck. Experimentelle Daten werden in der Regel in einem Labor erzeugt und simulierte Daten mittels spezieller Software in einem Computer.

Eine eigene Kategorie bilden Metadaten. Wir sprechen hier auch von *Daten über Daten*, d.h. zusätzliche Informationen darüber, wie Primär- oder Sekundärdaten entstanden sind. Ein populäres Beispiel für Metadaten sind EXIF-Daten. EXIF ist ein Standardformat der Japan Electronic and Information Technology Industries Association für das Abspeichern von Metadaten in digitalen Bildern. Werden Fotos mit einer Bildbearbeitungssoftware verarbeitet, so können auch die zu einem Bild bzw. Foto zugehörigen und von einer Digitalkamera aufgezeichneten Bilddaten (EXIF-Daten) angezeigt und für die Suche von bestimmten Bildern in einer Bilddatenbank verwendet werden (*Finde alle Bilder, die mit einer Objektiv-Brennweite von 50 mm aufgenommen wurden, u.a.*).

Vergleichbar ist in der Kristallstrukturforschung das seit Jahrzehnten etablierte CIF-Format (Crystallographic Information File), mit dessen Hilfe die Struktureigenschaften eines kristallinen Materials und deren experimentelle Bestimmung genau beschrieben werden. Dieses geschieht anhand festgelegter Konventionen, die auch viele Jahre später die Generierung, Reproduktion und Auffindbarkeit von Strukturdaten ermöglichen. Die Abbildung 2.2 zeigt als einfaches Beispiel einen Ausschnitt aus der CIF-Datei von Gold (Au), entnommen aus der frei verfügbaren Crystallography Open Database COD. Allein die Eingabe des Elementnamens *Au* reicht in diesem Fall aus, um über (`http://www.crystallography.net/cod/`) die gewünschten Informationen und das CIF-File aus der COD zu erhalten und für die Weiterverarbeitung nutzen zu können. Man erhält Informationen

darüber, wer wann die Strukturdaten bestimmt hat und in welcher Zeitschrift die Daten ursprünglich publiziert worden sind. Außerdem enthält das CIF-File Informationen über die chemische Zusammensetzung, Symmetrieinformationen, die Nummer des Datensatzes innerhalb der COD-Datenbank, u.v.m. Kurz gefasst handelt es sich bei Metadaten um eine strukturierte, digitale Form der Dokumentation.

Zusammenfassung:

Primärdaten sind Rohdaten und Sekundärdaten sind immer weiter verarbeitete Daten. Metadaten sind eine strukturierte, digitale Form der Dokumentation von Primär- und Sekundärdaten.

```
#$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
```

Abbildung 2.2.: CIF-Datei von Gold (Au).

2.2. Die Erzeugung FAIRer Forschungsdaten

Die Erzeugung FAIRer Forschungsdaten setzt als wichtigstes Kriterium voraus, das diese Daten auffindbar und zugänglich sind. Dabei spielt es keine Rolle, ob es sich um Primär- oder Sekundärdaten handelt. Weiterhin müssen Daten interoperabel und nach-nutzbar sein. Wir werden in den folgenden Kapiteln dieses Ratgebers immer wieder auf diese Kriterien eingehen. Hier soll zunächst kurz umrissen werden, welche Kriterien jeweils erfüllt sein müssen, damit Forschungsdaten FAIR sind.

2.2.1. Auffindbarkeit

Damit Forschungsdaten gefunden werden können, müssen zuerst ihre zugehörigen Metadaten öffentlich zugänglich gemacht werden. Daten müssen außerdem über Standardmechanismen eindeutig identifiziert werden können, zum Beispiel durch die Vergabe von dauerhaften, digitalen Kennungen PID: persistent and unique identifiers. Die praktische Vergabe solcher PID's erläutern wir in den Kapiteln zu den elektronischen Laborbüchern (ELN) und den Knowledgegraphen. Es sollten hierbei entweder bestehende Namenskonventionen für die jeweilige Fachdisziplin beachtet, oder falls diese bislang nicht existieren, neue Namenskonventionen geschaffen werden. Damit wird auch die spätere Suche in Datenbanken oder Repositorien über Schlüsselwörter wesentlich erleichtert. Idealerweise werden vorhandene Metadatenstandards berücksichtigt oder neue Standards geschaffen. Auch dazu werden wir in den folgenden Kapiteln ausführliche und praktische Hinweise geben.

2.2.2. Zugänglichkeit

Zu Beginn muss festgelegt werden, welche Daten überhaupt freigegeben bzw. publiziert werden sollen. Es ist nicht immer gewünscht oder ratsam, alle Daten eines Forschungsprojektes frei verfügbar zu machen, insbesondere wenn eventuelle Patente auf entwickelte Verfahren oder Materialien angemeldet werden sollen. Das gilt auch für Projekte, die ausschließlich mit Hilfe von öffentlichen Drittmitteln finanziert werden. Ist diese Entscheidung gefallen, muss überlegt werden, wie die Daten technisch zugänglich gemacht werden sollen. Welche Infrastruktur und welche Software soll hierzu genutzt werden? Ein populäre und relativ simple Möglichkeit ist das Hochladen in frei zugängliche Repositorien wie *Zenodo* (https://zenodo.org/). Zenodo ist ein Online-Speicherdienst, der hauptsächlich für wissenschaftliche Datensätze, aber auch für wissenschaftsbezogene Software, Publikationen, Berichte, Präsentationen, Videos etc. verwendet werden kann. Finanziert wird der Dienst über die Europäische Kommission und betrieben wird Zenodo vom CERN. Es gibt noch weitere Repositorien, die genutzt werden können und auch hierauf werden wir insbesondere im Kapitel über die dauerhafte Speicherung von Daten näher eingehen. Falls der Zugang zu Forschungsdaten und deren Metadaten nur eingeschränkt möglich sein soll, dann müssen hierfür die notwendigen Regularien und Schutzmechanismen festgelegt und geschaffen werden.

2.2.3. Verarbeitbarkeit

Daten können leichter verarbeitet werden, wenn sie und ihre Metadaten etablierten Standards gehorchen. Das können zum Beispiel gängige Dateiformate wie Tabellen im CSV-Format sein. Das Dateiformat CSV (Comma-separated values) beschreibt den Aufbau einer Textdatei, mit der einfach strukturierte Daten gespeichert oder ausgetauscht werden können. Bevorzugt sollten freie und allgemeine Dateiformate wie einfache Textdateien (ASCII-Code) genutzt werden. Die Verwendung propietärer Dateiformate schränkt die Nutzungsmöglichkeiten automatisch für eine Reihe potentieller Nutzer ein. Für Bilddateien sind die Dateiformate JPG, PNG, BMP oder TIFF sehr populär (je nach Anwendung), für Audio- und Videodateien die Formate MP3 und MP4. Die Aufzählung erhebt allerdings keinen Anspruch auf Vollständigkeit. Dieser Ratgeber wurde zum Beispiel mit dem Textsatzsystem LaTeX im frei verfügbaren TeX-Format geschrieben, welches mit einem beliebigen Editor angesehen und verändert werden kann.

2.2.4. Nach-Nutzbarkeit

Es kann sinnvoll sein, bestimmte Lizenzen für die Nach-Nutzung von Forschungsdaten festzulegen (z.B. Creative Commons CC BY 4.0 International), damit einmal publizierte Daten auch möglichst frei und global genutzt werden können, soweit das gewünscht ist. Machen Sie auch deutlich, ob die Nutzung zeitlich befristet oder dauerhaft möglich ist und ob ggf. Embargofristen eingehalten werden müssen. Dieses kann bei Industriekooperationen oder in Forschungsverbünden der Fall sein, insbesondere wenn ein Projekt noch nicht abgeschlossen ist. Dokumentieren sie idealerweise auch, wie die Datenqualität überprüft und sichergestellt worden ist. Diesen Punkt werden wir in den folgenden Kapiteln zur Datenqualität und Datennachverfolgung noch intensiv diskutieren. Das ist ein Kernthema des gesamten Forschungsdatenmanagements und sollte intensive Beachtung finden.

Zusammenfassung:

FAIRe Forschungsdaten müssen auffindbar, zugänglich, verarbeitbar und wiederverwendbar sein. Voraussetzung dafür sind gut strukturierte Metadaten und eine eindeutige Kennzeichnung von Datensätzen. Es muss festgelegt werden, wo und wie Forschungsdaten, die zugehörigen Metadaten, die notwendigen Dokumentationen und ggf. auch Software (Quellcodes) gespeichert sind.

3. Rechtliche Aspekte

Da es sich bei den Autoren dieses Ratgebers nicht um Juristen sondern um Naturwissenschaftler handelt, können an dieser Stelle auch keine rechtsverbindlichen Informationen vermittelt werden. In den folgenden Abschnitten werden frei verfügbare juristische Informationen kondensiert wiedergegeben, die als unverbindliche Hinweise angesehen werden sollten. Im Zweifelsfall sollte die juristische Abteilung und/oder der Datenschutzbeauftragte der jeweiligen Forschungseinrichtung konsultiert werden, bevor kritische Systeme in Betrieb genommen werden oder die Verarbeitung von Forschungsdaten nicht gesichert und eindeutig rechtskonform ist. Im Folgenden wird ausführlicher erläutert, in welchen Fällen besonders auf den rechtskonformen Umgang mit Forschungsdaten geachtet werden muss.

3.1. Urheberrechte bei der Nutzung fremder Daten

Nutzungs- und Urheberrechte sowie Recht auf fremdes geistiges Eigentum und Patentrechte müssen bereits vor der Nutzung fremder Daten geprüft werden. Die rechtlichen Verpflichtungen beziehen sich übergreifend auf die Aufnahme, Dokumentation, Speicherung, Archivierung und Nachnutzung verwendeter Daten [**DFG**]. Informationen selbst sind nicht durch das Urheberrecht geschützt. Das Gleiche gilt für Thesen und Lehrmeinungen, damit sie nicht durch das Urheberrecht einem Monopol unterliegen können, und somit eine freie wissenschaftliche Diskussion gewährleistet bleibt [**Lauber**]. Forschungsdaten können allerdings durchaus durch Urheber- oder Leistungsschutzgesetze geschützt sein. Nach §2 Abs. 1 UrhG kommt ein urheberrechtlicher Schutz in Betracht, sofern eine *persönliche geistige Schöpfung* vorliegt (§2 Abs. 2 UrhG). Hierfür muss eine Leistung besondere Individualität aufweisen [**Lauber**]. Rein handwerkliche, routinemäßige Leistungen sind damit ausgeschlossen sowie fachwissenschaftliche Gepflogenheiten. Es muss ausdrücklich ein Gestaltungsspielraum des Wissenschaftlers bestanden haben. Für detaillierte Informationen verweisen wir an dieser Stelle auf [**Lauber**, **Baumann**].

3.1.1. Text und Data Mining

Unter Text und Data Mining versteht der Gesetzgeber die *automatisierte Analyse von einzelnen oder mehreren digitalen oder digitalisierten Werken, um daraus Informationen insbesondere über Muster, Trends und Korrelationen zu gewinnen* (§ 44b Abs. 1 UrhG) [**Brehm**]. Das Gesetz bezieht sich damit grundsätzlich auf sämtliche urheberrechtlichen Schutzgegenstände (Texte, Grafiken, Bilder, Audioaufnahmen, Musik, Daten, Datenbanken, etc.). Natürlich können auch urheberrechtlich nicht geschützte Objekte Gegenstand von Text und Data Mining sein. Eine urheberrechtliche Erlaubnis ist dafür grundsätzlich nicht erforderlich. Umfangreiche Datenbanken, die in der Erstellung sehr aufwändig waren, sind hingegen unabhängig vom urheberrechtlichen Schutz der Inhalte selbst geschützt. Diese Darstellung ist auf wissenschaftliche Textpublikationen beschränkt [**Brehm**].
Umfangreiche Information finden sich in den Guidelines von Elke Brehm [**Brehm**]. Dort beschreibt sie ausführlich, unter welchen Bedingungen Text und Data Mining zu wissenschaftlichen Zwecken

bei wissenschaftlichen Publikationen auf der Basis von sog. Schrankenregelungen und/oder Verträgen durchgeführt werden darf und welche Risiken dabei bestehen können.

3.2. Verarbeitung personenbezogener Daten

Das Datenschutzrecht kommt dann zur Anwendung, wenn personenbezogene Daten verarbeitet werden. Das sind nach der Datenschutzgrundverordnung (Art. 4 Nr.1 DSGVO) alle Informationen, die sich auf eine identifizierte oder identifizierbare Person beziehen [**Lauber**]. Seit Einführung der DSGVO gibt es eine kontinuierliche Diskussion darüber, was in der Fotografie bzw. allgemein beim Umgang mit Bilddaten erlaubt ist und was nicht. Ein Personenbezug kann zum Beispiel auch bei Fotos mit unkenntlich gemachten Gesichtern hergestellt werden, wenn aufgrund des Hintergrunds, der Kleidung und Haltung der abgebildeten Personen sowie begleitender Informationen über Zeitpunkt und Ort der Aufnahme eine Identifizierung möglich ist [**Lauber**]. Bei medizinischen Forschungsdaten können in den Metadaten neben den Patienten- oder Probanden-Daten auch Informationen über die beteiligten Forscher enthalten sein. Auch hier greift die DSGVO, da es sich ebenfalls um personenbezogene Daten handelt.
In diesem Zusammenhang ist beim Einsatz von elektronischen Laborbüchern (ELN) darauf zu achten, das über die gespeicherten personenbezogenen Daten der Forscher in öffentlichen Einrichtungen keine Leistungskontrolle erfolgen darf. Im Idealfall wird durch eine Betriebsvereinbarung zwischen dem Betriebsrat und dem Arbeitgeber genau geregelt, wie entsprechende Daten zu behandeln sind [**ZB**, **Corti**, **Johannes**].
Gemäß § 75 Abs. 3 Nr. 17 BPersVG unterliegt die Einführung und Anwendung technischer Einrichtungen, die dazu bestimmt sind, das Verhalten oder die Leistung der Beschäftigten zu überwachen, der Mitbestimmung des Betriebsrats [**Bremecker**]. Als technische Geräte im Sinn eines Mitbestimmungstatbestands kommen allgemein auch Datenverarbeitungssysteme in Betracht. Die technische Einrichtung muss dazu bestimmt sein, das Verhalten und die Leistung der Beschäftigten zu überwachen. Nach der Rechtsprechung des Bundesverwaltungsgerichts in Leipzig (BVerwG) ist dieser Tatbestand bereits erfüllt, wenn die technische Einrichtung objektiv geeignet ist, Verhalten oder Leistung der Beschäftigten zu überwachen. Es reicht also aus, wenn die Einrichtung zunächst anderen Zwecken dienen soll und eine Kontrollabsicht nicht vorgesehen ist. Allein die technischen Möglichkeiten (Hard- und Software) für eine potentiell mögliche Überwachung bzw. Kontrolle machen eine Mitbestimmung erforderlich.
Folgt man der Rechtsauffassung von Paul C. Johannes [**Johannes**], dann sollte besonders die Freiheit der wissenschaftlichen Forschung und deren Akteure im Vordergrund stehen. Nach Art. 5 III GG (Grundgesetz) gilt die wissenschaftliche Freiheit als besonders schutzwürdig.
Im übernächsten Kapitel werden wir ausführlich auf die technischen bzw. digitalen Möglichkeiten beim Einsatz eines ELN eingehen und es wird dabei schnell deutlich werden, das mit einem ELN eine Leistungskontrolle durchaus möglich ist.

3.3. Rechtliche Rahmenbedingungen für die Weitergabe von Daten

Die Bedingungen, unter denen Forschungsdaten zur Nachnutzung freigegeben werden, sollten wenig restriktiv und möglichst transparent sein [**Lauber**]. Da nach deutschen Recht nicht vollständig

auf Urheberrechte verzichtet werden kann, was auch für Forschungsdaten gilt, muss man sich mit Lizenzverträgen behelfen. Umfassende, vergütungsfreie Nutzungsrechte werden den Nutzern durch sogenannte freie Lizenzen eingeräumt.
Repositorien wie z.B. Zenodo verwenden dafür Vertragsmuster. Weit verbreitet sind auch Creative-Commons-Lizenzen. Die Europäische Kommision empfiehlt die Lizenztypen CC-BY und CCO [**EU**]. In der folgenden Tabelle sind die wichtigsten Informationen zu diesen beiden Lizenztypen zusammengefasst. Ausführlicher werden weitere mögliche Lizenzen in [**Lauber**] dargestellt. Für Forschungsdaten können als Lizenzmodell auch die Open Data Commons (ODC: `https://opendatacommons.org/`) in Betracht kommen.

Tabelle 3.1.: Creative Commons Lizenzen.

Lizenz	Erlaubt:	Bedingung:
CC BY	Vervielfältigung, Weitergabe, Erstellung von Bearbeitungen sowie deren Vervielfältigung und Weitergabe für kommerzielle und nicht-kommerzielle Zwecke	Namensnennung: Bezeichnung des Erstellers; Nennung des Lizenztyps und Referenz auf Lizenztext durch URI/Hyperlink; URI/Hyperlink zum lizenzierten Material; Copyright-Vermerk, Hinweis auf Haftungsausschluss; Hinweis, wenn lizenziertes Material verändert wurde.
CCO	z.T. Verzicht auf Urheberrecht; da nach UrhG nicht möglich, weitestmögliche Einräumung von Nutzungsrechten.	Grundsätzlich keine Namensnennung erforderlich.

Zusammenfassung:

In allen Bereichen des Forschungsdatenmanagements sind rechtliche Rahmenbedingung zu beachten, insbesondere bei der Nutzung fremder Daten, bei der Verarbeitung personenbezogener Daten und bei der Weitergabe von Forschungsdaten. Im Zweifelsfall sollte immer die juristische Abteilung oder der Datenschutzbeauftragte der jeweiligen Einrichtung konsultiert werden.

4. Der Datenmanagementplan

Der Datenmanagementplan (DMP) ist zunächst ein formales Dokument, das den Umgang mit Forschungsdaten über den gesamten Lebenszyklus (siehe Kapitel 2) beschreibt, und zwar von der Projektvorbereitung, während der Projektlaufzeit und auch darüber hinaus, zum Beispiel als Planungsbasis für weitere Projekte. Der DMP sollte ein selbstverständlicher Teil der Projektplanung sein und während der Projektlaufzeit aktualisiert werden.
Inzwischen erwarten die meisten Drittmittelgeber einen DMP als Teil eines Projektantrages (siehe Kapitel 1) und es gibt erste verifizierte Fälle, in denen ein Projektantrag wegen eines fehlenden DMP abgelehnt wurde. Eine Übersicht zum Thema DMP findet sich unter anderem auf den Seiten von (`https://forschungsdaten.info/`).

Es bleibt bislang allerdings jedem selbst überlassen, welchen Aufwand man für die Erstellung und Pflege eines DMP betreiben möchte. Eine sehr umfängliche, aber auch recht aufwändige Methode ist der Einsatz von RDMO (Research Data Management Organiser; `https://rdmorganiser.github.io/`). RDMO ist eine freie Software, die zur Planung, Umsetzung und Verwaltung des Forschungsdatenmanagements eingesetzt werden kann.
Unter `https://rdmo.aip.de/` steht eine Demo-Instanz für ein erstes Ausprobieren zur Verfügung. Diese Software muss allerdings auf einem Server installiert sein und die Accounts werden zentral verwaltet. Idealerweise besitzt das betreibende Institut eine eigene RDMO-Instanz, bei der man Administratorrechte besitzt. Kosten und Aufwand für den Betrieb und die Pflege von RDMO sind unvermeidlich die Folge.

Die Praxis hat jedoch gezeigt, dass, selbst wenn die notwendige Infrastruktur zur Verfügung gestellt wird, so nutzen doch nur vergleichsweise wenige Wissenschaftler diese Möglichkeit. Deutlich einfacher ist dagegen die Erstellung eines DMP als einfache Textdatei, die sukzessive und einfach modifiziert werden kann. Für die meisten Drittmittelgeber ist diese einfache Lösung ausreichend und wird bei Projektanträgen akzeptiert.
Wir werden daher die notwendigen Informationen für das Schreiben eines einfachen DMP im Folgenden kurz skizzieren und mit Beispielen veranschaulichen. Diese Beschreibung orientiert sich an den Empfehlungen der DFG (`https://www.dfg.de/antragstellung/forschungsdaten`).
Die folgenden Fragen sollten möglichste präzise beantwortet werden, damit ein DMP nachvollziehbar wird. Gleichzeitig helfen diese Fragen bei der Projektplanung und erleichtern die Strukturierung von Projekten, denn überall dort, wo in einem Projekt Daten entstehen und verarbeitet werden, müssen auch entsprechende Experimente oder Simulationen durchgeführt werden. Bei der Planung der Datenströme wird automatisch eine sinnvolle zeitliche Abfolge des gesamten Projektes ersichtlich.

4.1. Datenbeschreibung

Auf welche Weise entstehen in Ihrem Projekt neue Daten? Werden existierende Daten wiederverwendet? Welche Datentypen, im Sinne von Datenformaten (z. B. Bilddaten, Textdaten oder Messdaten) entstehen in Ihrem Projekt und auf welche Weise werden sie weiterverarbeitet? In welchem Umfang fallen diese an bzw. welches Datenvolumen ist zu erwarten?

Beispiel:

Für das Vorhaben sind nach Recherchen in gängigen Datenrepositorien keine aktuellen bzw. geeigneten Forschungsdaten zur Nachnutzung verfügbar. Die im Projekt erzeugten Daten werden weitere Erkenntnisse im Bereich XY ermöglichen. Die erzeugten Datensätze werden durch das Projektteam mit unterschiedlichen Analyseverfahren, primär REM, XRD sowie TEM erstellt. Hauptsächlich fallen textuelle, tabellarische und Bilddaten an. Diese werden nach Möglichkeit in offenen Formaten gespeichert (Textuelle Daten txt, rtf, pdf; Tabellarische Daten csv; Bilddaten tiff). Während der Projektlaufzeit werden Analysen und Auswertungen mit der frei verfügbaren Programmiersprache Python sowie deren etablierten Open-Source-Bibliotheken durchgeführt. Das erwartete Datenvolumen wird maximal 100 GB betragen.

4.2. Dokumentation und Datenqualität

Welche Ansätze werden verfolgt, um die Daten nachvollziehbar zu beschreiben (z.B. Nutzung vorhandener Metadaten- bzw. Dokumentationsstandards oder Ontologien)? Welche Maßnahmen werden getroffen, um eine hohe Qualität der Daten zu gewährleisten? Sind Qualitätskontrollen vorgesehen und wenn ja, auf welche Weise? Welche digitalen Methoden und Werkzeuge (z.B. Software) sind zur Nutzung der Daten erforderlich?

Beispiel:

Die erzeugten Forschungsdaten und Skripte werden im Repositorium JülichDATA veröffentlicht. Im Sinne der FAIR-Prinzipien werden die Daten im Repositorium durch Metadaten beschrieben, orientiert am DataCite-Schema (u.a. Abstract, freie Schlagwörter und DDC-Klassifikation). Für die Datendokumentation wird das elektronische Laborbuch eLabFTW eingesetzt. Den Metadaten wird durch das Repositorium ein persistenter Identifier (DOI) hinzugefügt, der den Datensatz eindeutig referenzierbar macht. Die Benennung von Dateien und Verzeichnissen erfolgt nach einem einheitlichen Schema, z.B. werden Datumsangaben nach ISO 2014 formatiert:[JJJJ]-[MM]-[TT]. Das Schema wird zu Projektbeginn mit allen Projektmitarbeitern gemeinsam festgelegt. Die Dokumentation der tabellarischen Daten (CSV) erfolgt durch ein bzw. mehrere sogenannte Tabular Data Packages. In Form einer solchen Spezifikation erfolgt eine Dokumentation der Daten, der einzelnen Spalten (Variablen), deren erlaubten Datentypen und Wertebereichen sowie eine Beziehung von Spalten untereinander (auch über einzelne Dateien hinweg). Durch die formalisierte Datenbeschreibung und -dokumentation wird eine tool-basierte Qualitätskontrolle ermöglicht und regelmäßig durchgeführt (z.B. liegen alle Werte konkreter Spalten in erlaubten Wertebereichen). Die Validierung und Qualität der Daten erfolgt durch regelmäßige Messungen von Standardproben mit bekannten Eigenschaften an den verschiedenen Charakterisierungsanlagen. Die Nutzung der Daten und Skripte ist durch Open-Source-Standardtools möglich. Kosten für spezialisierte Software zum Lesen/Bearbeiten/Ausführen der Dateien fallen nicht an.

4.3. Speicherung und technische Sicherung während des Projektverlaufs

Auf welche Weise werden die Daten während der Projektlaufzeit gespeichert und gesichert? Wie wird die Sicherheit sensibler Daten während der Projektlaufzeit gewährleistet (Zugriffs- und Nutzungsverwaltung)?

Beispiel:

Das erwartete Datenvolumen von maximal 100 GB wird durch einen wissenschaftlichen Speicherbereich (im Folgenden als *Projektnetzlaufwerk* bezeichnet) des Forschungszentrum Jülich bereitgestellt. Während der Projektlaufzeit werden Daten und Metadaten auf dem Projektnetzlaufwerk gespeichert. Das Laufwerk wird von allen Projektmitarbeitern als Netzlaufwerk über das jeweilige Betriebssystem eingebunden. Das Projektnetzlaufwerk unterliegt einer automatisierten, regelmäßigen, dateibasierten Backup-Routine durch das Rechenzentrum. Die Daten werden regelmäßig und automatisch auf einem Fileserver gesichert. Im Falle, dass Daten und Skripte lokal auf den Arbeitsrechnern der Projektgruppe erzeugt werden, synchronisieren die Mitarbeiter diese einmal wöchentlich mit dem Projektnetzlaufwerk, um Datenverlust vorzubeugen. Hierzu werden die für das jeweilige Betriebssystem etablierten Open-Source-Tools eingesetzt. Da keine sensiblen Daten erhoben werden, erfolgt keine gesonderte Zugriffs- und Nutzungsverwaltung. Der Zugriff zum Projektnetzlaufwerk wird zentral durch das Rechenzentrum des Forschungszentrum Jülich verwaltet (Zugriff haben nur Mitglieder der Projektgruppe). Ein Zugriff auf die Daten durch Dritte ist während der Projektlaufzeit nicht erforderlich.

4.4. Rechtliche Verpflichtungen und Rahmenbedingungen

Welche rechtlichen Besonderheiten bestehen im Zusammenhang mit dem Umgang mit Forschungsdaten in Ihrem Projekt? Sind Auswirkungen oder Einschränkungen in Bezug auf die spätere Veröffentlichung bzw. Zugänglichkeit zu erwarten? Auf welche Weise werden nutzungs- und urheberrechtliche Aspekte sowie Eigentumsfragen berücksichtigt? Existieren wichtige wissenschaftliche Kodizes bzw. fachliche Normen, die Berücksichtigung finden sollten?

Beispiel:

In Bezug auf die Daten liegen keine rechtlichen Besonderheiten vor. Die Nachnutzung von Software anderer Urheber wird im Sinne der guten wissenschaftlichen Praxis zitiert.
Anmerkung: An dieser Stelle können ggf. weitere rechtliche Absicherungen eingesetzt werden, wie sie in Kapitel 3 beschrieben worden sind.

4.5. Datenaustausch und dauerhafte Zugänglichkeit der Daten

Welche Daten bieten sich für die Nachnutzung in anderen Kontexten besonders an? Nach welchen Kriterien werden Forschungsdaten ausgewählt, um diese für die Nachnutzung durch andere zur Verfügung zu stellen? Planen Sie die Archivierung Ihrer Daten in einer geeigneten Infrastruktur? Falls ja, wie und wo? Gibt es Sperrfristen? Wann sind die Forschungsdaten für Dritte nutzbar?

Beispiel:

Die erhobenen Daten und Skripte bieten sich für die Nachnutzung durch Dritte an. Daher werden Metadaten sowie teilweise Daten im Forschungszentrum eigenen Repositorium JülichDATA veröffentlicht. In der Veröffentlichung inbegriffen sind sämtliche erzeugten Rohdaten und Skripte sowie finale Versionen von Textdaten und Tabellen. Außerdem wird der Veröffentlichung eine Dokumentation beigefügt. Zwischenergebnisse von Verarbeitungs- und Analyseschritten, die sich sämtlich aus den bereitgestellten Daten und Skripten erzeugen lassen, sind nicht Teil der Veröffentlichung. Die Veröffentlichung folgt den Empfehlungen der Open-Access-Policy und Forschungsdaten-Policy des Forschungszentrums. Durch die Verwendung des elektonischen Laborbuchs eLabFTW und des Repositoriums JülichDATA werden mehrere der in den FAIR-Prinzipien adressierten Punkte sichergestellt. So werden die Metadaten über standardisierte Schnittstellen (wie OAI-PMH) in übergreifenden Nachweissystemen und Suchmaschinen indexiert (z.B. BASE, DataCite Search, OpenAIRE). Dadurch wird eine erhöhte Sichtbarkeit der Forschungsergebnisse erreicht. Die erstellten Metadaten werden durch die Redaktion des Repositoriums geprüft. Des Weiteren wird ein DOI für zu publizierende Datensätze vergeben. Im Sinne der Leitlinien zur Sicherung guter wissenschaftlicher Praxis werden die Daten für mindestens zehn Jahre durch das Repositorium öffentlich, d.h. ohne Zugangsbeschränkung, bereitgestellt. Eine separate Archivierung, unabhängig von der Veröffentlichung, ist nicht vorgesehen. Eine Sperrfrist ist nicht erforderlich. Die Veröffentlichung findet so schnell wie möglich, spätestens jedoch innerhalb der letzten drei Monate der Projektlaufzeit statt.

4.6. Verantwortlichkeit und Ressourcen

Wer ist verantwortlich für den adäquaten Umgang mit den Forschungsdaten (Beschreibung der Rollen und Verantwortlichkeiten innerhalb des Projekts)? Welche Ressourcen (Kosten; Zeit oder anderes) sind erforderlich, um einen adäquaten Umgang mit Forschungsdaten im Projekt umzusetzen? Wer ist nach Ende der Laufzeit des Projekts für das Kuratieren der Daten verantwortlich?

Beispiel:

Hauptverantwortlich für den Umgang mit den im Projekt erzielten Forschungsdaten ist Herr/Frau XY. Die Einhaltung und Aktualisierung des DMP wird durch diese Person sichergestellt. Eine fortlaufende Dokumentation und Aufbereitung der Daten und Skripte erfolgt bereits während der Projektlaufzeit; ihre Finalisierung erfolgt in den letzten drei Monaten. Nach Ende der Projektlaufzeit werden sämtliche zur Veröffentlichung vorgesehenen Daten am angegebenen Ort publiziert. Über die Laufzeit hinaus findet keine weitere Kuratierung der Daten statt.

Zusammenfassung:

Das Schreiben und die fortlaufende Pflege eines DMP erscheinen ggf. zunächst als nicht unbedingt notwendiger Mehraufwand bzw. die Vorteile mögen nicht direkt ersichtlich sein. Tatsächlich ist ein effizient geschriebener DMP (siehe oben) eine sinnvolle Hilfe bei der Planung und Durchführung eines Projektes. Auch die Generierung FAIRer Forschungsdaten wird durch einen DMP deutlich erleichtert.

5. Die Datenerfassung, Datenspeicherung und Dokumentation

5.1. Elektronische Laborbücher

Immer häufiger werden in Laboren Papier-Laborbücher durch elektronische Laborbücher ersetzt. Bei diesem Umstieg spielt es nicht nur eine Rolle, Papier durch eine digitale Anwendung zu ersetzen. Ebenso wichtig ist auch die Möglichkeit, die elektronische Form des Laborbuches in das Gesamtsystem eines digitalen Forschungsdatenmanagements (FDM) zu integrieren [**ZB**]. Im Lebenszyklus von Forschungsdaten hat das ELN eine hohe Bedeutung im Rahmen der Dokumentationsphase. Bei allen Überlegungen im Kontext der Anschaffung und Nutzung eines ELN sollten zu Beginn einige grundsätzliche Fragen stehen:
Wie gestaltet sich der gesamte Workflow für Forschungsdaten im Lebenszyklus? Welche IT-Anwendungen oder Tools sollen in welchem Schritt genutzt werden? Welche Funktion soll ein ELN in diesem Gesamtkontext erfüllen?
Neben der Gestaltung eines institutionellen Forschungsdatenmanagements kann ein ELN auch einen wesentlichen Beitrag zur guten wissenschaftlichen Praxis leisten, da durch seinen Gebrauch Forschungsprozesse und Forschungsergebnisse besser nachvollziehbar werden.
Der Einsatz eines ELN bietet gegenüber der traditionellen Papierform eine Reihe von entscheidenden Vorteilen, welche die Arbeit im Labor oder bei Computer-Simulationen deutlich effizienter werden lassen und auch die Datensicherheit erhöhen [**ZB**]:

- Direkte Einbindung/Verlinkung bereits digital vorliegender Daten (z.B. Messergebnisse, Bild-, Video-, Audiodateien, Texte, Tabellen).
- Kein Informationsverlust durch unleserliche Handschriften.
- Such- und Filterfunktionen.
- Funktionen für das kollaborative Arbeiten (Rechte-, Rollenmanagement).
- Team- und Labororganisation.
- Erstellung und Verwendung von Vorlagen (Templates, z.B. für sich wiederholende Prozesse).
- Einbettung in eine vernetzte digitale Forschungsumgebung, Schnittstellen, Import- und Exportfunktionen, Anbindung an Repositorien, Langzeitarchivierung, u.a.).
- Schnellere und leichtere Publikation von Datensätzen.

5.2. Populäre elektronische Laborbücher

In den beiden folgenden Tabellen sind eine Reihe aktuell populärer freier (Open-Source) und kommerzieller elektronischer Laborbücher aufgelistet. Diese Tabellen erheben keinen Anspruch auf Vollständigkeit. Ständig entstehen neue ELN's, die häufig an die Bedürfnisse bestimmter Fachrichtungen angepasst werden. Weitere Hinweise auf existierende ELN's findet man im ELN Finder der Universitäts- und Landesbibliothek der TU Darmstadt (`https://eln-finder.ulb.tu-darmstadt.de/home`), bei Wikipedia (`https://en.wikipedia.org`) oder über eine Vielzahl weiterer Quellen im Internet. Aktuell gibt es neue Entwicklungen, bei denen große Sprachmodelle (LLM - Large Language Model) für die Verwendung mit elektronischen Laborbüchern mit KI konzipiert und programmiert werden. Da sich diese Entwicklungen aber noch ganz am Anfang befinden und derzeit noch nicht als ausgereift bezeichnet werden können, möchten wir an dieser Stelle lediglich auf die entsprechende Literatur verweisen [**Jalali**]. Sollten LLM's bei elektronischen Laborbüchern zukünftig eine weitere Verbreitung finden, was bei geschlossenen ELN-Systemen nicht ganz unkritisch und unproblematisch bezüglich der Datensicherheit sein dürfte, dann könnten sich neue Möglichkeiten in der beschleunigten Datenerfassung und dem Auffinden vorhandener Daten ergeben. Auch die Übergabe von Daten direkt an eine KI zur Datenanalyse und Weiterverwendung könnte auf diesem Wege vereinfacht und beschleunigt werden. Erste Ansätze dazu gibt es bereits, auch in dem im Folgenden beschrieben ELN Kadi4Mat, dort allerdings noch ohne LLM.

Tabelle 5.1.: Freie elektronische Laborbücher (Open-Source).

Name	Webadresse
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/`

Tabelle 5.2.: Kommerzielle elektronische Laborbücher.

Name	Webadresse
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/

Für die Korrektheit und die Sicherheit der in den Tabellen aufgeführten Weblinks (URL's) sowie auch für die Inhalte der verlinkten Webseiten übernehmen wir in diesem Handbuch ausdrücklich keinerlei Verantwortung.

5.3. Elektronische Laborbücher in der Praxis

5.3.1. JuliaBase

Ihr wissenschaftliches Institut oder Ihre Arbeitsgruppe erstellt viele Proben, und Ihr Team braucht ein Werkzeug, um den Überblick zu behalten? JuliaBase ist an genau so einem Institut entstanden. Es ist eine Datenbanklösung für Proben, ihre Prozessierung und Charakterisierung, mit den folgenden Funktionen:

- vollständig quelloffen
- browserbasierte Schnittstelle, die auch auf mobilen Geräten funktionsfähig ist
- hohe Flexibilität zur Anpassung an vorhandene Produktions- und Messeinrichtungen und an Arbeitsabläufe
- fein abgestufte Zugriffskontrolle
- die Möglichkeit, mehr als eine Abteilung getrennt in einer einzigen Datenbank zu verwalten
- stellt eine Verbindung zu Ihrem LDAP-Server für die Benutzerverwaltung her

- Spaltung von Proben werden sauber nachgehalten, damit auch Daten von Mutter- und Tochterstücken stets sichtbar sind
- Unterstützung für die Vorabauswertung von Rohdaten und die Visualisierung von Daten
- automatische Benachrichtigung über Änderungen an Proben
- Gruppierung nach Probenserien, Themen und Tags
- komplexe Suchen leicht gemacht, z.B. „finde alle Proben mit Infrarotmessungen, die zusammen mit einer Probe auf einem Glassubstrat abgelagert wurden mit einer Leitfähigkeit größer als 10–6 S/cm; ach ja, und nur aus diesem Jahr und hergestellt von John“
- Export in Tabellenkalkulationen
- automatische tabellarische Laborbücher
- Datenbankinteraktion aus eigenen Programmen, z.B. zur direkten Anbindung eines Messaufbaus an die Datenbank
- vollständig übersetzbar; der Kern ist bisher in Englisch und Deutsch verfügbar
- Layout kann an Corporate Identity angepasst werden
- ausgereifte Codebasis seit 2008
- Einhaltung moderner Web- und Sicherheitsstandards

JuliaBase verfolgt den Ansatz, dass sich die Datenbank den existierenden Arbeitsabläufen anpassen sollte und nicht umgekehrt.
Allerdings hat die Flexibilität von JuliaBase ihren Preis: Es muss Python-Code erstellt werden für jede Prozess-Art, die man einbinden will. Typischerweise ist das zwar nur bis zu 100 Zeilen Code für jeden Prozess, und JuliaBase enthält sogar Code für typische Verarbeitungs- und Messaufbauten, die man als Ausgangspunkt verwenden kann. Dennoch ist diese Arbeit notwendig.

Ein Spaziergang durch JuliaBase

Auf `https://demo.juliabase.org` ist eine Demo von JuliaBase installiert, damit man ein wenig damit spielen kann. Es lassen sich Proben, Prozesse, Aufgaben usw. hinzufügen, Beispieldatenblätter oder ein Labornotizbuch ansehen und vieles mehr. Man kann sich anmelden mit verschiedenen Konten, um unterschiedliche Berechtigungen (Rollen) auszuprobieren.
Die Demoseite ist die JuliaBase-Installation des *Institute of Nifty New Materials (INM)*. Es ist ein sehr kleines Institut mit nur sechs Mitarbeitern. Alle Konten haben das Passwort *12345*.

Die Demo-Konten:

Sean Renard (s.renard) ist der leitende Wissenschaftler und Direktor dieses Instituts. Dementsprechend erlaubt ihm sein JuliaBase-Konto, alle Proben zu sehen, aber er hat auch noch andere Privilegien. Dazu später mehr.
Nick Burkhardt (n.burkhardt) ist schon sehr lange Techniker im INM. Er ist verantwortlich für den PDS-Aufbau (photothermische Ablenkungsspektroskopie), ein Messgerät. Er führt Messungen für Forscher durch. Er würde niemals einer anderen Person erlauben, seine PDS zu benutzen.

Hank Griffin (h.griffin) ist ebenfalls ein Techniker. Er ist für den Solarsimulator, eine weitere Messeinrichtung, verantwortlich. Er führt Messungen für Forscher durch, aber nach entsprechender Einweisung durch ihn können auch andere Personen das Gerät benutzen.
Eddie Monroe (e.monroe) ist ein Operateur. Dies ist ein Techniker, der ein Abscheidungssystem betreibt – in seinem Fall die Cluster-Tool-Deposition. Mit solchen Prozessen werden Proben hergestellt. Auch hier können andere Institutsmitglieder nach entsprechender Einweisung dieses System benutzen.
Rosalee Calvert (r.calvert) ist fest angestellte Wissenschaftlerin und erstellt selbst Proben in der 5-Kammer-Beschichtungsanlage. Derzeit ist sie die Einzige, die diese Anlage benutzt. Anschließend misst sie die Proben im Solarsimulator. Ihr aktuelles Projekt ist eine Kooperation mit der Universität von Paris.
Juliette Silverton (j.silverton) ist eine Doktorandin, die viel zu tun hat. Daher kann sie die Probenvorbereitung und die Messungen nicht selbst durchführen, sondern muss dieses anderen überlassen. Sie nutzt intensiv die Aufgabenlisten-Funktion von JuliaBase, um diese Arbeiten in Auftrag zu geben. Im Folgenden werfen wir einen genaueren Blick auf die typischen Arbeitsabläufe der Personen.

Rosalee: Typischer normaler Benutzer

Melden Sie sich als r.calvert an, d.h. als der typische normale Benutzer (Abbildung 5.1). Im Haupt-

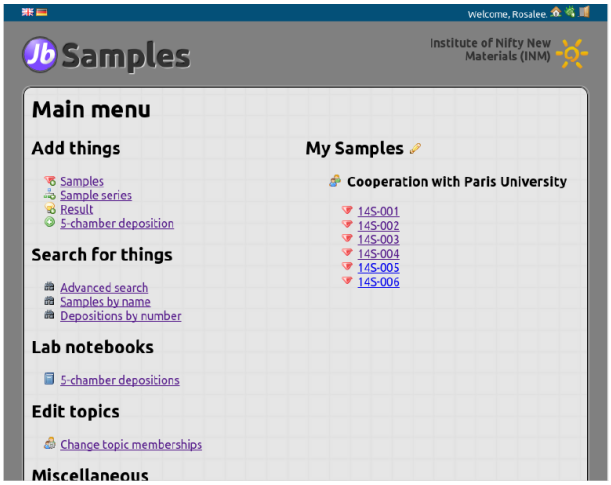


Abbildung 5.1.: Die tägliche Arbeit.

menü können Sie Rosalees *Meine Proben* sehen. Diese Liste enthält normalerweise nicht alle Proben eines Benutzers, sondern nur die, die für ihn/sie gerade interessant sind. Dennoch kann diese Liste recht lang werden und ist daher nach Themen und Probenserien gegliedert. Sie können auf die Icons vor den Serien- oder Themennamen klicken, um Abschnitte, die Sie ausblenden möchten, ein- und auszuklappen. Eine Probe gehört normalerweise zu genau einem Thema. Dies hilft, die Proben zu organisieren. (In der Liste von Rosalee gehören alle Proben zum Thema Zusammenarbeit mit der Universität Paris.) Zum einen kann man Themen aussagekräftige Namen geben, die den Zweck der Proben deutlich machen. Aber noch wichtiger ist, dass Themen festlegen, wer die Probe sehen kann. Die wichtigste Richtlinie in JuliaBase lautet: Sie können nur Proben aus Ihren Themen sehen. Personen können in beliebig vielen Themen gleichzeitig sein, aber eine Probe ist in genau einem Thema. Sie kann es während ihrer Lebenszeit wechseln. Leitende Teammitglieder haben u.U. die Berechtigung, *alle* Proben zu sehen. Sean Renard ist eine solche Person.

Probendatenblatt:

Schauen wir uns ein Beispiel an, indem wir auf *14S-001* klicken (Abbildung 5.2). Sie sehen das *Datenblatt* der Probe. Oben finden Sie einige allgemeine Informationen über die derzeit verantwortliche Person und das Thema. Dann sehen Sie eine Liste aller Arbeiten, die mit dieser Probe durchgeführt wurden, in chronologischer Reihenfolge. Es beginnt mit dem Substrat, wird mit der Abscheidung der Siliziumschichten fortgesetzt und endet mit einer Messung im Solarsimulator. Jeder dieser Schritte wird in JuliaBase als „Prozess“ bezeichnet. Sogar das Ausgangssubstrat ist ein Prozess, wenn auch nicht im wörtlichen Sinne. Jeder Prozess hat einen Operator und einen Zeitstempel. Sie können Prozesse einklappen, indem Sie auf die Überschrift klicken. Die Hauptarbeit bei der Anpassung von JuliaBase an ein neues Institut ist das Anlegen aller Prozesse, die das Institut benötigt, im Python-Code. Die Solarsimulator-Messung unten ist ein gutes Beispiel dafür, warum sich der Aufwand lohnt: Klicken Sie einfach auf die farbigen Quadrate, und Sie sehen sofort, wie sich die Daten und die Darstellung ändern. Solche Funktionen fehlen den meisten anderen ELNs. Dieses hohe Maß an Anpassungsfähigkeit und Flexibilität ist die Hauptstärke von JuliaBase. Blättern wir zurück an den Anfang. Sie sehen einen schematischen Querschnitt der Probe. Auch das ist eine Erweiterung für das INM (die nachnutzbar für andere Institute ist). Wenn Sie auf den Querschnitt klicken, erhalten Sie ihn als PDF. Das gilt auch für alle Plots in JuliaBase.

Proben bearbeiten:

Sie können eine Probe bearbeiten, indem Sie auf das Stiftsymbol neben dem Namen der Probe klicken. Das Bearbeiten einer Probe bezieht sich nur auf die Daten am oberen Rand des Datenblatts. Insbesondere sind keine Prozesse betroffen.

Prozess hinzufügen:

Außerdem befindet sich oben auf dem Musterdatenblatt das Zahnradsymbol, das dazu dient, einen neuen Prozess an die Probe anzuhängen. Wenn Sie darauf klicken, werden Sie gefragt, welche Art von Prozess Sie hinzufügen möchten, oder ob Sie eine Probe in Teile spalten wollen.

Proben und Prozesse löschen:

Es ist generell keine gute Idee, in einer Datenbank Dinge zu löschen. Aufgrund der großen Nachfrage von Seiten der Benutzer bietet JuliaBase dennoch die Möglichkeit, Proben und Prozesse nachträglich wieder zu entfernen. Die Regeln sind allerdings sehr streng: Man kann nur Prozesse löschen, die man bearbeiten kann *und* die nicht älter als eine Stunde sind.

Probenspaltung:

Um eine Probe in Stücke aufzuteilen, klicken Sie auf das Zahnradsymbol und wählen Sie *Probenspaltung*. Dann können Sie die neuen Namen der Probenstückchen eingeben. Wenn Sie sich das Datenblatt einer Probe ansehen, sehen Sie auch alle Prozesse der Eltern bzw. Vorfahren.

Der generische Prozess:

Ergebnisprozesse, oft einfach Ergebnis genannt, sind eine praktische Ad-hoc-Möglichkeit, generische Daten an das Datenblatt einer Probe anzuhängen. Wenn Sie ein Messergebnis hinzufügen möchten, für das bisher kein spezieller Prozess programmiert wurde, oder wenn Sie ein Diagramm, ein Bild oder einen Kommentar hinzufügen möchten, dann erstellen Sie ein Ergebnis. Das ist das Schweizer-Armeemesser-Verfahren, wenn nichts anderes passt.

Erweiterte Suche:

Rosalee möchte ihre besten Proben sehen (Abbildung 5.3). Wählen Sie dafür im Hauptmenü *Suche*

Welcome, Rosalee.
Sample 14S-001 was successfully changed in the database.

Samples

Institute of Nifty New Materials (INM)

Sample "14S-001"

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

100 nm
silver (800 nm)
n-a-Si:H (50 nm)
i-a-Si:H (120 nm)
p-a-Si:H (50 nm)
glass (1.1 mm)

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: **3 sccm**
Layer type: **p**	H_2: **1 sccm**
Chamber: **p**	Silane conc.: **64.29 %**
Date: **2014-10-01**	Temperature: **159 / 172 °C**
Layer number: **002**	SiH_4: **1 sccm**
Layer type: **i**	H_2: **0 sccm**
Chamber: **i2**	Silane conc.: **100 %**
Date: **2014-10-01**	Temperature: **113 / 176 °C**
Layer number: **003**	SiH_4: **9 sccm**
Layer type: **n**	H_2: **1 sccm**
Chamber: **n**	Silane conc.: **84.38 %**
Date: **2014-10-01**	Temperature: **131 / 156 °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 cm²**
Efficiency η: **6.97 %**
Short-circuit current density: **11.5 mA/cm²**
Data file: measurement-11.dat

current density in mA/cm²: 160, 140, 120, 100, 80, 60, 40, 20, 0, −20
voltage in V: −0.2, 0.0, 0.2, 0.4, 0.6, 0.8, 1.0

5.17
3.82
1.11
5.96
5.19
6.97
5.01
5.34
6.77

η in %

Abbildung 5.2.: Musterdatenblatt.

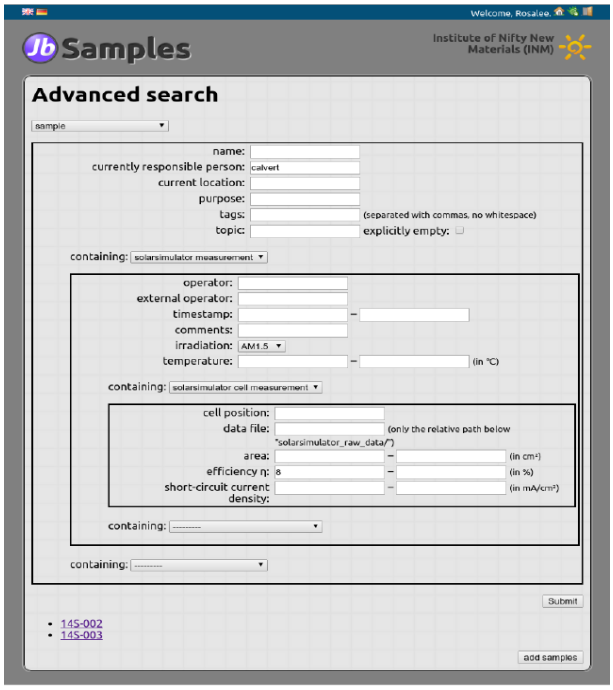


Abbildung 5.3.: Fortgeschrittene Suche.

nach Dingen – Erweiterte Suche. Führen Sie nun die folgenden Schritte durch und klicken Sie nach jedem Schritt auf „Absenden“:

1. Wählen Sie *Probe* im Dropdown-Menü.
2. Geben Sie *calvert* in „*currently responsible person*“ ein und wählen Sie „*solarsimulator measurement*“ im Dropdown-Menü aus.
3. Wählen Sie *AM1.5* in „*Bestrahlung*“ und wählen Sie „*Solarsimulator Zellmessung*“ im inneren Dropdown-Menü.
4. Geben Sie bei Wirkungsgrad η den Wert 8 ein.

Sie erhalten das Ergebnis wie auf dem Bild über diesem Text: Zwei ihrer Proben entsprechen den Kriterien, nämlich *14S-002* und *14S-003*. Das bedeutet, dass an beiden Proben mindestens eine Solarsimulatormessung unter AM 1.5-Bestrahlung gemacht wurde, wobei mindestens eine Zelle einen Wirkungsgrad von mehr als 8 % aufweist. Sie können die Ergebnisse komplexer Suchaufträge mit einem Lesezeichen versehen und sie so oft wie gewünscht erneut aufrufen. Jedes Mal erhalten Sie neue Ergebnisse für Ihre alten Suchkriterien.

Datenexport:

Rosalee braucht die Daten in ihrem Tabellenkalkulationsprogramm. Klicken Sie also ein weiteres Mal auf *Senden*. Sie können auf dem Beispieldatenblatt die Prozesse auswählen, die in den Export aufgenommen werden sollen. Wählen Sie die zweite Schicht der 5-Kammer-Beschichtung und die erste Solarsimulator-Messung. Klicken Sie auf „*Absenden*“. Nun können Sie die Felder dieser Prozesse auswählen, die Teil des Exports werden sollen. Wählen Sie „*SiH4/sccm*“ (das ist der Silan-Gasfluss) der Schicht und „*η der besten Zelle/%*“ der Solarsimulatormessung. Klicken Sie auf „*Absenden*“. Es

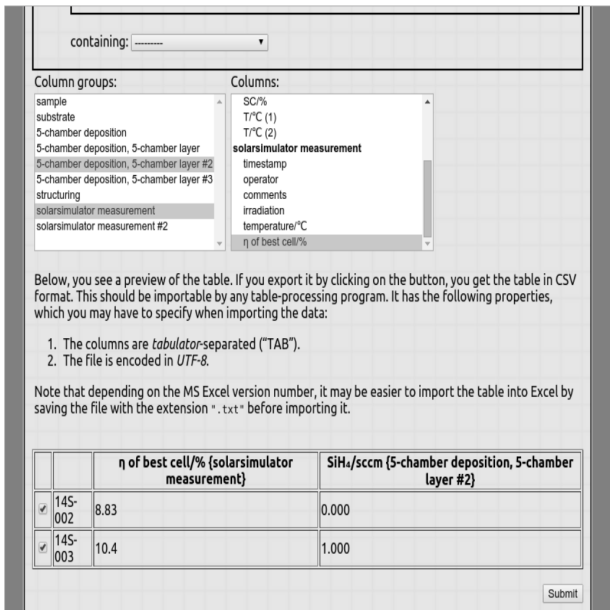


Abbildung 5.4.: Datenexport.

sollte dann wie Abbildung 5.4 aussehen. Die Tabelle enthält alle Daten, die exportiert werden. Klicken Sie ein letztes Mal auf „*Senden*“, und Sie können diese Tabelle als CSV-Datei herunterladen, die Sie mit Ihrem bevorzugten Tabellenkalkulationsprogramm öffnen können.

Proben hinzufügen:

Im Hauptmenü können Sie auf „*Dinge hinzufügen – Proben*“ klicken, um Proben hinzuzufügen. Beachten Sie, dass diese Seite sehr institutsspezifisch ist. Ihr Institut nutzt wahrscheinlich nicht so etwas wie Substrate, und schon hat schon gar nicht so etwas wie eine Reinigungsnummer. In jedem Fall müssen Sie die Anzahl der Proben sowie deren aktuellen Standort eingeben. Fügen Sie ein paar Proben hinzu, aber benennen Sie sie noch nicht um. Frische Proben haben in JuliaBase einen vorläufigen Namen. Er sieht aus wie **00034*, d.h. ein Sternchen gefolgt von einer fünfstelligen Zahl. Verwenden Sie diese Namen niemals auf Proben-Etiketten oder in Veröffentlichungen. Sie sollen so schnell wie möglich durch einen echten Namen ersetzt werden. Rosalees Proben erhalten ihre Namen nach der ersten Abscheidung von Silizium.

Tabellarische Laborbücher:

Öffnen Sie im Hauptmenü das Laborbuch für die Fünf-Kammer-Abscheidung. Sie sehen sechs Abscheidungen vom Oktober 2014. Wählen Sie eine von ihnen aus. JuliaBase zeigt Ihnen eine Seite an, die nur die Details dieser Abscheidung enthält. Klicken Sie oben auf der Seite auf auf das *Zahnradsymbol*, um einen neuen Run anzulegen, der diesen Run als Vorlage nutzt.

Neuen Abscheideprozess hinzufügen:

Rosalee nutzt alte Abscheidungs-Runs als Vorlage, weil diese nicht stark variieren. Auf diese Weise kann sie ohne viel Aufhebens neue Runs hinzufügen. Auf der Seite für die neue Abscheidungen muss sie nur die Proben für die Abscheidung auswählen (das sind die kürzlich hinzugefügten Proben mit diesen *... Namen), einige andere Dinge ändern, die bei diesem Durchlauf anders waren, und auf „Absenden“ klicken. Nun ist es im *Institut für Nifty New Materials* (INM) üblich, der Probe denselben Namen zu geben wie des Abscheinungs-Runs. (JuliaBase gestattet auch andere Namens-Policys.)

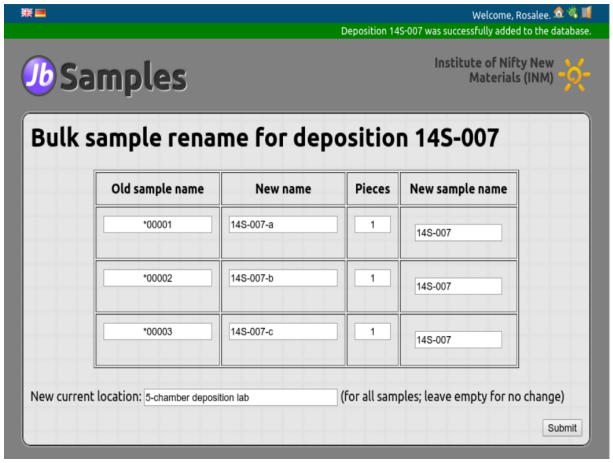


Abbildung 5.5.: Probenumbenennungen nach einem neuen Abscheideprozess.

Daher werden Sie unmittelbar nach dem Hinzufügen der Abscheidung auf eine Seite weitergeleitet, auf der Sie die neuen Probennamen überprüfen und ändern können. JuliaBase schlägt den Namen der Abscheidung für alle Proben vor (im Fall des Screenshots für drei Proben). Da die Namen jedoch eindeutig sein müssen, fügt Rosalee an den Namen . . . -a, . . . -b, und . . . -c an (siehe Screenshot, zweite Spalte). Klicken Sie auf „*Absenden*“, das war's! Die neu hinterlegten Proben erscheinen mit ihrem richtigen Namen unter „*Meine Proben*“ auf der Hauptmenüseite. Es kann natürlich sein, dass Ihr Institut einen anderen Arbeitsablauf ohne eine solche Umbenennung hat, was ohnehin eine schlechte Idee ist – Namen in einer Datenbank sollten sich nie ändern. Sie können also die Umbenennungsseite in Ihrem eigenen Code einfach weglassen.

Berechtigungen für Prozesse ändern:

Wie bereits erwähnt, ist Rosalee für Experimente an der 5-Kammer-Deposition zuständig. Aber neh-

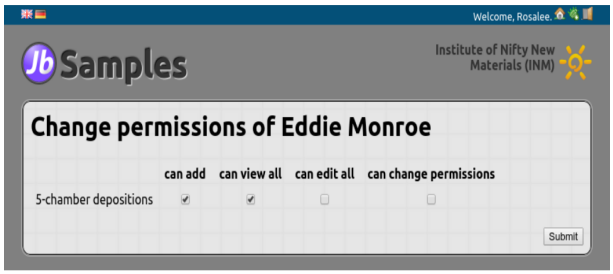


Abbildung 5.6.: Berechtigungen für Prozesse ändern.

men wir mal an, dass Eddie auch solche Ablagerungen machen will und eine Einführung bekommt? Dann sollte er auch die Erlaubnis haben, solche Abscheidungen zu JuliaBase hinzuzufügen (Abbildung 5.6). Rosalee geht im Hauptmenü auf „*Verschiedenes – Berechtigungen für Prozesse*“, wählt Eddie aus dem Dropdown-Menü aus und klickt auf „*Absenden*“. Sie setzt Häkchen in die ersten beiden Kontrollkästchen und klickt erneut auf „*Absenden*“. Jetzt hat Eddie die folgenden zusätzlichen Berechtigungen erhalten:

- Er kann neue 5-Kammer-Abscheidungen hinzufügen.
- Er kann seine eigenen 5-Kammer-Abscheidungen bearbeiten (diejenigen, deren Betreiber er ist).

- Er kann alle 5-Kammer-Abscheidungen einsehen. Dies bedeutet insbesondere, dass er das Laborbuch einsehen kann.

Probeneigentümer:

Manchmal müssen Proben den Besitzer wechseln, z.B. wenn jemand das Institut verläßt. Nehmen

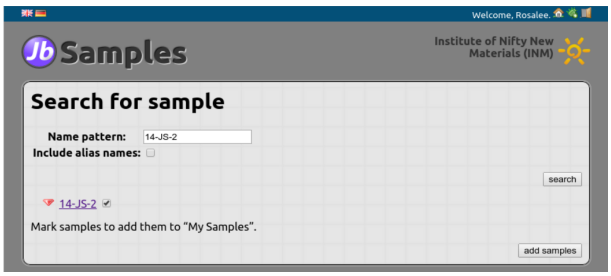


Abbildung 5.7.: Probensuche.

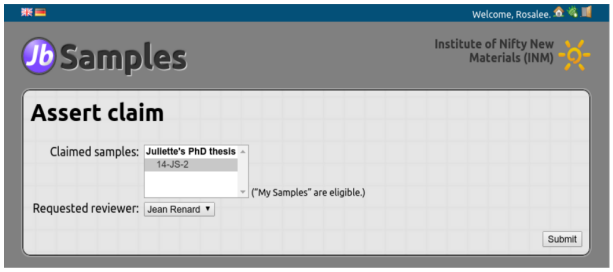


Abbildung 5.8.: Probeneigentümer.

wir an, es gibt eine Probe von Juliette, die Rosalee übernehmen möchte. Rosalee hat diese Probe bereits über die Suche gefunden (Abbildung 5.7). Im Prinzip könnte Juliette die derzeit verantwortliche Person der Probe auf Rosalee setzen, aber Juliette hat keine Zeit dafür, oder arbeitet vielleicht nicht mehr am INM. Es könnte auch um Proben gehen, die durch einen Altdatenimport in die Datenbank gespühlt wurden und noch keinen Besitzer haben. Wie auch immer: In diesen Fällen bietet JuliaBase über die „Beanspruchung von Proben“ (siehe Abbildung 5.8) die Möglichkeit, Proben quasi an sich zu ziehen. Rosalee muss dafür noch einen berechtigten Gutachter wählen – ihr Chef Sean Renard bietet sich an – der dann über die Probenbeanspruchung entscheidet.

Juliette: Die Arbeitsverteilerin

Juliette hat viel zu tun und kann sich nicht um schnöde Dinge wie die Probenvorbereitung und Probencharakterisierung selbst kümmern. Deshalb überträgt sie Aufgaben an andere Personen und analysiert die Ergebnisse. Melden Sie sich ab und melden Sie sich erneut als *j.silverton/12345* an.

Einen Auftrag hinzufügen:

Nehmen wir an, Juliette möchte eine PDS-Messung für ihre Probe *14-JS-1* durchführen lassen. Gehen Sie deshalb im Hauptmenü auf „*Sonstiges – Aufgabenlisten*“ (Abbildung 5.9). Wählen Sie als erstes dort die Prozesse aus, an denen Sie interessiert sind: Markieren Sie „*PDS-Messungen*“ und klicken Sie auf „*Senden*“. Nun fügen Sie eine neue Aufgabe für den PDS-Operateur (das ist Nick Burkhardt) hinzu, indem Sie auf das „*Plus-Symbol*“ für „*PDS-Messungen*“ klicken. Wählen Sie die Probe *14-JS-1*, klicken Sie auf „*Absenden*“ und Sie sind fertig. Sie können den neuen Auftrag in der Liste der

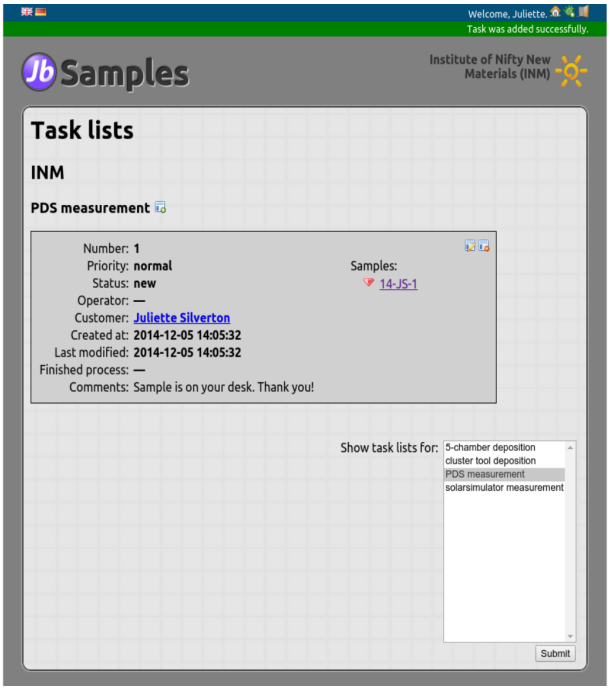


Abbildung 5.9.: Eine Aufgabe hinzufügen.

Aufträge sehen.

Eine Probe an einen anderen Nutzer senden:

Juliette möchte Nick die Probe *14-JS-1* zeigen, damit er sie sich ansehen kann. Natürlich könnte Nick auch selbst nach der Probe suchen, aber da die Probe zum Thema von Juliettes Doktorarbeit gehört und nicht zu Nick, kann er das Datenblatt der Probe nicht einsehen. Um die Probe an Nick zu senden, klicken Sie auf das „*Bleistift*“-Symbol neben „*Meine Proben*“ auf der Hauptmenüseite. Wählen Sie auf der linken Seite die Probe *14-JS-1* aus. Wählen Sie dann auf der rechten Seite in der Mehrfachauswahl „*Zu Benutzer kopieren*“ Nick aus und geben Sie z.B. „*Bitte schau dir diese Probe mal an*“ unter „*Kommentar für Empfänger*“ ein. Schließlich setzen Sie „*Freigabe*“ auf alle bisherigen Vorgänge, denn Juliette möchte, dass Nick das gesamte Datenblatt von *14-JS-1* sehen kann.

Nick: Übersendete Probe einsehen

Melden Sie sich nun als *n.burkhardt/12345* an. Sie sehen *14-JS-1* unter „*Meine Proben*“ und können das Datenblatt einsehen. Die Übersendung der Probe hat funktioniert!

Der News-Feed:

Ausserdem wurde Nick über die Übersendung im *Hauptmenü – Verschiedenes – Newsfeed* benachrichtigt. Dort kann er auch sehen, dass Juliette einen neuen Auftrag für eine PDS-Messung angelegt hat. Der *Newsfeed* enthält alle wichtigen Nachrichten für den jeweiligen Benutzer: Änderungen in seinen Proben, neue Proben in seinen Themen, auf ihn übertragene Proben, neue Aufträge und vieles mehr. Der *Newsfeed* ist eigentlich nicht dafür gedacht, im Browser angezeigt zu werden. Sie können es zwar tun, aber es ist ein wenig umständlich. Verwenden Sie stattdessen ein RSS-Feed-fähiges Programm wie Thunderbird. Dieses Programm kann Ihnen überdies anzeigen, welche Einträge im Feed wirklich neu sind.

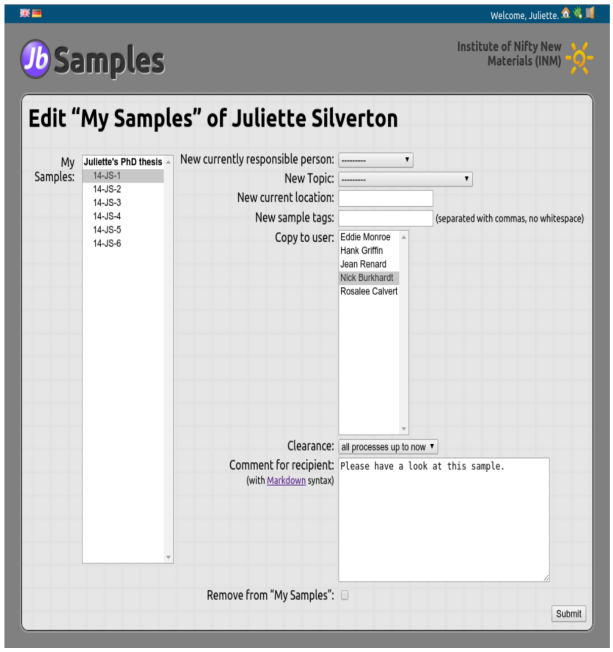


Abbildung 5.10.: Eine Probe an einen anderen Nutzer senden.

Aufträge:

Da Nick gelesen hat, dass Juliette eine neue PDS-Aufgabe eingereicht hat, besucht er selbst die Seite

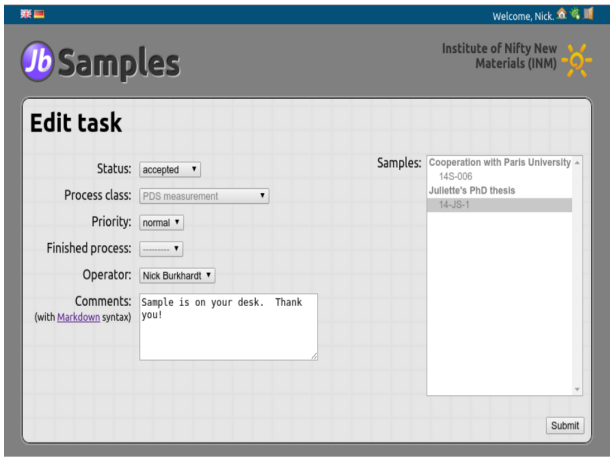


Abbildung 5.11.: Auftrag für Nick.

„*Auftragslisten*“ (Abbildung 5.11). Wie oben bereits erklärt, müssen Sie beim ersten Mal auf dieser Seite die *PDS* auswählen und auf „*Senden*“ klicken, damit Nick die PDS-Aufträge sehen kann. In der Regel arbeiten mehr als eine Person an einer Anlage wie der PDS. Manchmal sind die Leute abwesend (Urlaub, Krankheit, usw.). Daher ist es nicht von vornherein klar, wer einen Auftrag tatsächlich erledigen wird, und der Auftrag muss von jemandem ausdrücklich angenommen und auch jemandem zugewiesen werden. Klicken Sie dazu auf das *Stiftsymbol*, um ihn zu bearbeiten. Setzen Sie den *Status* auf „*akzeptiert*“ und übertragen Sie die Aufgabe an Nick selbst. Juliette wird hierüber benachrichtigt. Wenn Nick die Messung tatsächlich durchführt, kann er den Status des Auftrags auf „*in Bearbeitung*“ und danach auf „*abgeschlossen*“ setzen. Ein fertiger Auftrag kann sogar mit der

konkreten PDS-Messung verbunden werden. Einige dieser Schritte sind optional. Sie hängen von Arbeitsablauf in Ihrem Institut ab.

Sean: Der Teamleiter

Melden Sie sich als *s.renard/12345* an. Sean ist der Teamleiter und hat erweiterte Rechte. Diese sind:

- Alle Proben ansehen
- Neue Themen erstellen
- Ändern von Mitgliedschaften in allen Themen
- Erteilen und Entziehen von Berechtigungen für alle Anlagen und Experiment
- Probenbeanspruchungen genehmigen oder ablehnen

Das *Institute of Nifty New Materials* (INM) hat nur zwei Ebenen: Der Teamleiter und der Rest. Sie können weitere Ebenen in Ihrer Einrichtung hinzufügen, und Sie können die Berechtigungen auf andere Weise festlegen. Wir haben jedoch die Erfahrung gemacht, dass komplexe Berechtigungsregelungen eher ein Klotz am Bein sind.

Genehmigen Sie einen Probenanspruch:

Navigieren Sie zu „*Hauptmenü Verschiedenes – Musteranträge*“. Unten auf dieser Seite sehen Sie die

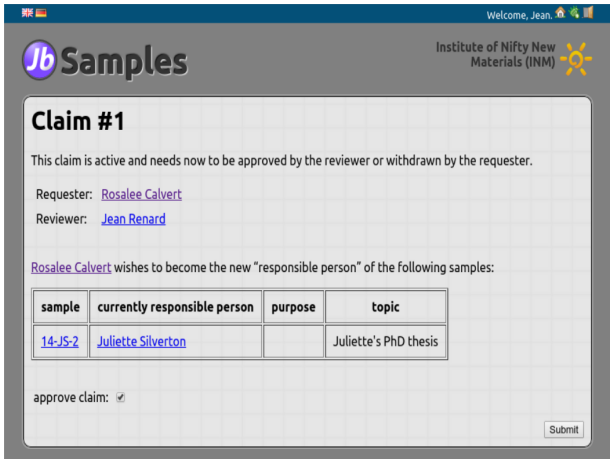


Abbildung 5.12.: Probenanspruch genehmigen.

Beanspruchung von Juliettes Proben von Rosalee (Abbildung 5.12). Klicken Sie sie an. Sean kann nun die Besanspruchung im Detail überprüfen und ihn genehmigen oder ablehnen.

5.3.2. eLabFTW

Das elektronische Laborbuch eLabFTW (`https://www.elabftw.net/`) ist eine weitere, freie Alternative eines ELN und dieses System gewinnt international zunehmend an Popularität. eLabFTW wird ständig weiter entwickelt (die aktuelle Version im Juli 2025 ist 5.2.8), wobei die Entwickler bewusst den Kontakt zu den Nutzern suchen und Verbesserungsvorschläge der Community gerne in die Entwicklung mit aufnehmen. Wer kommerziellen Support für seine eLabFTW-Instanz haben möchte, kann diesen ebenfalls durch das kleine Unternehmen Deltablot (`https://www.deltablot.com/`) bekommen. In diesem Handbuch stellen wir nur die notwendigsten Menüs und Arbeitsabläufe dar, damit ein erfolgreicher Start mit diesem sehr mächtigen System einfach gelingen kann. eLabFTW bietet sehr viele Optionen und Möglichkeiten, um Experimente zu organisieren, durchzuführen und auch zu automatisieren. Das Ziel ist eine echte Arbeitserleichterung für seine Nutzer sowie ein sicherer Umgang mit experimentellen Daten. Eine ausführliche Anleitung in englischer Sprache bietet das stets aktuelle Online-Handbuch von eLabFTW, aufgeteilt in Bereiche mit unterschiedlichen Informationen für Nutzer, Admins und Sysadmins (`https://doc.elabftw.net/index.html`). Die mögliche Detailtiefe geht weit über das hinaus, was wir in diesem Best-Practice-Handbuch darstellen können und wollen. Wir möchten die Leser dieses Handbuches nur mit den praktisch notwendigen Informationen versorgen, um den Einstieg möglichst zu erleichtern.

Warum sollte man eLabFTW verwenden?

eLabFTW ist Open Source und bietet eine Reihe von Vorteilen, welche dieses System für viele Nutzer attraktiv machen. Es bietet viele Gestaltungs- und Anpassungsmöglichkeiten an individuelle Bedürfnisse und kann für die meisten Laborumgebungen mit vertretbarem Aufwand konfiguriert werden. Das System ist webbasiert, wobei keine Clients installiert werden müssen. Ein vernetzter Rechner und ein Browser sind ausreichend. Durch die Verwendung eines responsiven Designs kann eLabFTW auf vernetzten Geräten mit beliebiger Bildschirmgröße eingesetzt werden, vom Handy bis zum Riesenbildschirm. Weitere, wichtige Vorteile sind:

- Es können sichere Zeitstempel für Experimente verwendet werden (RFC 3161 als Standard, oder über eine Blockchain).
- Die Authentizität eines Experimentes kann zusätzlich mit einer persönlichen Signatur abgesichert werden.
- Das System kann über Schnittstellen (REST-API) Daten austauschen.
- Für den Import und Export von Daten stehen verschiedene, gängige Dateiformate zur Verfügung: PDF, ZIP, CSV, JSON, QR-Code.
- Es existiert ein umfangreiches Rollen- und Rechtemanagement.
- Für sich wiederholende Experimente können Vorlagen (Templates) individuell kreiert und verwendet werden.
- Es können Datenbanken für Produkte/Chemikalien und Protokolle angelegt werden (Inventar-Verwaltung).
- Es können ToDo-Listen angelegt werden.

- Labore und ganze Teams können über einen Terminplaner organisiert werden.
- Es gibt einen JSON- und einen Molekül-Editor.
- eLabFTW ist inzwischen in 21 Sprachen übersetzt worden und jeder Nutzer kann seine bevorzugte Sprache selbst einstellen.
- Man kann eLabFTW in seinem eigenen, abgeschlossenen Netzwerk laufen lassen, wenn besondere Sicherheitsanforderungen existieren und man die Wartung und den Betrieb des ELN vollständig selbst realisieren möchte.
- Es können eine Vielzahl verschiedener Teams gleichzeitig mit einer installierten Instanz arbeiten, ohne dass es dabei zu Überschneidungen oder Konflikten kommt (einer der Autoren dieses Handbuchs arbeitet aktuell in seinem Institut mit mehr als 20 verschiedenen Teams mit einer Instanz).

Notwendige Infrastruktur

Die Installation von eLabFTW erfolgt am einfachsten mit Hilfe eines Docker-Containers auf einem dafür vorgesehenen Linux-Server. Weitere Möglichkeiten sind die Installation in einer Cloud-, auf einem NAS-Server sowie auf einem Mac- oder einem Windows-System. Die Autoren dieses Handbuches bevorzugen die erstere Installation in einem Docker-Container auf einem Linux-Server, da diese Option recht einfach und sicher betrieben, gewartet und auch bei Bedarf erweitert werden kann (z.B. mit zusätzlichem Speichervolumen). Für Details der Installation verweisen wir an dieser Stelle auf die ausführliche eLabFTW-Onlinehilfe (`https://doc.elabftw.net/`).

Das Rollen- und Rechtemanagement

eLabFTW verfügt über ein ausgefeiltes Rollen- und Rechtemanagement, basierend auf einer streng hierarchischen Struktur. Das folgende Schema verdeutlicht die verschiedenen Ebenen, wobei die Rechte von oben nach unten abnehmen. Der Systemadministrator (kurz: Sysadmin) hat die volle

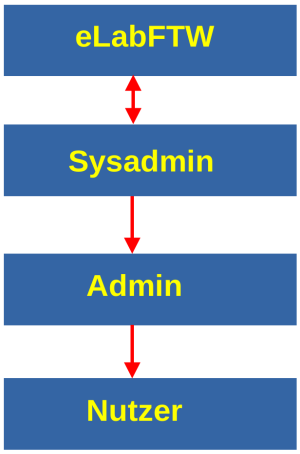


Abbildung 5.13.: Das Rollenmanagement in eLabFTW.

Kontrolle über die installierte Instanz und kann gleichzeitig nach unten auf die Ebenen der Administratoren (kurz: Admin) und der Nutzer einwirken. Er überwacht den laufenden Betrieb und beseitigt

Störungen im System. Für die Organisation der praktischen Arbeiten sind allerdings die Admins und Nutzer zuständig. Das folgende Diagramm zeigt mögliche Aufteilungen von Teams, wie sie von den Admins vorgenommen werden können: Wenn neue Nutzer in das System aufgenommen werden sol-

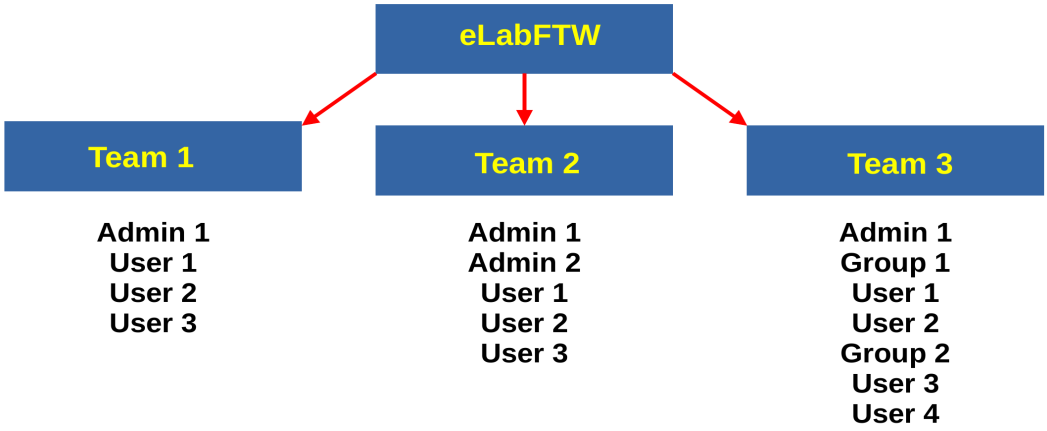


Abbildung 5.14.: Drei Beispiele für mögliche Einteilungen von Teams. Die Aufteilung der Nutzer in bestimmte Teams werden durch die jeweiligen Admins vorgenommen. Weitere Variationen sind möglich.

len, dann muss Ihnen vorher mitgeteilt werden, welchem Team sie zugeordnet werden. In einem Team kann der erste Admin weitere Admins benennen, die anschließend über die gleichen Rechte verfügen wie er selbst. Die Nutzer eines Teams können weiter in Gruppen eingeteilt werden, zum Beispiel weil sie gemeinsam an einem Forschungsprojekt arbeiten sollen. Weiterhin ist es für einen Admin auch möglich, Nutzer aus anderen Teams zur Mitarbeit in einer Gruppe einzuladen. Die folgende Tabelle fasst die unterschiedlichen Rechte von Admins und Nutzern zusammen. Wir werden im Folgenden näher auf diese Unterschiede eingehen.

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

Abbildung 5.15.: Die unterschiedlichen Rechte von Admins und Nutzern in eLabFTW.

Einführung in die Arbeitsweise von eLabFTW: Login und EXPERIMENTE

Für den Zugang zu eLabFTW muss man als Nutzer zunächst von seinem Admin im System registriert und einem Team zugeordnet sein (siehe Abbildung 5.16). Mit einem ersten Passwort, wel-

ches man nach dem ersten Login ändern sollte, gelangt man in das System. In diesem Beispiel wird die öffentlich zugängliche Demo-Instanz von eLabFTW verwendet (`https://demo.elabftw.net/`). Wir können den Einstieg mit dieser Demo-Version sehr empfehlen. Hiermit ist es risikolos möglich, die meisten Funktionen auszuprobieren und kennenzulernen, ohne dass dabei irgendwelche Daten beschädigt werden können. Außerdem kommt man in den Genuss, die Funktionen und Menüs der aktuellsten Version auf einfache Weise ausprobieren zu können. Es erscheint der Bildschirm mit

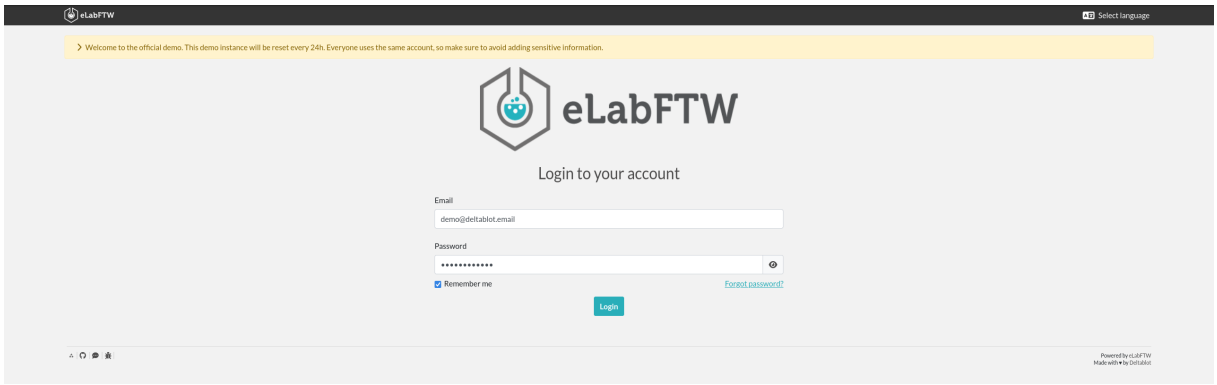


Abbildung 5.16.: Browser-Oberfläche von eLabFTW: Login.

dem Bereich der EXPERIMENTE (Abbildung 5.17). Ganz oben auf dem Bildschirm ist ein Reihe mit Reitern zu erkennen (weiße Schrift auf schwarzen Hintergrund). Mit einem Mausklick auf das eLabFTW-Logo oder das Haus-Symbol öffnet sich ein Dashboard, welches einen aktuellen Überblick über die eigenen Arbeiten bietet. Rechts neben dem Reiter EXPERIMENTE (in diesem Menü

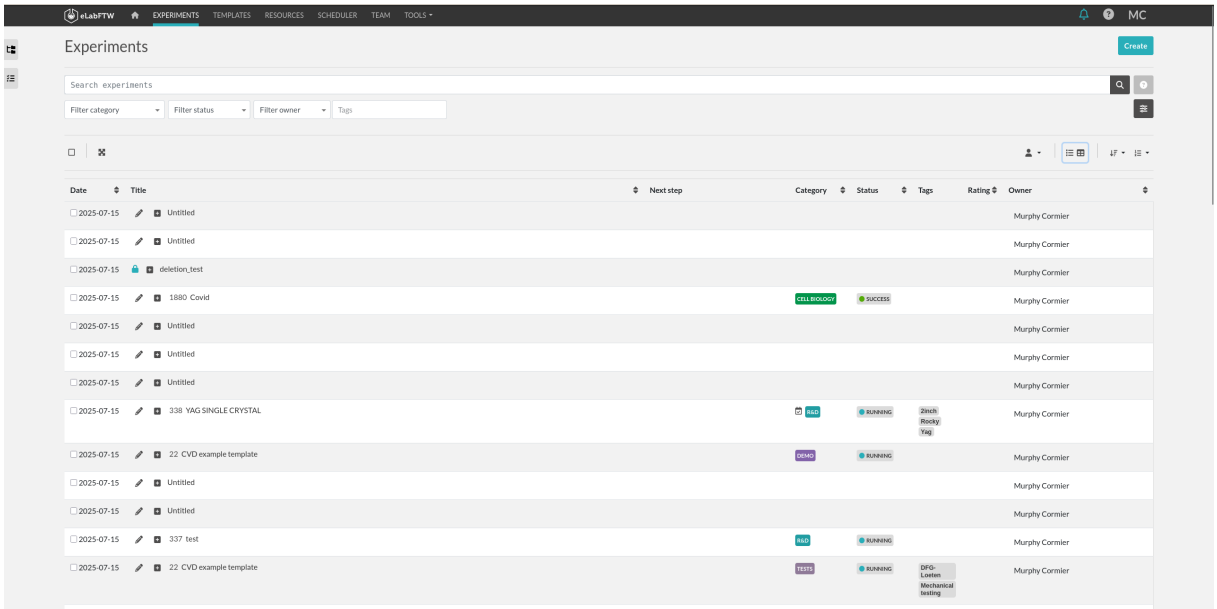


Abbildung 5.17.: Browser-Oberfläche von eLabFTW: EXPERIMENTE.

bewegen wir uns aktuell), erscheint der nächste Reiter TEMPLATES. Dieser zeigt die für ein Team verfügbaren Templates an, also Makros für die vereinfachte und wiederholte Durchführung von Experimenten. Eine Weiterentwicklung der Experimente sind die sogenannten RESSOURCEN. Auf die Unterschiede werden wir im Folgenden noch näher eingehen. Bei dem nächsten Reiter SCHEDULER

handelt es sich um einen Zeitplaner, mit dessen Hilfe sich ein Team organisieren und Experimente, Meetings usw. planen kann. Der nächste Reiter TEAM listet alle Mitglieder eines Teams auf, inklusive der jeweiligen Email-Adresse. So ist zum Beispiel eine einfache Kommunikation innerhalb von eLabFTW möglich. Der Reiter TOOLS funktioniert als Pull-Down-Menü und stellt nützliche Zusatzprogramme zur Verfügung. Auch dazu später mehr.

Wenn wir in der Liste der EXPERIMENTE ein bestimmtes Experimente auswählen, indem wir

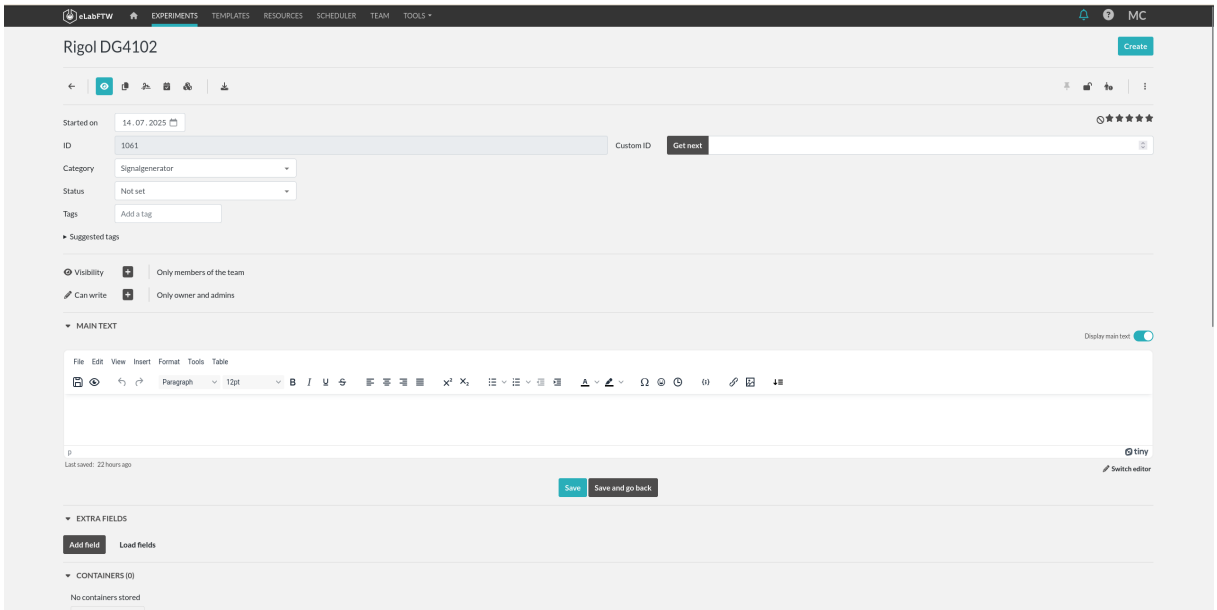


Abbildung 5.18.: Browser-Oberfläche von eLabFTW: Ein einzelnes Experiment.

mit der Maus auf den jeweiligen Namen klicken, dann erscheint die Oberfläche für ein einzelnes Experiment mit neuen Menüs und Symbolen, die wir in diesem Kapitel einzeln erläutern werden (Abbildung 5.18). Zentral erscheint in diesem Menü ein Editor, der für jedes Experiment zur Verfügung steht und ähnlich einer Textverarbeitung funktioniert. Was diesen Editor von einer normalen Textverarbeitung allerdings unterscheidet, ist die Möglichkeit, auch mit dem Textsatzsystem LaTeX arbeiten zu können. Für weitere Details verweisen wir an dieser Stelle auf das Handbuch zu eLabFTW (`https://doc.elabftw.net/index.html`).

In eLabFTW werden prinzipiell zwei Hauptobjekttypen unterschieden: EXPERIMENTE und RESSOURCEN. In diesem Kapitel wollen wir uns zunächst auf Experimente konzentrieren. Sie sind grundsätzlich Eigentum des Nutzers, der ein Experiment angelegt bzw. gestartet hat. Für die vereinfachte Durchführung wiederkehrender Experimente können Templates (Vorlagen) erstellt werden. Abgeschlossene Experimente können aus Gründen der Rechtssicherheit, zum Beispiel für die Anmeldung von Patenten, mit einem Zeitstempel versehen werden. Weiterhin ist es möglich, Experimente mit einer persönlichen Signatur zu versehen. Beides erhöht die Sicherheit in Bezug auf die Eigentumsrechte eines Nutzers. Beide Verfahren werden im Laufe dieses Kapitels noch näher erläutert werden.

Die folgende Grafik (Abbildung 5.19) zeigt einen Teil der Browser-Oberfläche von eLabFTW, in dem Experimente angezeigt werden. Dabei hat man die Wahl zwischen dem Showmodus (oben) und dem Listenlayout (unten). Durch einen linken Mausklick auf die rot umrandete Schaltfläche kann zwischen diesen beiden Modi umgeschaltet werden. Durch das Anklicken des Wortes EXPERIMENTE in der obersten Zeile im Browserfenster (siehe Abbildung 5.20) gelangt man in das entsprechende Menü. Man kann an dieser Stelle ein neues Experiment gestalten und starten. In diesem Beispiel ist

Leo Riem der Besitzer dieses Experimentes. Das Experiment wurde am 07.08.2024 begonnen und es handelt sich um eine Röntgenbeugungsmessung (Titel: Empyrean), mit einem Röntgendiffraktometer mit dem Namen Empyrean (daher der Titel). Im Fenster kann man einige Symbole erkennen, die im Folgenden einzeln erklärt werden.

- 1 - Anzeigemodus.
- 2 - Experimente duplizieren.
- 3 - Signatur hinzufügen.
- 4 - Zeitstempel.
- 5 - Zeitstempel mittels Blockchain.
- 6 - Experiment in externes Format exportieren.
- 7 - Pin: Der Eintrag wird an den Anfang der Liste der Experimente gesetzt.
- 8 - Sperren/Entsperren eines Elements.
- 9 - Aktivieren/Deaktivieren des exklusiven Bearbeitungsmodus.
- 10 - Aktion durch andere Nutzer anfordern.
- 11 - Ellipsis-Menü mit weiteren, möglichen Funktionen.

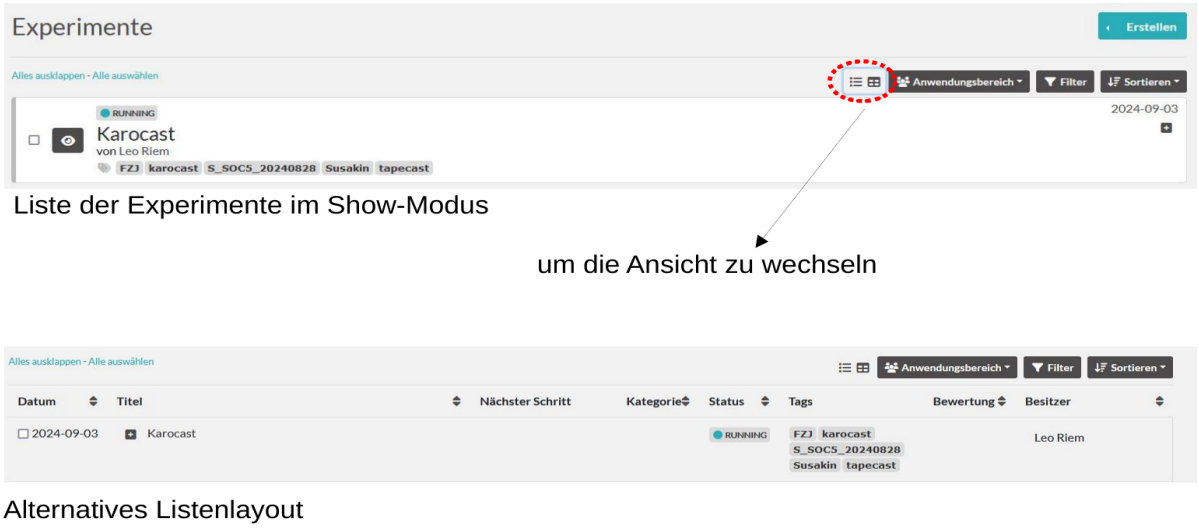


Abbildung 5.19.: Browser-Oberfläche von eLabFTW: Showmodus und Listenlayout.

Mit dem Symbol 5 können Experimente in verschiedene Dateiformate exportiert werden, um sie anschließend ggf. weiter verarbeiten zu können. Im Ellipsis-Menü (Symbol 11) können weitere Einstellungen vorgenommen werden, wie zum Beispiel das Übertragen der Eigentümerschaft eines Experimentes oder das Archivieren und Wiederherstellen eines Experimentes. An dieser Stelle kann man ein anderes Teammitglied auswählen, das man als neuen Eigentümer für ein Experiment eintragen möchte. Man kann jedem Experiment eine Kategorie zuweisen (muss es aber nicht), aber nur der

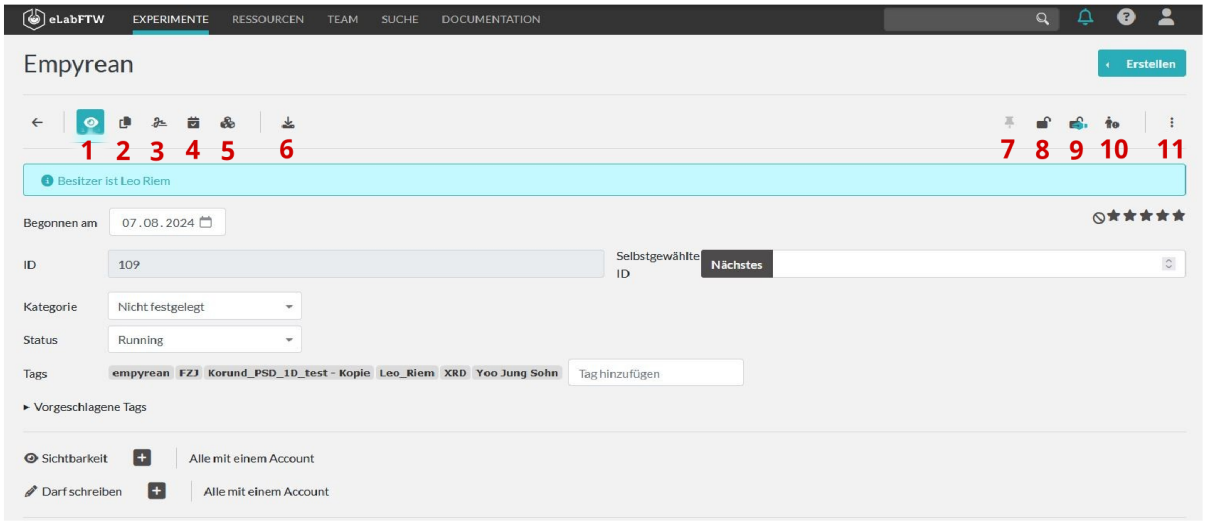


Abbildung 5.20.: Browser-Oberfläche von eLabFTW: Die 11 wichtigsten Symbole für ein Experiment.

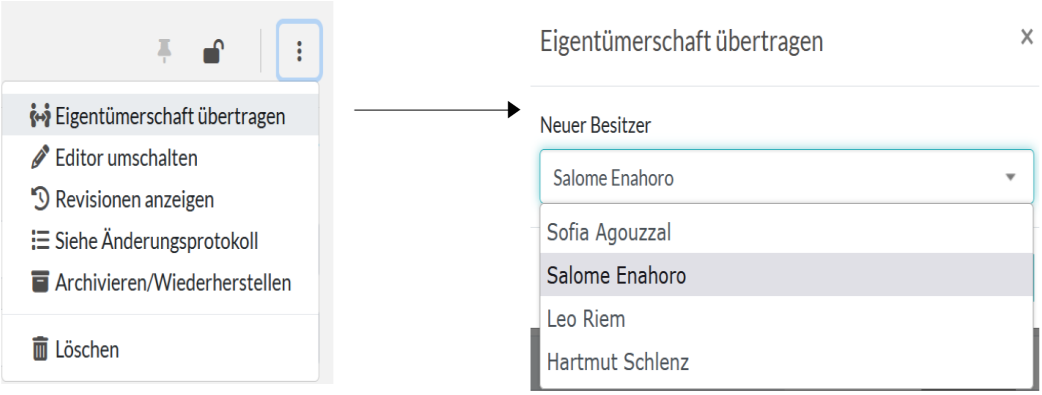


Abbildung 5.21.: Browser-Oberfläche von eLabFTW: Das Ellipsis-Menü.

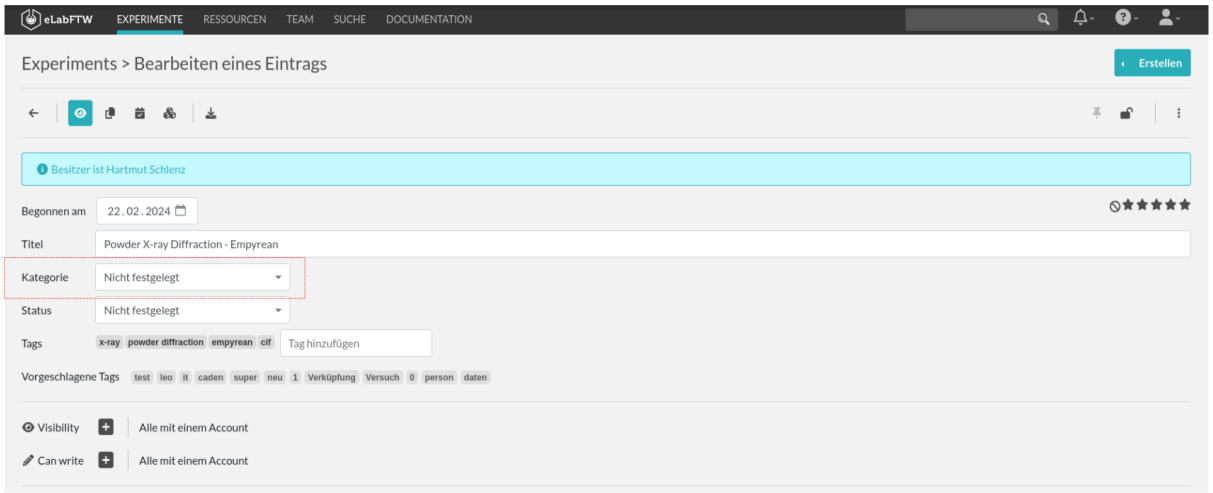


Abbildung 5.22.: Browser-Oberfläche von eLabFTW: Kategorie festlegen.

Administrator kann festlegen, welche Kategorien für ein Team überhaupt zur Verfügung stehen. Das können zum Beispiel Kategorien wie bestimmte Projekte und deren Namen sein, oder nur Demo- oder Testexperimente, Produktionsverfahren, usw. Mit dem Status kann festgelegt werde, in welcher

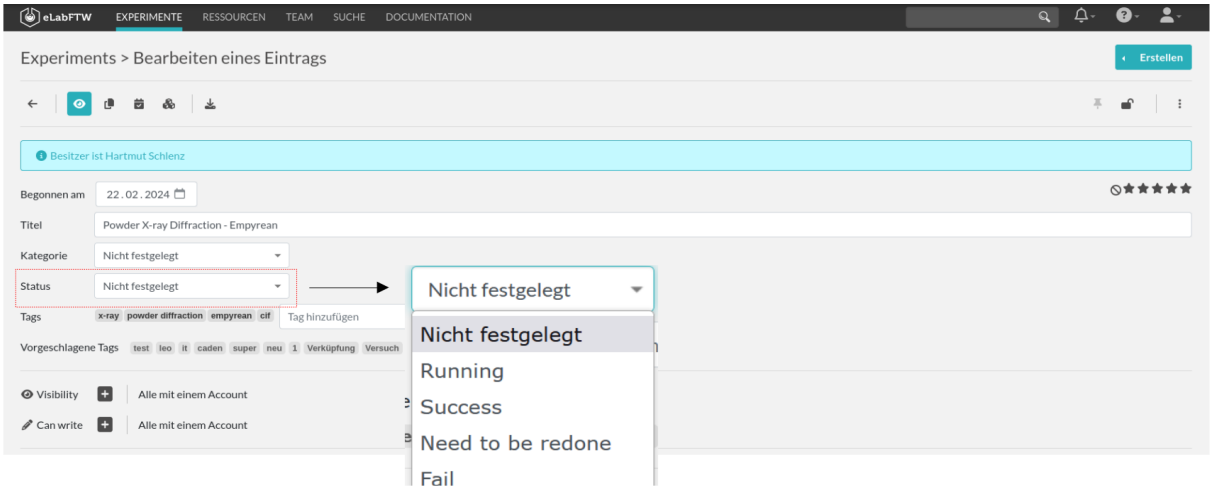


Abbildung 5.23.: Browser-Oberfläche von eLabFTW: Status festlegen.

Phase sich ein Experiment aktuell befindet. So kann ein Experiment noch laufen, es kann erfolgreich abgeschlossen worden sein oder es muss ggf. wiederholt werden. Mit Tags (Schlagwörter) können

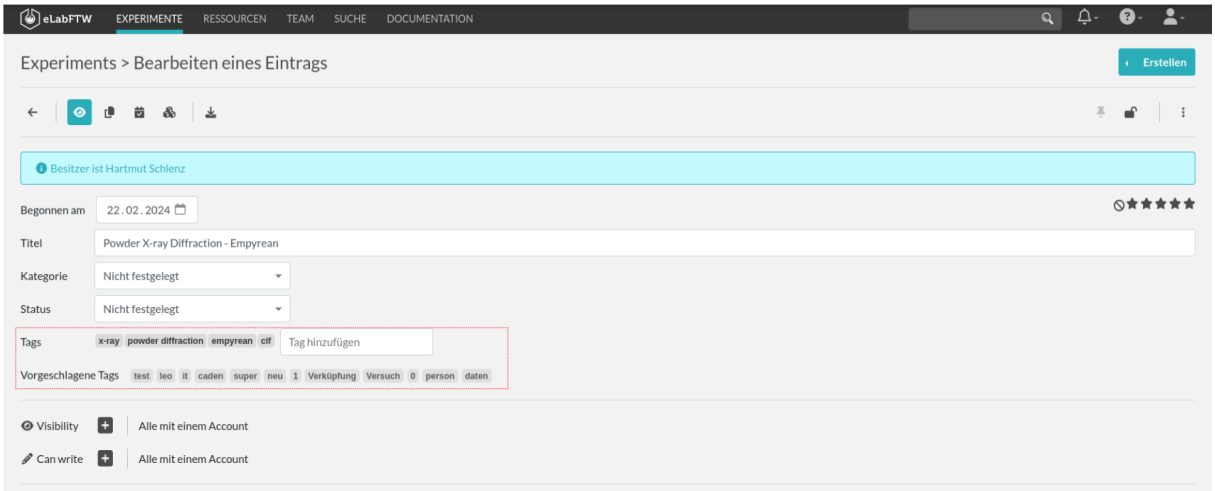


Abbildung 5.24.: Browser-Oberfläche von eLabFTW: Tags auswählen.

Experimente einfach gekennzeichnet und gruppiert werden. Dabei ist die Anzahl der Tags nicht begrenzt. Auf diese Weise werden die Metadaten zu einem Experiment erweitert, ähnlich wie bei einer Software zur Bildbearbeitung von Photos, wo Tags ebenfalls zum leichteren Wiederfinden von Bildern verwendet werden. Alle Experimente mit demselben Tag werden durch Anklicken oder Suchen nach diesem Tag sofort zugänglich. Im Bearbeitungsmodus reicht ein Klick auf ein Tag, um es zu entfernen. Man erkennt im Menü eine Vorschlagsliste der zuletzt verwendeten Tags. Dabei stehen alle Tags für ein Team gemeinsam zur Verfügung. Bei dem zunächst letzten Punkt handelt es sich bereits um Einstellungen für das Rollen- und Rechtemanagement (dazu später mehr). Unter *Visibilty*

(Sichtbarkeit) und *Can write* (darf schreiben) kann jeder Nutzer selbst festlegen, wer seine Experimente in welcher Form sehen und/oder verändern darf. Möchte ein Nutzer nicht, dass irgendjemand sein Experiment überhaupt nur sehen kann, dann ist eine entsprechende Einstellung an dieser Stelle einfach möglich. Selbst ein Administrator hat dann keine Möglichkeit, ein Experiment zu sehen oder zu verändern. Ein solches Vorgehen macht allerdings keinen Sinn, wenn man als Team gemeinsam an einem Experiment arbeitet. Dann kann derjenige Nutzer, der das Experiment angelegt und gestartet hat, festlegen, welche Teammitglieder mit an diesem Experiment arbeiten können. Damit haben

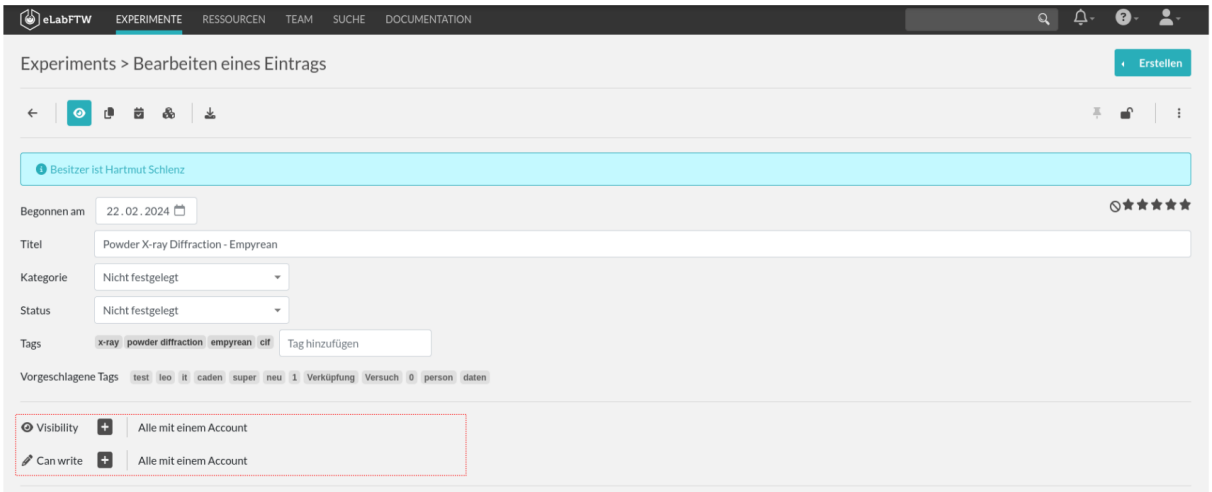


Abbildung 5.25.: Browser-Oberfläche von eLabFTW: Berechtigungen festlegen.

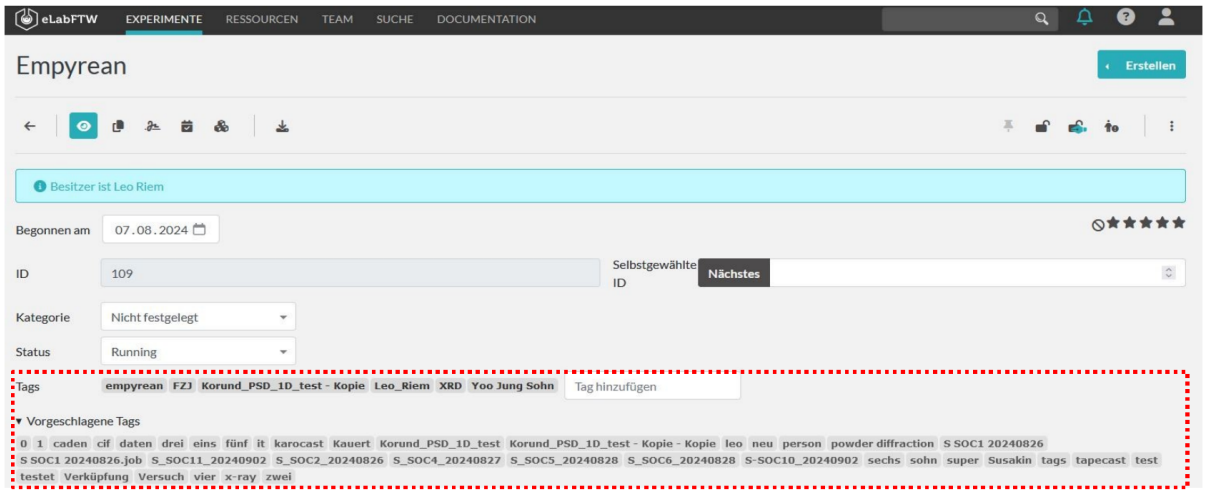


Abbildung 5.26.: Browser-Oberfläche von eLabFTW: Vorschläge für Tags durch eLabFTW.

wir kurz die wichtigsten Einstellungen auf der ersten Seite eines (neuen) Experimentes in eLabFTW erläutert, sobald sie dieses über den Reiter EXPERIMENTE angelegt haben. Im nächsten Abschnitt wird es um die Erstellung und Verwendung von Templates (Vorlagen) gehen, mit deren Hilfe man die Arbeit mit eLabFTW recht komfortabel vereinfachen und automatisieren kann.

TEMPLATES

In der Abbildung 5.27 wird die einfachste Form eines Templates gezeigt, mit Bezug zu dem Beispiel einer CIF-Datei im Kapitel 2.1. Hier wird einfach der in jedem Experiment zur Verfügung stehende Editor verwendet, der sehr ähnlich einer üblichen Textverarbeitung funktioniert, um die notwendigen Informationen strukturiert aufzuschreiben und die für ein Experiment notwendigen Parameter einfach und komfortabel ändern zu können. Das Anlegen eines Templates kann aber auch

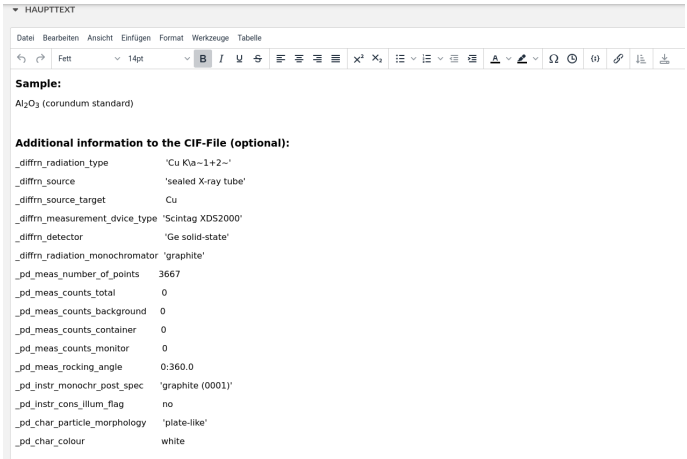


Abbildung 5.27.: Browser-Oberfläche von eLabFTW: Einfaches Template im Editor.

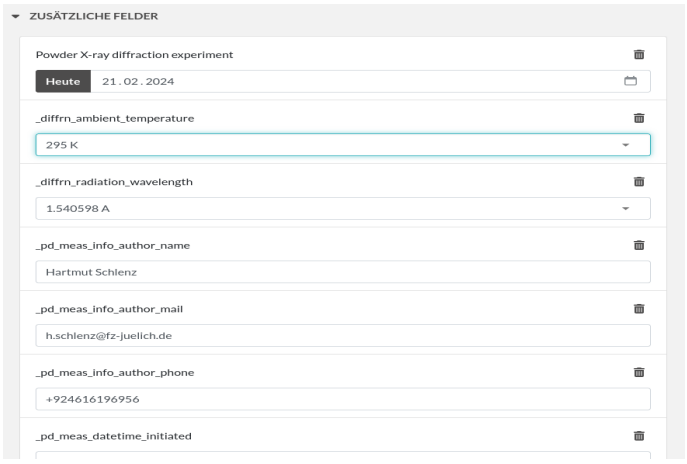


Abbildung 5.28.: Browser-Oberfläche von eLabFTW: Template mit Pull-down Menüs.

deutlich komfortabler gestaltet werden, zum Beispiel durch die Erzeugung von Pulldown-Menüs. In der Abbildung 5.28 können unter anderem verschiedene Temperaturen ausgewählt werden, bei denen ein Experiment durchgeführt werden soll. Für die Generierung solcher Templates ist der interne JSON-Editor notwendig. Das Online-Handbuch von eLabFTW zeigt an Beispielen sehr anschaulich und ausführlich den einfachen Gebrauch diese Editors für die Erstellung eigerner Templates, die anschließend allen Team-Mitgliedern zur Verfügung gestellt werden können (`https://doc.elabftw.net/user-guide.html#templates`).

TEAM

Mit den Mitgliedern des eigenen Teams sowie auch mit anderen Nutzern einer Instanz von eLabFTW können Nachrichten und Informationen über das interne Email-System ausgetauscht werden, sofern ein Nutzer im System bekannt ist. Abbildung 5.29 zeigt die intuitive Oberfläche der Email-Funktion von eLabFTW. Das Menü enthält in diesem Beispiel unter dem Menüpunkt MITGLIEDER die Namen und Mailadressen der eigenen Teammitglieder und durch einfaches Anklicken der Mailadresse wird das notwendige Menü zum Schreiben geöffnet. Alternativ wählt man den Menüpunkt E-MAIL und kann direkt eine Nachricht verfassen. Über den Menüpunkt VORLAGEN können die in einem

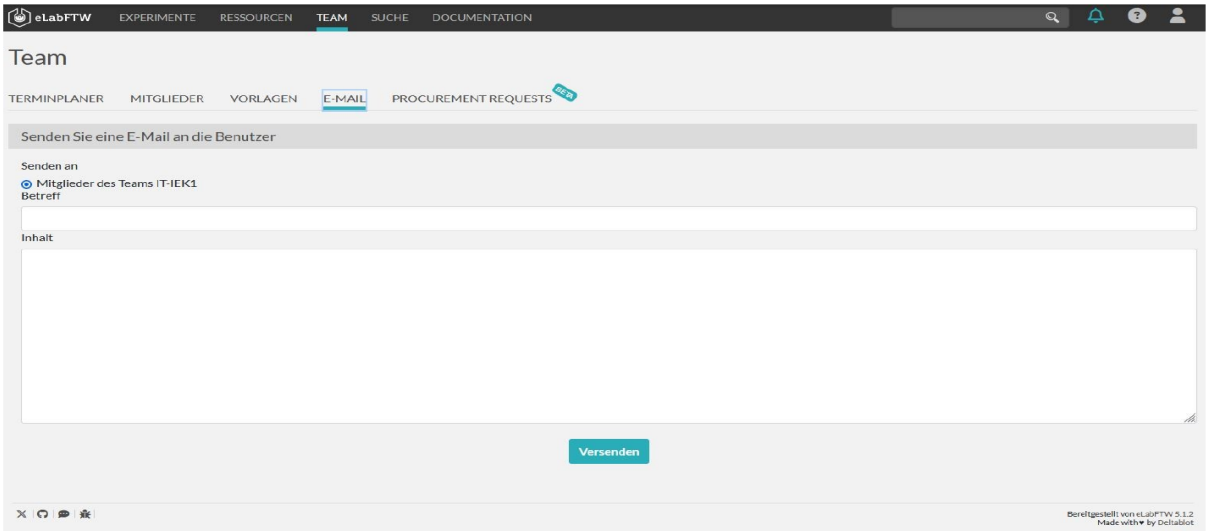


Abbildung 5.29.: Browser-Oberfläche von eLabFTW: Kommunikation im Team (Email-Funktion).

Team vorhandenen Templates gefunden werden. Hierzu muss der Unter-Menüpunkt SETTINGS angeklickt werden.

SUCHE

Die Suchfunktion in eLabFTW (siehe Abbildung 5.30) erlaubt eine sehr detaillierte Suche nach allen möglichen Begriffen bzw. Parametern. Da alle Experimente mit ihren Metadaten immer in der Datenbank von eLabFTW gespeichert werden, findet die Suchfunktion alle in einer Instanz gespeicherten Informationen bzw. Daten. So ist es unter anderem möglich nach dem Namen eines Experimentes zu suchen, zum Beispiel verknüpft mit einem Datum oder mit einer chemischen Komponente. Es ist auch eine einfache Suche mittels Tags möglich, was eine der effektivsten Suchmethoden darstellt.

RESSOURCEN

Ressourcen sind Experimenten ähnlich, aber sie erfüllen einen anderen Zweck. Ressourcen enthalten Listen und die Organisation von *Dingen*, die für ein Experiment notwendig und in einem Paket zusammengefasst sind. Diese Ressourcen können von den Nutzern gebucht werden, welche vorher von dem jeweiligen Admin erstellt wurden. Normale Nutzer können nicht selbst neue Ressourcen erstellen, sondern nur vorhandene nutzen. In einer Ressource kann jede Art von Elementen gespeichert werden: Ganze Projekte, Chemikalien, Geräte und Messapparaturen, Prüfstände, usw. Auf diese

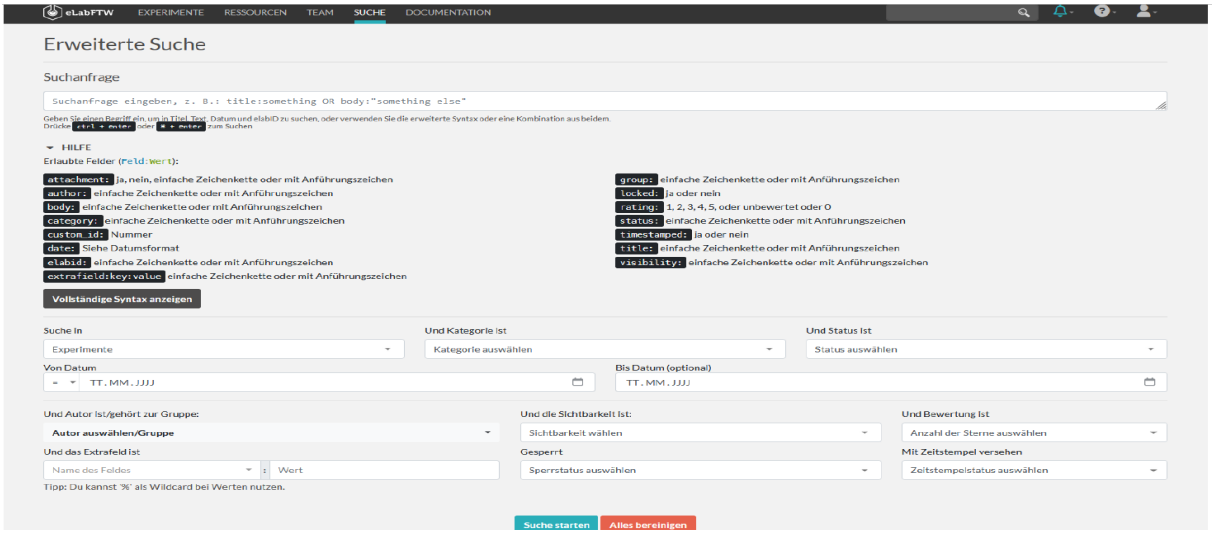


Abbildung 5.30.: Browser-Oberfläche von eLabFTW: Detaillierte Suchfunktion.

Weise kann zum Beispiel neuen Mitarbeitern in einem Paket zusammengefasst alles zur Verfügung gestellt werden, was sie für einen erfolgreichen Arbeitsbeginn benötigen.

PERSÖNLICHES PROFIL UND EINSTELLUNGEN

Durch das Anklicken des Buttons für das eigene Profil (rote Markierung oben rechts in Abbildung 5.31) erhält man die Möglichkeit, das System nach seinen eigenen Vorlieben zu konfigurieren. An dieser Stelle kann die Sprache gewählt werden (21 Sprachen sind möglich), oder die Darstellung auf dem Bildschirm. Man kann eigene Tastenkürzel definieren, um die Bedienung zu beschleunigen oder auch Ausgabeoptionen für den Export von PDF-Dateien auswählen. In Abbildung 5.31 ist ebenfalls zu sehen, dass über das persönliche Profil auch To-DO-Listen eingesehen werden können, sowohl für den jeweiligen Nutzer, wie auch für das gesamte Team. Man erhält auf diese Weise einen schnellen Überblick, welche Schritte in einem Experiment oder einer Ressource noch unerledigt sind.

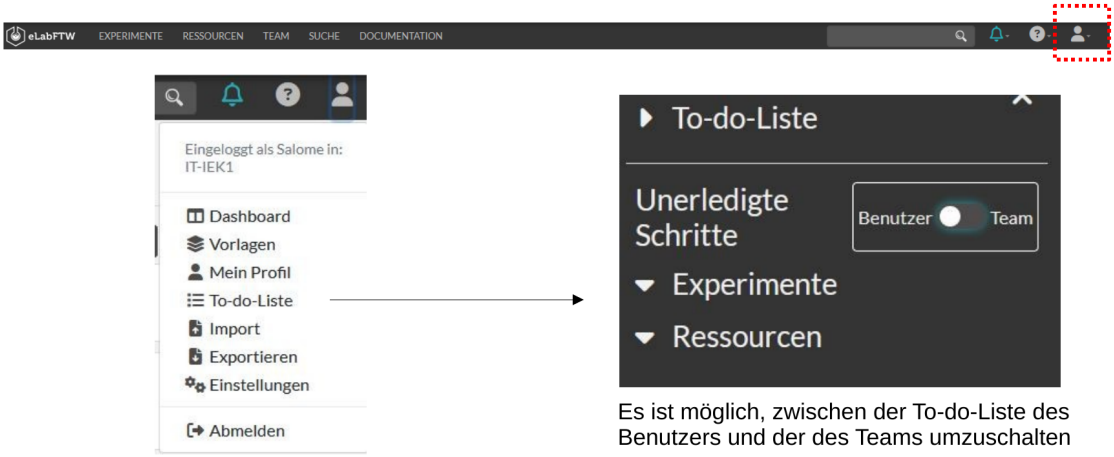


Abbildung 5.31.: Browser-Oberfläche von eLabFTW: Mögliche Einstellungen im persönlichen Profil.

TOOLS (NEU in Version 5.2.x von eLabFTW)

Abschließend möchten wir beispielhaft noch zwei nützliche Zusatzprogramme nennen, die über den Reiter TOOLS gestartet werden können. Das ist zum einen die Möglichkeit, Chemikalien, die für Experimente benötigt und verwendet werden, zu inventarisieren. Für solche Zwecke ist das Menü COMPOUNDS gedacht (siehe Abbildung 5.32). Ebenfalls nützlich kann, je nach Anwender, auch der Molekül-Editor sein (Abbildung 5.33), der in der aktuellen Version von eLabFTW 5.2.8 (Juli 2025) umfangreiche Möglichkeiten bietet, inklusive der Aufnahme von Molekülen und Verbindungen in die Datenbank der chemischen Komponenten (COMPOUNDS) sowie der weiteren Möglichkeit der Suche nach ähnlichen Strukturen in der Datenbank.

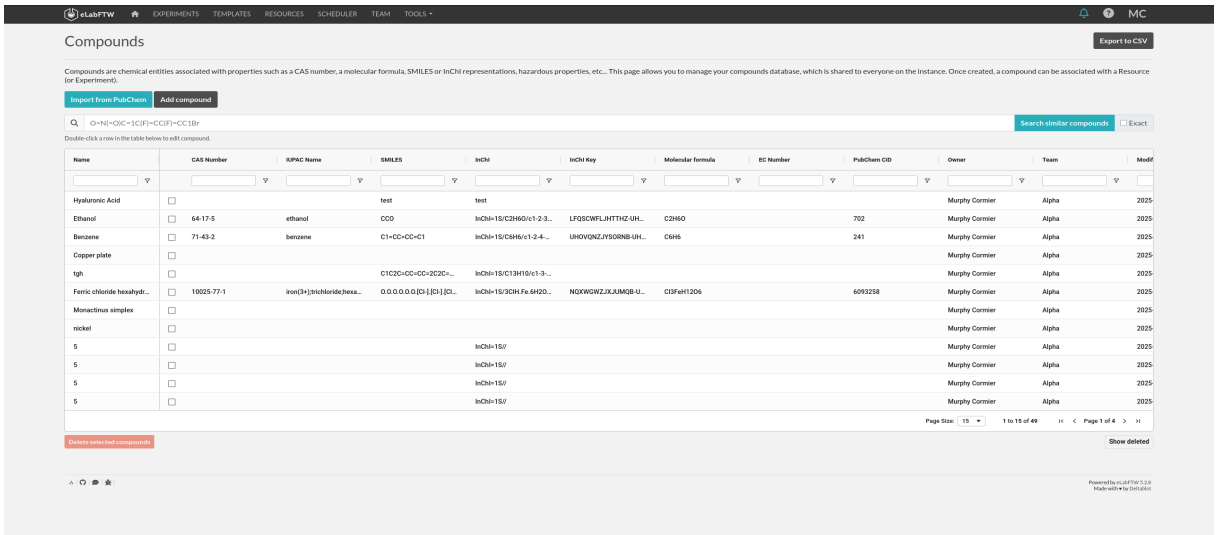


Abbildung 5.32.: Browser-Oberfläche von eLabFTW: Inventarisierung von Chemikalien.

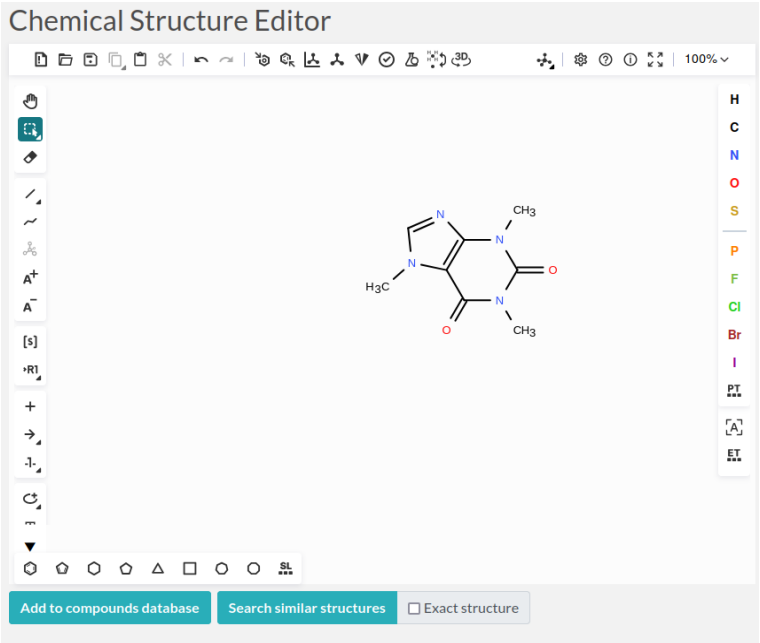


Abbildung 5.33.: Browser-Oberfläche von eLabFTW: Molekül-Editor.

5.3.3. Kadi4Mat

Kadi4Mat ist eine generische und quelloffene virtuelle Forschungsumgebung. Ursprünglich im Kontext der Materialwissenschaft entwickelt, kann Kadi4Mat für die Verwaltung jeglicher Art von Forschungsdaten innerhalb verschiedener Forschungsdisziplinen und Anwendungsfälle verwendet werden. Sein Ziel ist es, die Fähigkeit zur Verwaltung und zum Austausch von Daten, die Repository-Komponente, mit der Möglichkeit zur Analyse, Visualisierung und Transformation dieser Daten, der ELN-Komponente (Electronic Lab Notebook), zu kombinieren. Bei der Repository-Komponente liegt der Fokus auf warmen, d.h. unveröffentlichten und noch weiter zu analysierenden Daten, bei der ELN-Komponente auf der automatisierten und dokumentierten Ausführung von heterogenen Workflows über eine Application Programming Interface (API). Auf diese Weise wird ein anpassbarer Rahmen geschaffen, der gute Praktiken der Forschungsdatenverwaltung und die Zusammenarbeit zwischen Forschern erleichtert. Insofern geht Kadi4Mat noch über die Möglichkeiten der bereits vorgestellten ELN's JuliaBase und eLabFTW hinaus und bietet noch mehr Möglichkeiten. Im Folgenden werden wir die wesentlichen Komponenten und deren Anwendung knapp erläutern, damit interessierte Nutzer in die Lage versetzt werden, erfolgreich mit der Arbeit mit diesem System zu beginnen. Abbildung 5.34 veranschaulicht die Basis-Struktur von Kadi4Mat. Weitere Informationen finden sich unter anderem in den Publikationen von Brandt et al. (2021), Griem et al. (2023) und Al-Salman et al. (2023) [**Brandt**, **Griem**, **Al-Salman**]. Ähnlich wie für eLabFTW gibt es auch für Kadi4Mat ei-

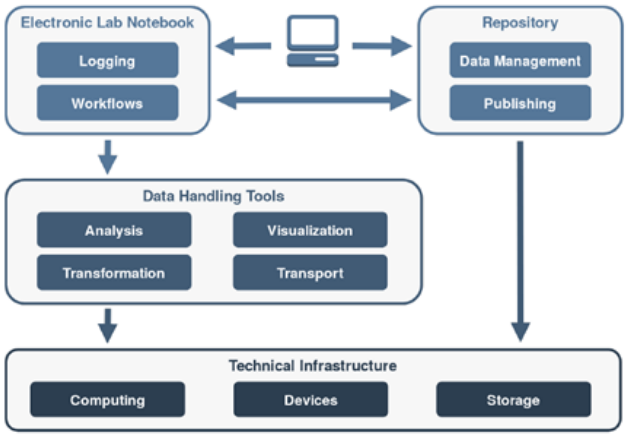


Abbildung 5.34.: Die Struktur von Kadi4Mat.

ne Demo-Version (`https://demo-kadi4mat.iam.kit.edu/`). Wir können diese Demo für das Kennenlernen des Systems nur dringend empfehlen, denn so werden viele Zusammenhänge fast spielerisch deutlich und man bekommt ein gutes Gefühl für die verschiedenen Menüs und das Zusammenspiel der verschiedenen Komponenten.
Kadi4Mat stellt tatsächlich ein ganzes Ökosystem an verschiedenen Programmen dar, welche aufeinander aufbauen und optimiert zusammenwirken. Abbildung 5.35 zeigt die verschiedenen (Basis-)Komponenten von Kadi4Mat, d.h. **Kadi-Web** (damit haben Sie als neuer Nutzer hauptsächlich zu tun), **Kadi-Studio**, **Kadi-AI** (Komponenten für künstliche Intelligenz), **Kadi-FS** und **Kadi-APY**. Zum Kennenlernen dieser Komponenten empfehlen wir die Demo-Version und die darin enthaltenen Hilfestellungen, oder direkt die Beantragung eines Nutzerkontos.

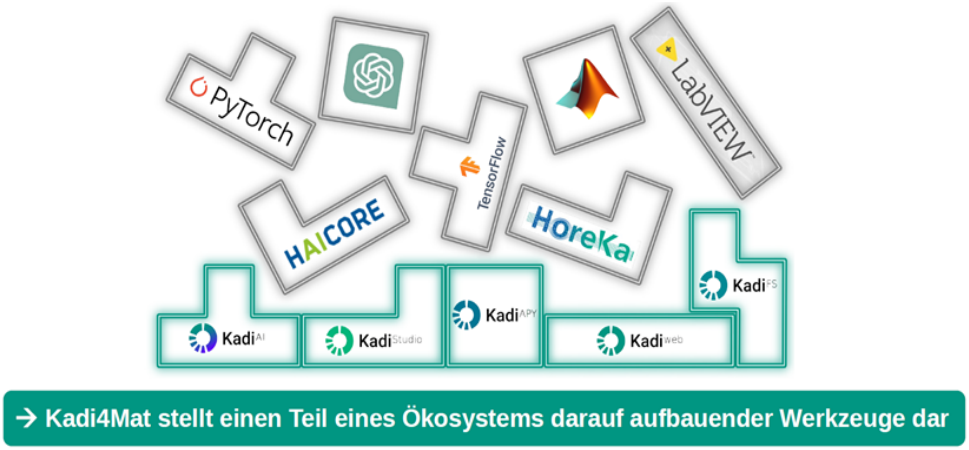


Abbildung 5.35.: Das Kadi4Mat-Ökosystem.

Login

Es gibt verschiedene Möglichkeiten, sich bei Kadi4Mat anzumelden. Jede Kadi4Mat-Instanz kann einen oder mehrere Authentifizierungsanbieter registriert haben, die im Folgenden kurz beschrieben werden. **1. Berechtigungsnachweise:** Diese Methode der Authentifizierung verwendet separate Konten, die für eine bestimmte Kadi4Mat-Instanz lokal sind. Die Credentials eines jeden Benutzers bestehen aus einem eindeutigen Benutzernamen und einem Passwort. Je nach Konfiguration einer Instanz kann die Registrierung neuer Konten durch einzelne Benutzer oder nur durch Systemadministratoren möglich sein. **2. LDAP:** Diese Authentifizierungsmethode ermöglicht die Anmeldung über bestehende Konten, die von einer bestimmten LDAP-Installation (Lightweight Directory Access Protocol) verwaltet werden. LDAP ist ein Verzeichnisdienst, der üblicherweise für die Authentifizierung von Benutzern und die Verwaltung verschiedener Arten von Informationen verwendet wird. Ein solcher Dienst kann auf verschiedenen Ebenen eingesetzt werden, z. B. für einzelne Arbeitsgruppen oder ganze Institutionen. **3. OpenID Connect:** Diese Authentifizierungsmethode ermöglicht es, sich mit bestehenden Konten bei einem oder mehreren webbasierten Diensten von Drittanbietern über das OpenID Connect-Authentifizierungsprotokoll anzumelden. Jeder Dienst muss von einem Systemadministrator separat in Kadi4Mat konfiguriert werden. Bei der erstmaligen Authentifizierung mit einem dieser Dienste müssen die Benutzer zustimmen, dass Kadi4Mat auf ihre Benutzerdaten zugreift. **4. Shibboleth:** Diese Authentifizierungsmethode ermöglicht es, sich mit bestehenden Konten der Heimatinstitution eines Benutzers über das Shibboleth-Authentifizierungsprotokoll anzumelden. Jede Einrichtung muss separat in Kadi4Mat von einem Systemadministrator konfiguriert werden. Es ist zu beachten, dass die Identity Provider mancher Institutionen nicht alle benötigten Benutzerattribute standardmäßig an Kadi4Mat weitergeben. In diesem Fall muss ggf. der Administrator des Shibboleth Identity Providers kontaktiert werden. Befindet man sich im System, so sieht man zunächst die folgende Browser-Oberfläche (Abbildung 5.36): In der obersten Leiste mit weißer Schrift auf schwarzen Grund, finden sich (ähnlich wie bei eLabFTW) verschiedene Reiter, unter denen sich entsprechende Menüs verbergen. Abbildung 5.37 hebt diese Menüs zur Verdeutlichung noch einmal besonders hervor. Auf die einzelnen Menüs werden wir im Folgenden noch einmal detailliert eingehen.

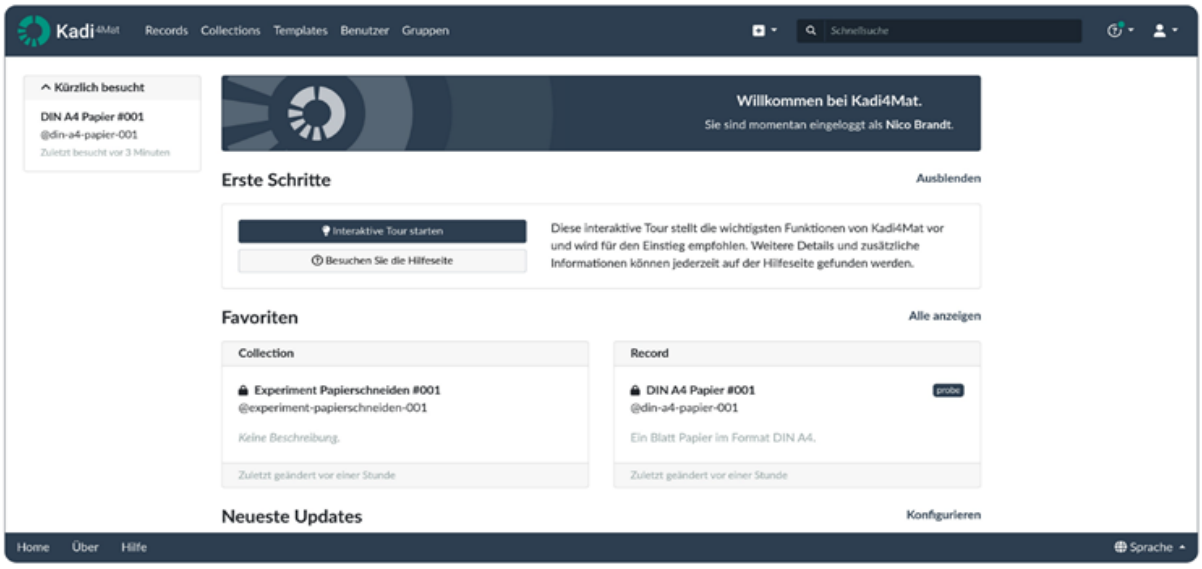


Abbildung 5.36.: Die Oberfläche von Kadi4Mat.

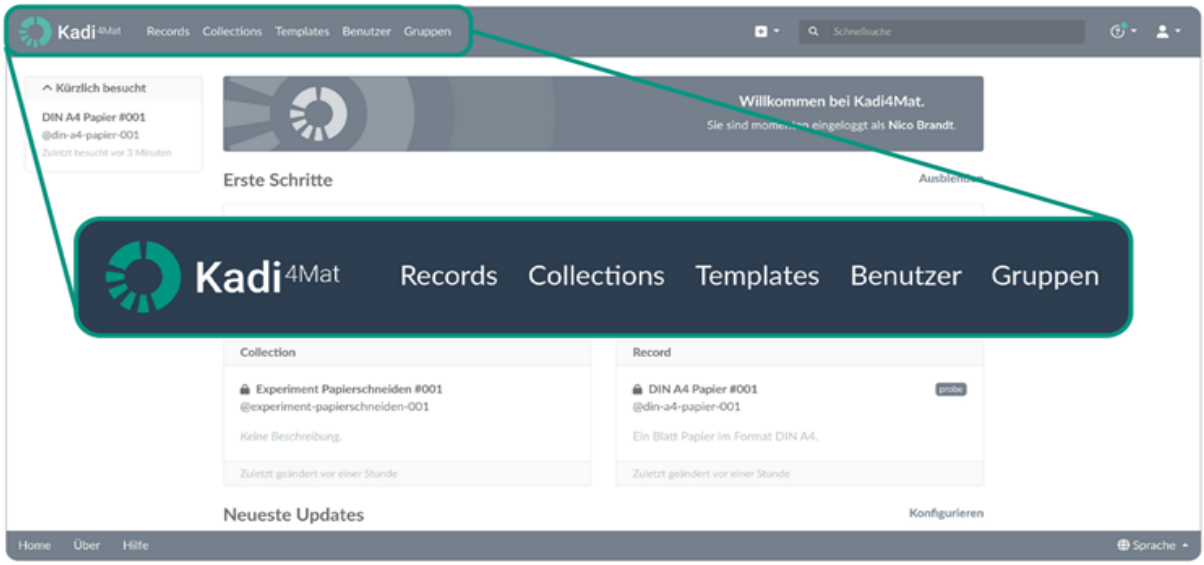


Abbildung 5.37.: Die Menüs in der Oberfläche von Kadi4Mat.

Navigation

Nach dem Einloggen ermöglicht das Menü oben links in der Navigationsleiste (Abbildung 5.37) den Zugriff auf Datensätze (Records), Sammlungen (Collections), Vorlagen (Templates), Benutzer und Gruppen. Jeder Navigationspunkt führt zu der entsprechenden Hauptseite der jeweiligen Ressource, auf der bestehende Ressourcen durchsucht und gegebenenfalls neue Ressourcen erstellt werden können. Details zu den einzelnen Ressourcentypen werden in den folgenden Abschnitten beschrieben. Die ersten beiden Punkte oben rechts in der Navigationsleiste bieten ebenfalls einen schnellen Zugriff auf die Erstellung neuer Ressourcen und die Suche nach bestehenden Ressourcen. Über die beiden Dropdown-Menüs ganz rechts können Sie schnell zu verschiedenen Informationsseiten navigieren und auf die Profilseite des aktuellen Benutzers zugreifen, einschließlich seiner erstellten und gelöschten Ressourcen (siehe auch Benutzer). Letzteres ermöglicht außerdem den Zugriff auf die Einstellungen und das Abmelden. Zusätzlich zur Navigationsleiste gibt es auf allen Seiten unten eine Navigationsleiste, unabhängig davon, ob man eingeloggt ist oder nicht. Diese Fußzeile ermöglicht eine schnelle Navigation zu verschiedenen Informationsseiten und enthält auch eine Auswahl zum Umschalten der aktuellen Sprache.

Datensätze

Für die Entstehung und Weiterverarbeitung von Daten bzw.Datensätzen in Kadi4Mat möchten wir diesen Abschnitt mit einem einfachen Beispiel beginnen. Unser Gerät für die Durchführung eines Experimentes soll eine einfache Schere sein und unsere Probe ist ein Blatt Papier im DIN A4-Format. Wir zerschneiden das Papier und erhalten zwei neue Proben, jeweils in der Größe eines halben DIN A4-Blattes (Abbildung 5.38). Der zugehörige Graph in Kadi4Mat für dieses einfache Experimente in

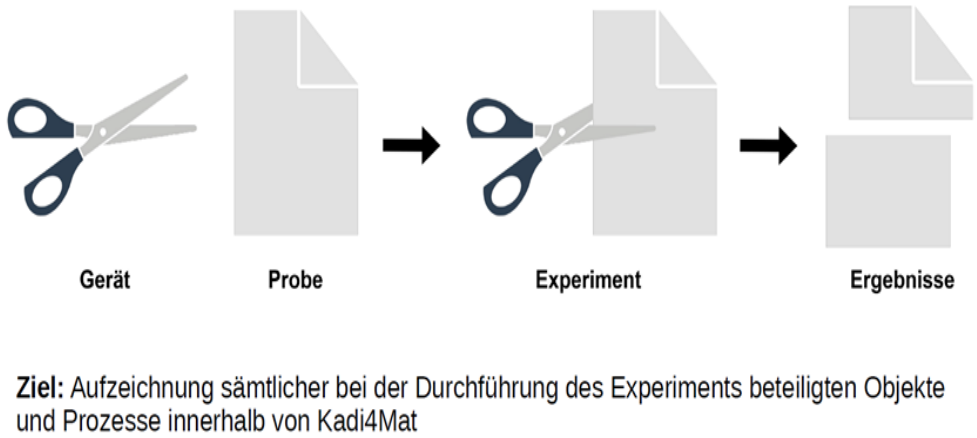


Abbildung 5.38.: Ein einfaches Experiment.

Abbildung 5.39 zeigt, wie das System ein solches Experiment strukturiert. Wir führen ein Experiment mit dem Namen *Papierschneiden* durch, verwenden dazu ein Gerät mit dem Namen *Schere*, bearbeiten eine Probe (*DIN A4-Papier*), und erzeugen als Resultat zwei neue Proben, nämlich zwei halbe Blätter Papier. Abbildung 5.40 zeigt die generelle Struktur, wie Daten in Kadi4Mat erzeugt und bearbeitet werden. Die Metadaten für das *Papierschneiden* zeigt die Abbildung 5.41. Wie bereits mehrfach in diesem Handbuch ausgeführt, gehören experimentelle (Mess-)Daten und die entsprechenden Metadaten immer zusammen. In Abbildung 5.40 können wir erkennen, wie beide gemeinsam von Kadi4Mat analysiert, verarbeitet und ggf. auch publiziert bzw. weiter verwendet werden.

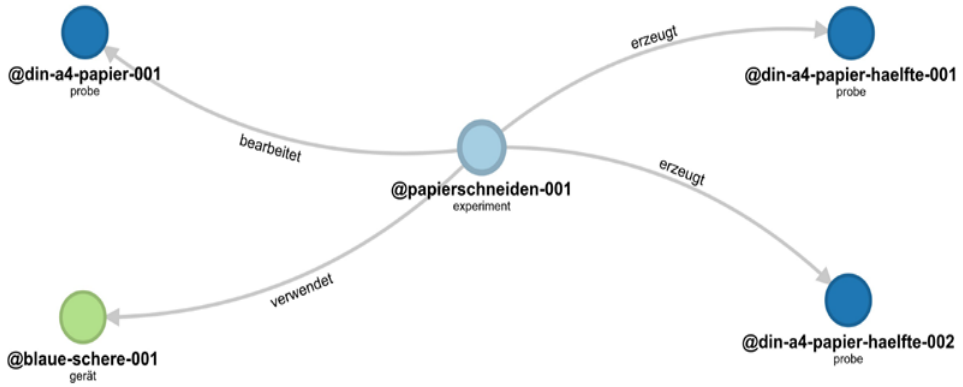


Abbildung 5.39.: Der Graph in Kadi4Mat für das einfache Experimente in Abbildung 5.38.

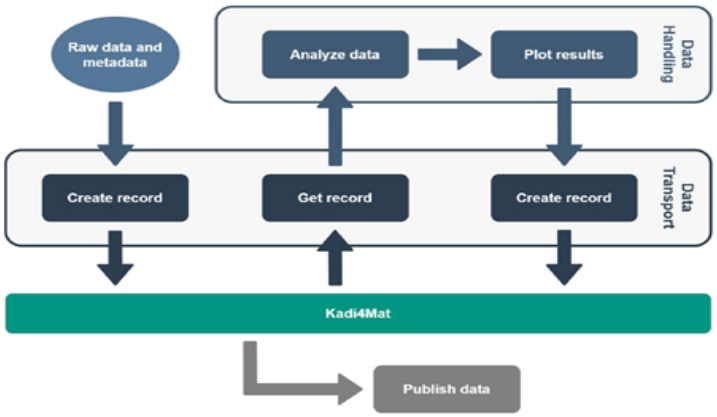


Abbildung 5.40.: Die Erzeugung und Verarbeitung von Datensätzen in Kadi4Mat.

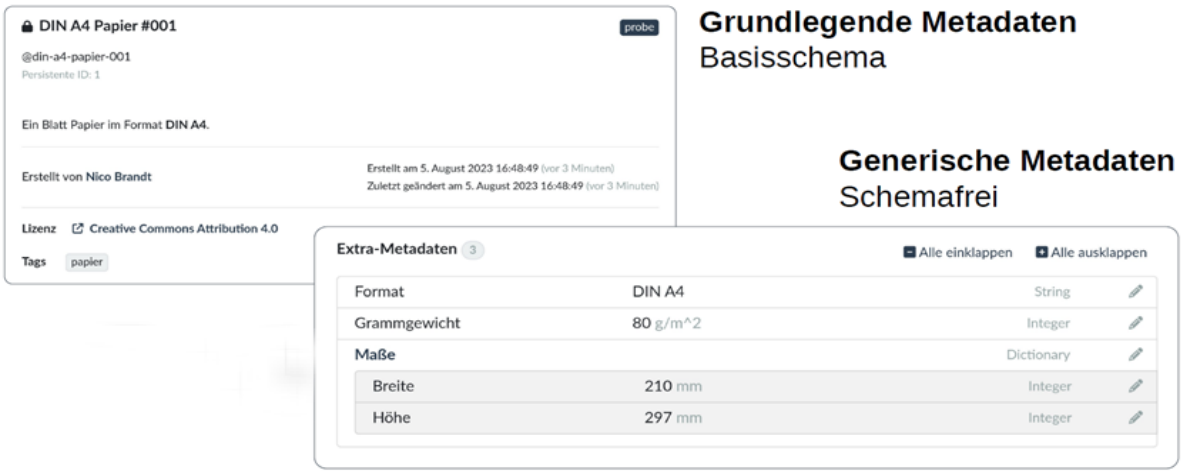


Abbildung 5.41.: Die Metadaten des einfachen Experimentes *Papierschneiden*.

Datensätze sind die Basiskomponenten von Kadi4Mat und können jede Art von digitalen oder digitalisierten Objekten repräsentieren, z.B. beliebige Forschungsdaten, Proben, Versuchsgeräte oder auch einzelne Bearbeitungsschritte. Datensätze bestehen aus Metadaten, die entweder für sich alleine stehen oder mit einer beliebigen Anzahl von korrespondierenden Daten verknüpft werden können. Sie können auch in Sammlungen gruppiert oder mit anderen Datensätzen verknüpft werden, wie später beschrieben. Um einen neuen Datensatz zu erstellen, müssen zunächst die Metadaten eingegeben werden. Dazu gehören grundlegende Informationen wie Titel, (eindeutiger) Bezeichner und Beschreibung. Darüber hinaus können generische zusätzliche Metadaten angegeben werden, die für die verschiedenen Arten von Datensätzen spezifisch sind. Diese Metadaten bestehen aus erweiterten Schlüssel-Wert-Paaren, wobei jeder Eintrag mindestens einen eindeutigen Schlüssel, einen Typ und einen entsprechenden Wert hat. Optional können auch eine Beschreibung, ein zusätzlicher Begriff IRI (Internationalized Resource Identifier) und Validierungsanweisungen angegeben werden. Es ist auch möglich, Templates für die generischen Metadaten zu erstellen, wie unter Templates beschrieben. Die folgenden Wertetypen können für diese Metadaten verwendet werden:
String - Ein einzelner Textwert.
Ganzzahl - Ein einzelner Ganzzahlwert. Begrenzt auf Werte zwischen -$(2^{53}-1)$ und $(2^{53}-1)$. Integer-Werte können optional eine Einheit haben, die sie weiter beschreibt.
Float - Ein einzelner Gleitkommawert mit doppelter Genauigkeit (64 Bit). Float-Werte können optional mit einer Einheit versehen werden, die sie näher beschreibt.
Boolean - Ein einzelner boolescher Wert, der entweder wahr oder falsch sein kann.
Datum - Ein einzelner Datums- und Zeitwert.
Dictionary - Ein verschachtelter Wert, der verwendet werden kann, um mehrere Metadateneinträge unter einem einzigen Schlüssel zu kombinieren.
Liste - Ein verschachtelter Wert, der ähnlich wie Wörterbücher funktioniert, mit dem Unterschied, dass keiner der Werte in einer Liste einen Schlüssel hat.
Abgesehen von den Metadaten ist es möglich, die Sichtbarkeit eines Datensatzes entweder auf privat oder öffentlich zu setzen, wobei letzteres jedem angemeldeten Benutzer die Möglichkeit gibt, nach dem Datensatz zu suchen und seinen Inhalt einzusehen, ohne dass dafür explizite Leseberechtigungen erforderlich sind. Schließlich kann ein Datensatz direkt mit anderen Ressourcen verknüpft werden, und seine Berechtigungen können direkt festgelegt werden, wobei beide Aspekte auch nach der Erstellung des Datensatzes verwaltet werden können. Sobald die Metadaten eines Datensatzes erstellt wurden, können die eigentlichen Daten des Datensatzes in einer separaten Ansicht hinzugefügt werden, zu der die Anwendung standardmäßig automatisch weiterleitet. Dies ist nur ein Teil der verschiedenen Ansichten zur Verwaltung von Datensätzen. Im nächsten Abschnitt wird der Zweck der anderen Ansichten beschrieben, die ausgewählt werden können, nachdem Sie zur Datensatzübersichtsseite zurückgekehrt sind. Für die Verwaltung bestehender Datensätze stehen auf der Datensatzübersichtsseite verschiedene Ansichten zur Verfügung, die jeweils über die entsprechende Registerkarte im Navigationsmenü eines Datensatzes aufgerufen werden können. Die einzelnen Registerkarten und ihre Inhalte werden im Folgenden kurz beschrieben:

Überblick: Diese Registerkarte bietet einen Überblick über einen Datensatz, hauptsächlich in Bezug auf seine Metadaten. Hier ist es möglich, einen Datensatz zu bearbeiten oder zu kopieren, wenn die entsprechenden Berechtigungen erfüllt sind, wobei das Bearbeiten eines Datensatzes auch das Löschen desselben ermöglicht. Beachten Sie, dass der Datensatz dabei zuerst in den Papierkorb verschoben wird; siehe auch Benutzer. Außerdem können Datensätze in verschiedenen Formaten exportiert, veröffentlicht oder favorisiert werden. Beachten Sie, dass die Veröffentlichungsfunktion nur verfügbar ist, wenn mindestens ein Veröffentlichungsanbieter bei der Anwendung registriert ist.

Dateien: Diese Registerkarte bietet einen Überblick über die mit einem Datensatz verbundenen Dateien. Mit den entsprechenden Berechtigungen können neue Dateien hinzugefügt werden, was in der Regel durch das Hochladen lokal gespeicherter Dateien geschieht. Bestimmte Dateitypen können jedoch auch direkt über die Webschnittstelle erstellt werden. Bestehende Dateien können entweder als Ganzes oder einzeln über die Schnellnavigation der jeweiligen Datei heruntergeladen werden. Je nach Berechtigung zeigt diese Navigation auch zusätzliche Aktionen zur schnellen Verwaltung von Dateien an. Ein Klick auf eine Datei führt zu einer separaten Übersichtsseite der entsprechenden Datei, die alle zusätzlichen Metadaten der Datei anzeigt. Darüber hinaus verfügen viele Dateitypen über eine integrierte Vorschaufunktion. Hier ist es auch möglich, die Metadaten oder den Inhalt einer Datei zu bearbeiten, wobei die Bearbeitung der Metadaten auch das Löschen der Datei ermöglicht. Bei einigen Dateitypen ist eine direkte Bearbeitung des eigentlichen Dateiinhalts möglich, ansonsten kann die reguläre Upload-Funktionalität genutzt werden.

Verknüpfungen: Diese Registerkarte bietet einen Überblick über die Ressourcen, mit denen ein Datensatz verknüpft ist, einschließlich anderer Datensätze und Sammlungen. Sammlungen stellen logische Gruppierungen von mehreren Datensätzen dar, während Verknüpfungen zwischen Datensätzen deren Beziehung spezifizieren und auch zusätzliche Metadaten enthalten können. Darüber hinaus können Datensatzverknüpfungen in einem interaktiven Diagramm visualisiert werden. Ein Klick auf einen Datensatz-Link führt zu einer separaten Ansicht, die einen detaillierteren Überblick über den Link und die zugehörigen Datensätze bietet. Die Verknüpfung von Ressourcen erfordert eine Verknüpfungserlaubnis in beiden Ressourcen, die miteinander verknüpft werden sollen. Beachten Sie, dass Benutzer weiterhin nicht in der Lage sind, verknüpfte Ressourcen einzusehen, wenn sie keine explizite Berechtigung dazu haben. Eine begrenzte Anzahl von Informationen über Datensatzverknüpfungen (die ID der Verknüpfung, der verknüpfte Datensatz, der Name und der Term IRI der Verknüpfung) wird jedoch immer als Teil der Datensatzüberarbeitungen angezeigt.

Berechtigungen: Diese Registerkarte bietet einen Überblick über die Zugriffsberechtigungen, die einzelnen Benutzern oder Gruppen von mehreren Benutzern für einen bestimmten Datensatz gewährt werden. Neue Berechtigungen können erteilt werden, wenn die entsprechenden Berechtigungen dafür erfüllt sind, was derzeit über vordefinierte Rollen funktioniert. Details zu den spezifischen Berechtigungen und Aktionen, die jede Rolle bietet, finden Sie, wenn Sie auf das Popover Rollen klicken. Beachten Sie, dass Gruppenrollen für Benutzer, die Berechtigungen verwalten können, immer angezeigt werden, auch wenn die Gruppe normalerweise nicht sichtbar wäre. Auf diese Weise können bestehende Gruppenrollen jederzeit geändert und/oder entfernt werden. Solche Gruppenrollen enthalten nur sehr begrenzte Informationen über die Gruppe selbst (ihre ID, ihren Titel, ihren Bezeichner und ihre Sichtbarkeit).

Überarbeitungen: Diese Registerkarte bietet einen Überblick über die Änderungen an den Metadaten eines Datensatzes, an den Metadaten einer Datei und an den Verknüpfungen zu anderen Datensätzen. Wenn Sie auf Revision eines Revisionseintrags anzeigen klicken, wird eine separate Ansicht geöffnet, die einen detaillierteren Überblick über die jeweilige Revision und die entsprechenden Änderungen bietet.

Änhlich wie bei anderen ELN's können auch in Kadi4Mat Datensätze in verschiedene Dateiformate exportiert werden (Abbildung 5.42). Zum Beispiel in die Formate JSON, PDF, QR-Code, RDF (Turtle) und RO-Crate. Für Detailfragen werfen Sie bitte einen Blick in die entsprechenden Hilfe-Funktionen in Kadi4Mat.

Abbildung 5.42.: Export von Datensätzen (Records).

Sammlungen

Sammlungen stellen logische Gruppierungen (z. B. Projekte, Simulationsstudien oder Experimente) von mehreren Datensätzen oder anderen Sammlungen dar. Das Anlegen einer neuen Sammlung ist ähnlich wie das Anlegen neuer Datensätze, mit dem Unterschied, dass nur einige grundlegende Metadaten erforderlich sind. Darüber hinaus ist es möglich, eine Datensatzvorlage anzugeben, die beim Hinzufügen neuer Datensätze zu einer Sammlung als Standard verwendet wird. Beachten Sie, dass eine begrenzte Untermenge an Informationen über solche Vorlagen immer als Teil der Sammlungsrevisionen (die ID der Vorlage) und bei der Bearbeitung der Sammlung (die ID und der Bezeichner der Vorlage) angezeigt wird.
Ähnlich wie bei Datensätzen haben Sammlungen ihr eigenes Navigationsmenü, das verschiedene Registerkarten zur Anzeige und Verwaltung von Sammlungen bietet. Da die meisten Inhalte denen von Datensätzen ähnlich sind, werden in den folgenden Abschnitten nur die wichtigsten Unterschiede aufgeführt:

Übersicht: Diese Registerkarte bietet einen Überblick über eine Sammlung, hauptsächlich in Bezug auf ihre Metadaten. Außerdem werden die Datensätze, die Teil einer Sammlung sind, direkt in dieser Übersicht aufgeführt.

Verknüpfungen: Neben der Verknüpfung von Sammlungen mit Datensätzen können Sammlungen auch mit anderen Sammlungen verknüpft werden, was auf dieser Registerkarte angezeigt wird. Diese Funktion kann verwendet werden, um einfache Hierarchien von übergeordneten und untergeordneten Sammlungen zu erstellen, um die Strukturierung mehrerer Ressourcen zu verbessern, z. B. durch die Darstellung von Projekten und entsprechenden Unterprojekten. Beachten Sie, dass jede Sammlung nur eine übergeordnete Sammlung haben kann, während die erforderlichen Berechtigungen zur Verknüpfung von Sammlungen ähnlich wie bei anderen Ressourcenverknüpfungen gehandhabt werden.

Berechtigungen: Neben der Verwaltung der Berechtigungen von Sammlungen selbst ist es auch möglich, die Rollen von Benutzern und Gruppen aller verknüpften Datensätze in einer Sammlung auf der entsprechenden Registerkarte zu verwalten, da die Sammlungsberechtigungen derzeit nicht an ver-

knüpfte Ressourcen vererbt werden. Dies gilt insbesondere für alle verknüpften Datensätze, bei denen der aktuelle Benutzer Berechtigungen verwalten kann. Wenn Sie eine leere Rolle auswählen, werden stattdessen alle bestehenden Berechtigungen des entsprechenden Benutzers oder der Gruppe entfernt.

Templates

Vorlagen ermöglichen die Erstellung von Entwürfen für verschiedene Ressourcen. Es gibt mehrere Arten von Vorlagen, die den tatsächlichen Inhalt definieren, den eine Vorlage enthalten kann. Derzeit gibt es die folgenden Arten von Vorlagen:

Datensatz: Datensatzvorlagen können alle Metadaten enthalten, die für einen Datensatz angegeben werden können, einschließlich seiner allgemeinen zusätzlichen Metadaten, verknüpften Sammlungen, Datensatzlinks und Berechtigungen. Beachten Sie, dass bei verknüpften Ressourcen oder Gruppen deren IDs für alle Benutzer, die auf die Vorlage zugreifen können, sichtbar sind. Datensatzvorlagen können an den meisten Stellen ausgewählt werden, an denen neue Datensätze erstellt werden können, sowie für die Erstellung von allgemeinen zusätzlichen Metadaten.

Extras: Extras-Vorlagen sind auf die allgemeinen zusätzlichen Metadaten eines Datensatzes ausgerichtet. Diese Art von Vorlagen kann überall dort ausgewählt und kombiniert werden, wo solche Metadaten angegeben werden können, auch bei der Erstellung anderer Vorlagen.
Die Erstellung einer neuen Vorlage ähnelt der Erstellung anderer Arten von Ressourcen. Für jede Vorlage müssen zumindest ein Titel und ein Bezeichner für die Vorlage selbst festgelegt werden, während die eigentlichen Vorlagendaten von der jeweiligen Art der Vorlage abhängen. Ähnlich wie bei anderen Ressourcen gibt es für Vorlagen ein eigenes Navigationsmenü mit verschiedenen Registerkarten zur Anzeige und Verwaltung von Vorlagen.

Nutzer

In dieser Ansicht werden alle registrierten Benutzer der aktuellen Kadi4Mat-Instanz angezeigt. Ein Klick auf einen Benutzer führt zu einer separaten Seite, die ein weiteres Navigationsmenü enthält, das Zugang zu verschiedenen Unterseiten bietet. Die jeweiligen Inhalte dieser Seiten werden im Folgenden kurz beschrieben:

Profil: Auf dieser Seite werden die grundlegenden Informationen eines Benutzers angezeigt, z. B. der Benutzername und der Kontotyp. Die Benutzer können einige der auf dieser Seite angezeigten Informationen über die Einstellungen steuern.

Ressourcen: Auf dieser Seite werden alle zugänglichen Ressourcen angezeigt, die ein bestimmter Benutzer erstellt hat. Wenn Sie andere Benutzer betrachten, werden auch Ressourcen angezeigt, die (explizit) mit einem Benutzer geteilt wurden, entweder direkt oder über Gruppen, einschließlich gemeinsamer Gruppen.

Papierkorb: Diese Seite ist nur für den aktuellen Benutzer sichtbar. Hier finden sich gelöschte Ressourcen, nämlich gelöschte Datensätze, Sammlungen, Vorlagen oder Gruppen. Die Ressourcen können entweder wiederhergestellt oder endgültig gelöscht werden. Letzteres geschieht ebenfalls automatisch nach 1 Woche. Bis dahin kann die Kennung der gelöschten Ressourcen nicht für neu erstellte Ressourcen wiederverwendet werden. Beachten Sie, dass derzeit nur der Ersteller einer Ressource diese wiederherstellen oder dauerhaft löschen kann.

Gruppen

Mit Gruppen können mehrere Benutzer zusammengefasst werden, was vor allem die Zugangsverwaltung erleichtert. Das Anlegen einer neuen Gruppe erfolgt ähnlich wie das Anlegen anderer Ressourcen. Neben den grundlegenden Metadaten kann auch ein Gruppenbild hochgeladen werden, das in der Gruppenübersicht und auf den Suchergebnisseiten angezeigt wird. Die Verwaltung bestehender Gruppen ist ähnlich wie bei anderen Ressourcen. Neben den üblichen Inhalten gibt es zusätzliche Registerkarten, die die Anzeige von Ressourcen ermöglichen, die mit einer Gruppe geteilt wurden:

Überblick: Diese Registerkarte bietet einen Überblick über eine Gruppe, hauptsächlich in Bezug auf ihre Metadaten. Außerdem werden die Mitglieder einer Gruppe in dieser Übersicht angezeigt. Die Verwaltung der Mitglieder einer Gruppe ist vergleichbar mit der Verwaltung der Zugriffsberechtigungen anderer Ressourcen, da die Gruppenzugehörigkeit mit den Zugriffsberechtigungen einer Gruppe verknüpft ist. Solange ein Benutzer eine beliebige Rolle in einer Gruppe innehat, ist er auch Mitglied dieser Gruppe, wobei er natürlich auch die entsprechenden Rechte hat, die seine Rolle innerhalb der Gruppe vorsieht. Neben der direkten Verwaltung von Mitgliedern ist es auch möglich, die Zuweisung von Rollen innerhalb einer Gruppe zu automatisieren, indem Regeln mit verschiedenen Bedingungen festgelegt werden. Jede Regel wird bei der Registrierung eines neuen Benutzers angewendet, kann aber auch rückwirkend auf alle bestehenden Benutzer angewendet werden.

Einstellungen

In den Einstellungen können Sie alles verwalten, was nicht direkt mit der eigentlichen Erstellung und Verwaltung von Ressourcen zu tun hat. Das Navigationsmenü der Einstellungen bietet wiederum Zugang zu verschiedenen Unterseiten:

Profil: In diesem Menü können Sie die grundlegenden Benutzerinformationen ändern, die im Profil eines Benutzers angezeigt werden. Beachten Sie, dass einige Optionen je nach Art des Benutzerkontos deaktiviert sein können.

Passwort: In diesem Menü können die Benutzer ihr Passwort ändern. Je nach Art des Benutzerkontos kann dieser Menüpunkt ausgeblendet sein.

Voreinstellungen: In diesem Menü können die Benutzer ihre Einstellungen bezüglich des Verhaltens oder des Aussehens von Kadi4Mat ändern.

Zugangs-Tokens: In diesem Menü können persönliche Zugangstoken (PATs) erstellt und verwaltet werden. Die Erstellung eines PATs ermöglicht die direkte Interaktion mit der HTTP-API, die Kadi4Mat zur Verfügung stellt. Detaillierte Informationen über die API und PATs sind in der Dokumentation von Kadi4Mat zu finden. Einige Teile der API können auch direkt über den Webbrowser getestet werden, indem man zu `https://demo-kadi4mat.iam.kit.edu/api/v1` navigiert, während eine OpenAPI-Spezifikation der API über `https://demo-kadi4mat.iam.kit.edu/openapi.json?v=v1` abgerufen werden kann. Beachten Sie, dass Sie für beide Links eingeloggt sein müssen.

Anwendungen: In diesem Menü können Sie registrierte oder autorisierte Anwendungen registrieren und verwalten. Eine Anwendung kann verwendet werden, um einen anderen Dienst mit der HTTP-API zu integrieren, die Kadi4Mat zur Verfügung stellt. Für die Autorisierung werden OAuth2-Tokens verwendet, die eine Anwendung durch Implementierung des entsprechenden OAuth2-Autorisierungsflusses anfordern kann. Detaillierte Informationen zur API und zu den OAuth2-Tokens finden sich in der Dokumentation von Kadi4Mat.

Verbundene Dienste: In diesem Menü können die Benutzer ihre Verbindungen mit verschiedenen Drittanbieterdiensten verwalten. Jeder Dienst muss als Plugin in der Anwendung registriert werden und kann für verschiedene Aufgaben verwendet werden, z.B. für die Veröffentlichung von Ressourcen. Wenn in der aktuellen Kadi4Mat-Instanz keine Dienste verfügbar sind, wird dieser Menüpunkt ausgeblendet.

Die Publikation von Datensätzen

Kadi4Mat bietet eine besonders komfortable Möglichkeit zur Publikation von Datensätzen. Direkt aus dem System heraus können Datensätze bei Zenodo (`https://zenodo.org`) hochgeladen und publiziert werden, d.h. sie erhalten dort eine eigene DOI, auf die später referenziert werden kann. Sobald ein Benutzerkonto in Kadi4Mat mit einem entsprechenden Nutzerkonto bei Zenodo verbunden ist, kann ein Upload erfolgen (Abbildung 5.43). Mehr Informationen dazu liefern wir im Kapitel *Datenpublikation*.

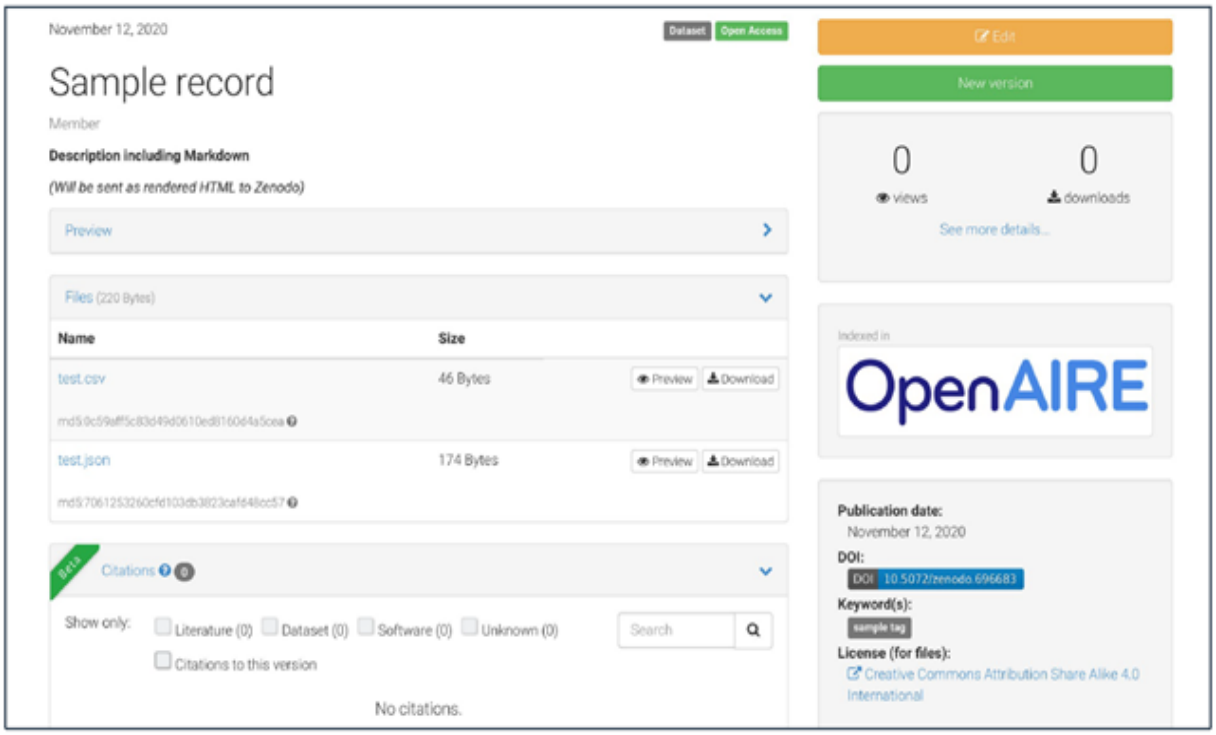


Abbildung 5.43.: Publikation von Datensätzen bei Zenodo.

Die Suchfunktionen in Kadi4Mat

Vergleichbar mit den Suchfunktionen von zum Beispiel eLabFTW, so bietet auch Kadi4Mat umfangreiche Möglichkeiten, um nach Informationen bzw. Daten zu suchen. Durch die Verwendung von Filtern kann die Suche optimiert und beschleunigt werden. Abbildung 5.44 vermittelt einen optischen Eindruck, wie direkt aus dem Browser heraus eine Suche gestaltet und gestartet werden kann.

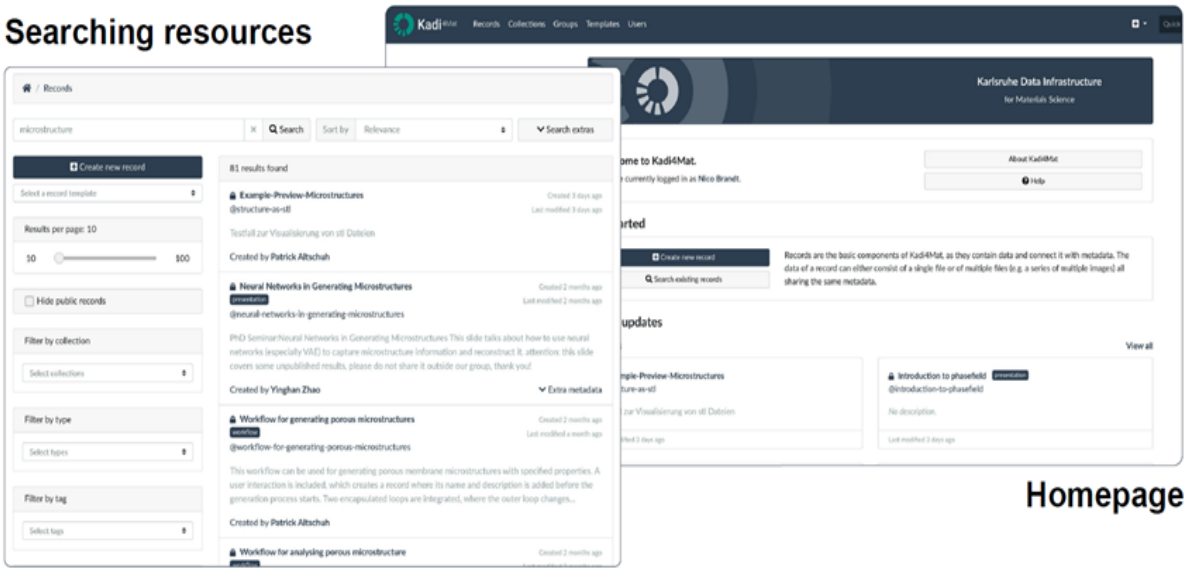


Abbildung 5.44.: Die Suchfunktionen von Kadi4Mat.

6. Datenqualität

In den letzten Kapiteln haben wir demonstriert, wie man systematisch Forschungsdaten erfassen kann. Nun muss es darum gehen, die Datenqualität sicher zu stellen. Dieser Punkt ist sehr entscheidend für die weitere Verwendung von Forschungsdaten, sei es für die Entwicklung von Folgeprodukten oder auch für die Analyse mittels KI. Generell gilt "*Garbage in, garbage out*", d.h. die Qualität des Dateninputs entscheidet über die Qualität des Ergebnisses.
Bei der Datenanalyse müssen wir zwischen systematischen und statistischen Fehlern von Messdaten bzw. Forschungsdaten unterscheiden. Systematische Fehler können zum Beispiel durch Ungenauigkeiten eines Messgerätes oder durch Bedienungsfehler entstehen. Idealerweise können solche systematischen Fehler durch sog. Ausreißer (Outlier) im Gesamt-Datensatz erkannt werden. Die Autoren dieses Handbuchs arbeiten aktuell im Rahmen Ihrer Tätigkeiten im Konsortium NFDI4ING an einer KI-gestützen, intrinsischen Datenanalyse, welche automatisiert und direkt in einem elektronischen Laborbuch abläuft, ohne dass Nutzerinnen und Nutzer selbst tätig werden müssen. Das Ziel ist das automatische Auffinden von Outliern in Messdaten, wobei der Anwender über diese Outlier informiert wird und entscheiden kann, wie mit solchen Daten weiter verfahren werden soll. Da sich diese Arbeiten aktuell noch in der Entwicklung befinden, werden wir erst in einer Folgeausgabe dieses Handbuches ein entsprechende Kapitel integrieren können.
Datenanalyse kann sehr umfangreich sein und umfasst unter anderem die mathematischen Gebiete der linearen Algebra, Statistik und Wahrscheinlichkeitsrechnung. Auch wenn es heute große und etablierte Softwarepakete (kommerziell oder Open-Source) auf dem Markt gibt, die einem viel Arbeit abnehmen können, so muss man bei der Verwendung dennoch verstehen, wie solche Auswertungen funktionieren und wie die Ergebnisse zustande kommen. Wer sich tiefer in die Materie einarbeiten möchte, dem empfehlen wir die beiden folgenden Lehrbücher von Thomas Nield [**Nield**] und das Standardwerk von Lothar Papula [**Papula**] als verständliche Einführungen zum Selbststudium.

Kommerzielle Software für die Analyse von Forschungsdaten kann zum Beispiel das etablierte Programm *Origin*® (`https://www.originlab.com/`) sein, oder die wesentlich günstigere Variante mit ähnlichen Möglichkeiten, das Programm *QtiPlot*© (`https://www.qtiplot.com/`). Letzteres hat den Vorteil, dass es nicht nur für Windows verfügbar ist, sondern auch für andere Betriebssysteme wie Linux. Diese beiden Programme stellen professionelle Alternativen zum etablierten MS-Excel dar, mit deutlich mehr Möglichkeiten für die Datenanalyse. Es gibt aber auch hervorragende, freie Programme für die Datenanalyse. Wir möchten an dieser Stelle nur zwei Beispiele nennen, nämlich die beiden Statistikpakete *R* (`https://www.r-project.org/`) und *Gretl* (`https://gretl.sourceforge.net/`). Mit *Gretl* werden wir im Folgenden ein einfaches Beispiel für die Datenanalyse demonstrieren und empfehlen dieses Programm auch gerne für den einfachen Einstieg in die statistische Datenanalyse. *R* bietet, auch über Zusatzmodule, noch deutlich weitergehende Möglichkeiten.

6.1. Messdaten und ihre Fehler

Für ein einfaches Beispiel gehen wir von einer beliebigen, normalverteilten Messreihe mit den Datenpunkten $x_1, x_2, x_3, ..., x_n$ aus, d.h. wir betrachten n unabhängige Messwerte mit gleicher Genauigkeit. Das setzt voraus, dass alle Datenpunkte mit der gleichen Messmethode, dem gleichen Messinstrument und von demselben Beobachter bestimmt worden sind. Den Mittelwert $\overline{x}$ einer solchen Messreihe berechnet man folgendermaßen:

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

Die Standardabweichung einer Einzelmessung ergibt sich dann aus:

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

Für die Standardabweichung des Mittelwertes folgt daraus:

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

Ein Messergebnis x ist daher folgendermaßen korrekt zu formulieren:

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

Der Zahlenfaktor t hängt vom gewählten Vertrauensniveau γ (z.B. $\gamma = 95\%$) und der Anzahl n der Einzelmessungen ab. Die folgende Tabelle enthält einige Werte für t. Ausführlichere Tabellen sind in der oben genannten Literatur zu finden:

Tabelle 6.1.: Werte für t, in Abhängigkeit von der Anzahl n der Messwerte und dem gewählten Vertrauensniveau 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. Datenanalyse und die Visualisierung von Daten

In diesem Abschnitt möchten wir anhand eines einfachen, künstlich generieren Datensatzes (anscombe.gdt) demonstrieren, wie dieser mit Hilfe der bereits oben genannten Statistiksoftware *Gretl* (`https://gretl.sourceforge.net/`) recht komfortabel und schnell analysiert werden kann. Dieser Datensatz steht als Beispieldatensatz in der aktuellen Version von *Gretl 2025b* (Juli 2025) zur Verfügung und wir möchten alle Leserinnen und Leser gerne dazu animieren, mit diesem Datensatz

unsere Analyse zu reproduzieren sowie auch verfeinerte und komplexere Analysen durchzuführen, um die Möglichkeiten des Programms kennenzulernen. Dabei wird auch deutlich werden, dass die numerische und grafische Datenanalyse eng miteinander verknüpft sind und sich sinnvollerweise ergänzen sollten. Das menschliche Auge bzw. unser Gehirn kann grafische Unstimmigkeiten in der Regel sehr schnell erfassen, wodurch die Datenanalyse stark beschleunigt werden kann. Outlier können so häufig bereits optisch erfasst werden.
Abbildung 6.1 zeigt den Startbildschirm von *Gretl*. In diesem Menü haben wir bereits unseren Beispieldatensatz *anscombe.gdt* gewählt und erhalten so einen ersten Eindruck von der Datenstruktur.

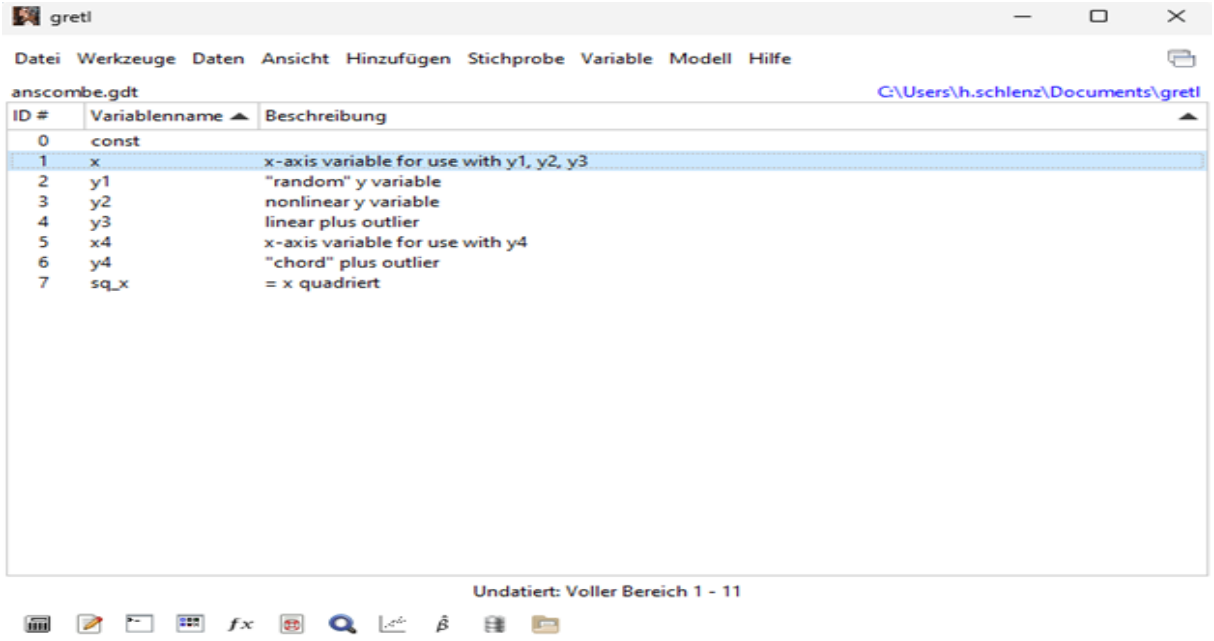
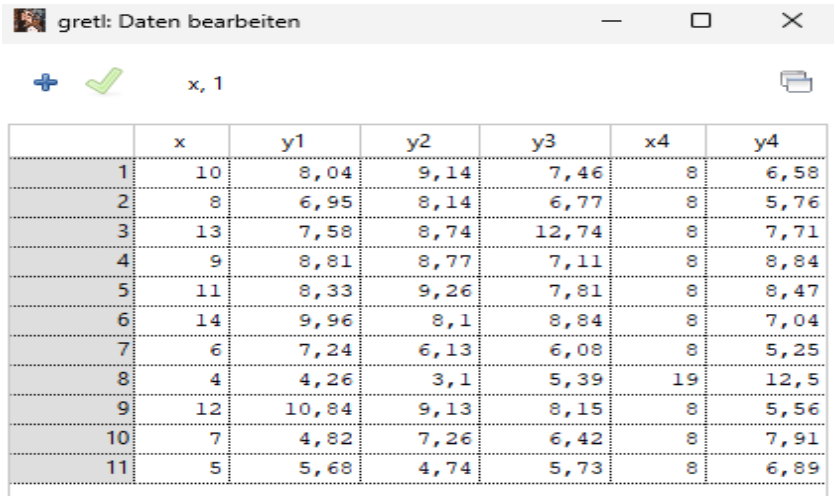


Abbildung 6.1.: Die Oberfläche des Programms *Gretl*.



	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

Abbildung 6.2.: Der Datensatz mit dem Namen anscombe.gdt.

In der folgenden Abbildung 6.2 sehen wir explizit den Datensatz mit allen Werten. Der Variable x können verschiedene y-Werte zugeordnet werden (y_1 bis y_3). Zusätzlich gibt es eine Spalte mit quadrierten x-Werten (x_4) und einem jeweils zugehörigen y_4-Wert. Für unsere Beispielanalyse wählen

wir die Wertepaare (x,y_1). Für die Analyse der Daten wollen wir eine Ausgleichskurve berechnen. Die einfachste Form einer *Ausgleichskurve* ist eine Regressionsgerade. Eine Gerade mit der Formel $y = a\,x + b$ passt sich optimal an Messpunkte $P_i = (x_i, y_i)$ an, mit $i = 1, 2, ..., n$ und $n \geq 3$. Die Steigung a dieser Geraden (der Regressionskoeffizient) und der Achsenabschnitt b können folgendermaßen berechnet werden:

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

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

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

Um die Genauigkeit der Anpassung beurteilen zu können, wird in der Regel der *Korrelationskoeffizient* r berechnet. Die n Messpunkte liegen immer dann nahezu auf einer Geraden, wenn r sich nur wenig von -1 oder +1 unterscheidet. Im Fall von |r| = 1 liegen die Messpunkte exakt auf einer Geraden. Häufig wird bei linearen Regression aber nicht der Korrelationskoeffizient r, sondern sein Quadrat, das sog. *Bestimmtheitsmaß* R^2 verwendet (und dann irrtümlich als Korrelationskoeffizient bezeichnet). R^2 gibt an, wie gut ein lineares Modell die beobachteten Daten anpasst. D.h. in unserem einfachen Beispiel soll die Regressionsgerade $y = a\,x + b$ unserer erstes mathematisches Modell zur Beschreibung der Daten sein, wobei zur Vollständigkeit noch die Steigung a und der Achsenabschnitt b numerisch bestimmt werden müssen.

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

Die folgende Abbildung 6.3 zeigt die lineare Anpassung mit *Gretl*. Durch das Anklicken des achten Symbols von links, in der unteren Symbolleiste von *Gretl* (Abbildung 6.1), gelangt man in das Grafikmenü und wählt dort für die Darstellung einfach die Variablen x und y_1 aus. Die lineare Regression ist voreingestellt und man erhält direkt die gezeigte Grafik. Für die Steigung a erhalten wir den Wert 0,5

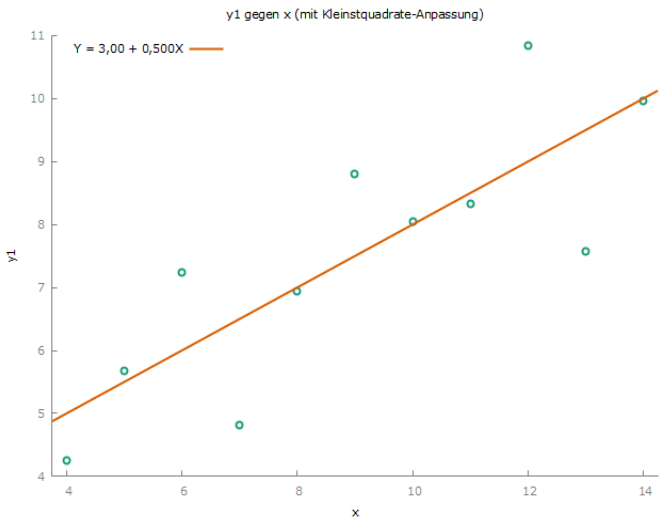


Abbildung 6.3.: Lineare Regression für die Datenpunkte P(x, y_1).

und für den Achsenabschnitt b ist der Wert 3,0. Über das Grafikmenü können wir auch den Wert für R^2 von 0,6665 erhalten. Aus diesem Wert und auch der optischen Betrachtung erschließt sich, dass die Anpassung mit diesem einfachen Modell nicht optimal ist. Daher versuchen wir es als Nächstes mit einer quadratischen Anpassung $y = a + bx + cx^2$. Das Resultat sehen wir in Abbildung 6.4. Jetzt erhalten wir einen Wert $R^2 = 0,6873$, also eine leichte Verbesserung in Richtung 1, mit $a = 0,755$, $b = 1,07$ und $c = -0,0316$. Auch diese Anpassung ist noch nicht optimal. Versuchen Sie es selbst und probieren einfach mal aus, ob eine kubische Anpassung oder eine andere Funktion ein besseres Modell zur Beschreibung des Datensatzes liefert. Abschließend besteht zum Beispiel noch die

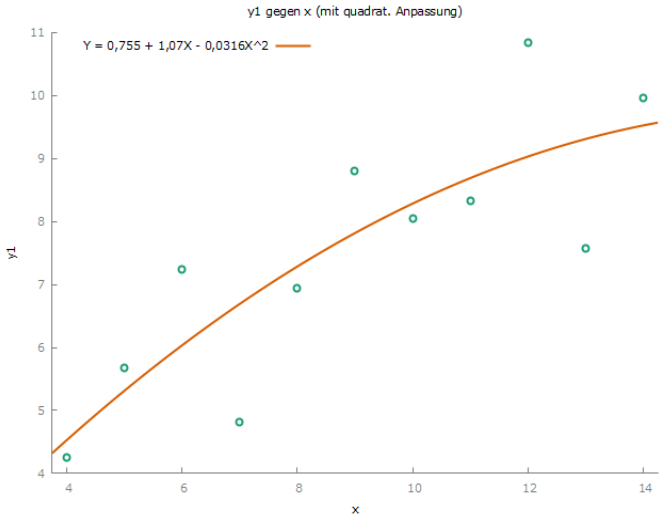


Abbildung 6.4.: Quadratische Regression für die Datenpunkte P(x, y_1).

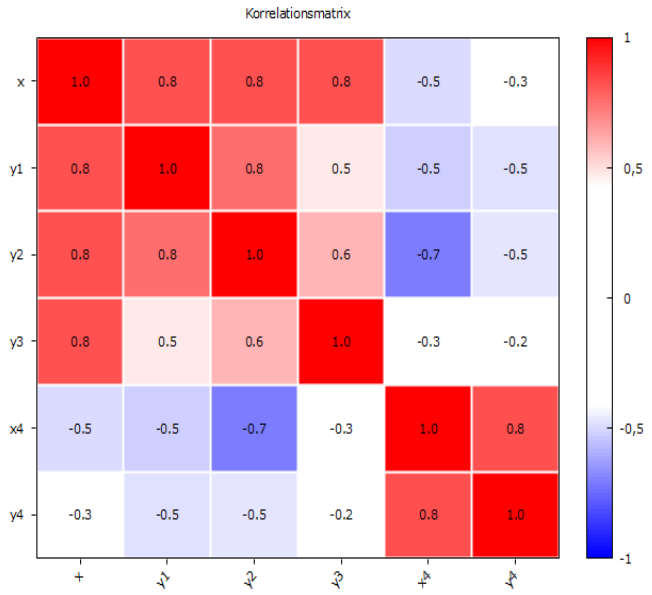


Abbildung 6.5.: Korrelationsmatrix (engl. Heatmap) für den gesamten Datensatz.

Möglichkeit, mit *Gretl* eine Korrelationsmatrix (engl. Heatmap) für den gesamten Datensatz zu generieren, um einen Überblick über alle bestehenden positiven und negativen Korrelationen zu erhalten.

In Abbildung 6.5 sehen wir eine starke, positive Korrelation von x und y_1 mit einem Wert von 0,8, d.h. wenn der Wert von x ansteigt, dann sehen wir auch eine entsprechende positive Zunahme für den Wert von y_1, was zum Beispiel anhand von Abbildung 6.3 leicht nachvollzogen werden kann.

Mit diesem einfachen Beispiel für eine Datenanalyse möchten wir die Grundprinzipien verdeutlichen. Selbstverständlich gestaltet sich die Analyse großer Datensätze ggf. noch wesentlich komplexer und umfangreicher, aber das Prinzip bleibt immer ähnlich. Die notwendigen Modelle zur Beschreibung großer und komplexer Datensätze erfordern Methoden der Wahrscheinlichkeitsrechnung, der deskriptiven und inferenziellen Statistik, der linearen Algebra, der logistischen Regression und Klassifikation oder sogar den Einsatz von KI-Methoden wie neuronale Netze. Die Darstellung dieser Verfahren würde allerdings weit über die Ambitionen diese Handbuches hinausgehen, und wir verweisen an dieser Stelle auf die im Anhang angegebene Literatur. Lediglich im Kapitel über *Maschinelles Lernen (KI)* werden wir noch etwas darauf eingehen.

Für die Visualisierung von Forschungsdaten gibt es eine Vielzahl an kommerziellen und freien Programmen. *Gretl* verwendet intern das freie Programm *Gnuplot* (`http://www.gnuplot.info/`), welches auch als selbstständiges Programm für nahezu jedes Betriebssystem verfügbar ist. Weitere freie Varianten sind *R* (`https://www.r-project.org/`), das recht komfortable *Veusz* (`https://veusz.github.io/`), *Labplot* (`https://labplot.org/`), oder für programmieraffine Nutzerinnen und Nutzer zum Beispiel *Python*-Bibliotheken wie *Matplotlib*. Auch an kommerziellen Systemen herrscht auf dem Markt kein Mangel, und die zwei Vertreter *Origin*® (`https://www.originlab.com/`) und *QtiPlot*© (`https://www.qtiplot.com/`)wurden bereits oben genannt.

Es können aber auch Computer-Algebra-Systeme wie *Mathematica* (`https://www.wolfram.com/mathematica/`), *Matlab* (`https://de.mathworks.com/?s_tid=gn_logo`), *Maple* (`https://www.maplesoft.com/products/maple/index.aspx`) und andere für die Datenanalyse und Visualisierung verwendet werden, je nach Geldbeutel und persönlichen Vorlieben. Wichtig ist jedoch immer, unabhängig von der verwendeten Software, dass Sie sich immer auch ein Bild (im wahrsten Sinne des Wortes) von Ihren Daten machen, um diese bestmöglich beurteilen zu können.

7. Datenaustausch und Datennachverfolgung

Diese Kapitel richtet sich an fortgeschrittene Anwender, die tiefer in das Forschungsdatenmanagement einsteigen möchten. Der Inhalt der vorhergehenden Kapitel wird für das Verständnis vorausgesetzt. Hier möchten wir beschreiben, wie Forschungsdaten zwischen verschiedenen elektronischen Laborbüchern ausgetauscht werden können, auch zwischen unterschiedlichsten ELN's rund um den Globus. Ein weiterer Fokus soll auf der Datennachverfolgung (engl. Provenance Tracking) liegen, d.h. der idealerweise lückenlosen Nachverfolgung einzelner Prozessschritte, zum Beispiel in der Materialforschung oder in der Verfahrenstechnik. Aber auch andere Anwendungsgebiete sind hier denkbar. Die Datennachverfolgung kann innerhalb einer Instanz eines ELN's erfolgen, aber auch zwischen verschiedenen ELN's, die miteinander kommunizieren können.

7.1. Datenaustausch zwischen elektronischen Laborbüchern mit SciMesh

Wir haben an dieser Stelle das noch recht neue System SciMesh (`https://scimesh.org`) aus zwei Gründen für die Darstellung der Vorgänge Datenaustausch und Datennachverfolgung gewählt: 1. Die Autoren dieses Handbuches sind (teilweise) auch die Entwickler von SciMesh; 2. SciMesh erscheint uns als eine ideale Lösung, um die o.g. Aufgaben effektiv lösen zu können. Im Laufe dieses Kapitels werden wir aber aus Gründen der Vollständigkeit auch noch kurz auf alternative Lösungen eingehen.

Einführung in SciMesh

SciMesh ist eine Reihe von Spezifikationen, die die Darstellung von wissenschaftlichen Ergebnissen in Form eines Wissensgraphen definieren. Während sich viele derartige Datenformate auf Simulationsdaten konzentrieren, schließt SciMesh physische Exemplare ausdrücklich als Bürger erster Klasse ein. Auf diese Weise kann die Herkunft sowohl von Daten als auch von Exemplaren dokumentiert werden. Der unmittelbare Zweck von SciMesh ist die Darstellung von Inhalten elektronischer Labornotizbücher (ELNs) für den Austausch wissenschaftlicher Ergebnisse zwischen ELN-Instanzen. Auf diese Weise können Kooperationspartner, die unterschiedliche ELNs (sogar unterschiedliche ELN-Software) verwenden, eine einheitliche Sicht auf ihre gemeinsamen Ergebnisse ohne Medienbrüche erhalten.
SciMesh stellt wissenschaftliche Erkenntnisse als einen Wissensgraphen dar. In seiner derzeitigen Ausführung konzentriert es sich auf probenbasierte Arbeitsabläufe, bei denen Proben (physische Proben oder Datenartefakte) eine Abfolge von Verarbeitungs- und Messschritten durchlaufen. Er ist jedoch nicht darauf beschränkt. Ganz allgemein stellt sie eine wissenschaftliche Erkenntnis dar, indem sie eine Beziehung zwischen Ursache und Wirkung erklärt. Mit anderen Worten: Wenn bestimmte

Voraussetzungen gegeben sind, dann werden bestimmte Beobachtungen gemacht. Zur Verdeutlichng soll Abbildung 7.1 dienen. Ausgehend von einem Ausgangszustand *Null* (*nil*) verändern die Prozesse

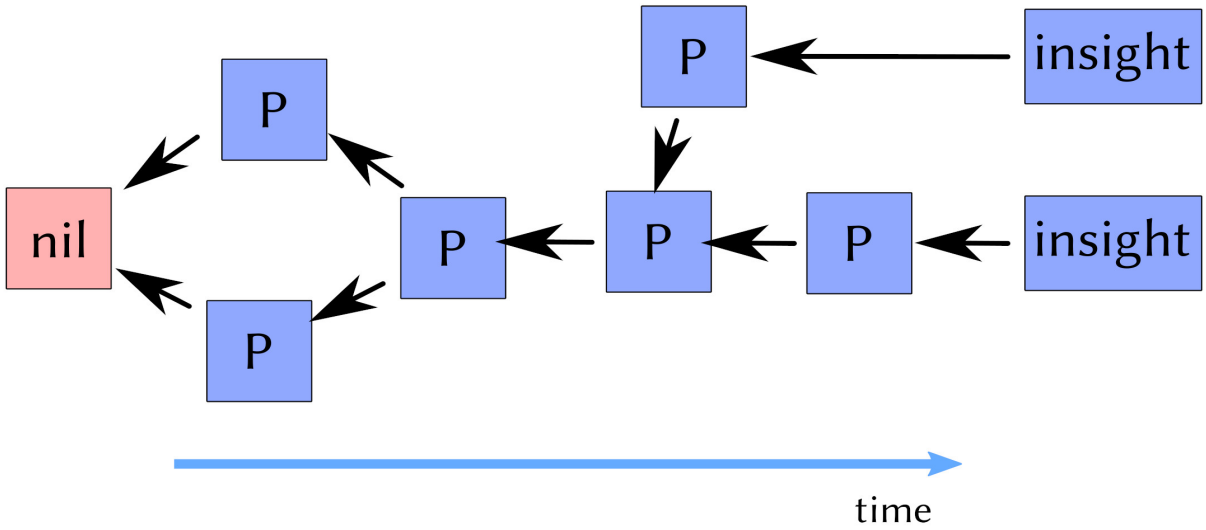


Abbildung 7.1.: Beispielgraph eines Prozesses in SciMesh.

diesen Zustand. Von links nach rechts ist ein monotoner zeitlicher Anstieg zu verzeichnen. Folglich ist diese Zeitachse auch in einer Kette von Prozessen enthalten. Ein Prozess kann lauten: *Siliziumsubstrat aus dem Regal nehmen*, *Probe erhitzen*, *zwei Substanzen zusammenmischen* oder *Sonnenfinsternis abwarten*. So allgemein kann das gehalten werden. Jeder Prozess hat eine oder mehrere Ursachen. Wenn es eine einzige ist, kann es der Anfangszustand sein. Der Anfangszustand ist völlig leer. Er enthält keinerlei Informationen, weshalb die Kette der Prozesse die dazwischen liegenden Zustände so vollständig wie nötig definieren muss, um für wissenschaftliche Schlussfolgerungen nützlich zu sein. Umgekehrt kann ein Prozess die Ursache für mehrere andere sein. Auf diese Weise können Ketten abgezweigt werden, möglicherweise von verschiedenen Wissenschaftlern Jahre nach der Arbeit an dem Hauptstamm. Abbildung 7.2 zeigt, wie das Wesen der wissenschaftlichen Arbeit, das Verhältnis von Wirkung und Ursache, in der Grafik dargestellt wird: Die Prozesse bis zu einem bestimmten Punkt sind die Ursache, und die Beobachtungen an diesem Punkt sind die Wirkung. Daher ist es gültig, diesen Punkt als „Erkenntnis“ zu bezeichnen, die einen eigenen Knoten im Graphen darstellt. Zwei Dinge sind hier wichtig. Erstens handelt es sich bei dem Prozess, an den die Beobachtungsdaten geknüpft sind, um eine Messung, und häufig ändert eine Messung den Zustand nicht. Dennoch handelt es sich in allen Belangen um einen Prozess im eigentlichen Sinne. Wir denken, dass es nicht notwendig ist, zwischen Messungen und Prozessen zu unterscheiden, die tatsächlich etwas verändern. Außerdem ändert eine Messung den Zustand, auch wenn die Änderung vielleicht nicht signifikant ist. Und zweitens kann ein bestimmter Graph von Prozessen viele Erkenntnisse enthalten, die auf ihn hinweisen. Insbesondere können die Prozesse, die auf eine Messung folgen, zu einer zweiten Messung mit neuen Ergebnissen führen, also zu einer neuen Erkenntnis. Apropos Erkenntnisse: Abbildung 7.3 zeigt die Beziehungen zwischen den Dingen, die auf einen bestimmten Zustand (auch Prozess genannt) in einem Diagramm hinweisen können. Alle diese Dinge sind Erkenntnisse, aber wenn es sich um eine Kette konkreter Prozesse zu bestimmten Zeitpunkten handelt, sollte dies als Experiment bezeichnet werden. Wenn all dies mit derselben Stichprobe geschah, kann das Experiment mit dieser Stichprobe identifiziert werden. Handelt es sich bei den Vorgängen nicht um konkrete Vorgänge, sondern um allgemeine Entwürfe von Vorgängen, ist das Experiment eigentlich ein Rezept für Experimente. Oder es

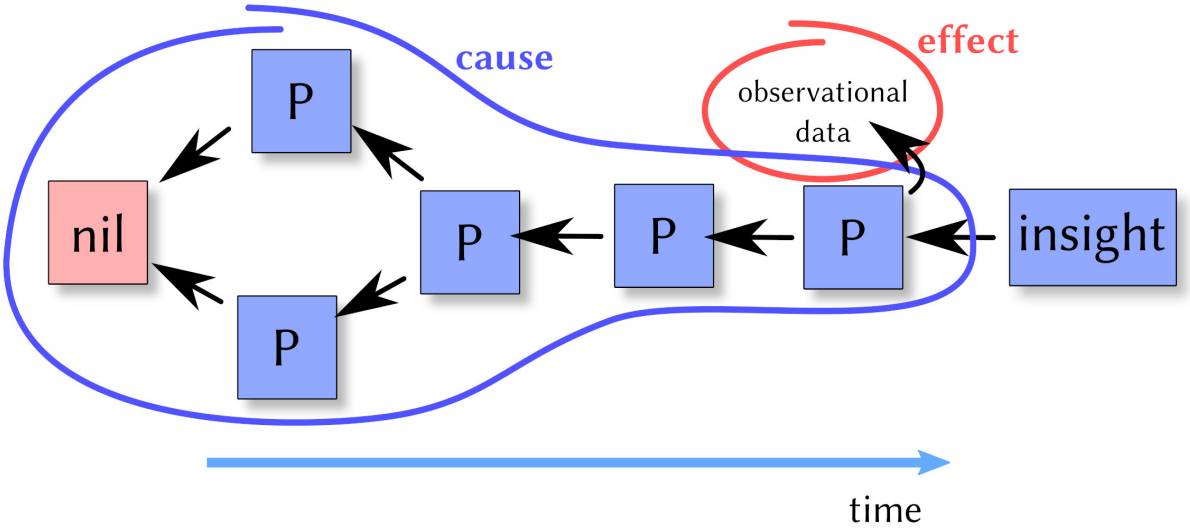


Abbildung 7.2.: Beispielgraph eines Prozesses, der Ursache und Wirkung zeigt.

handelt sich um eine wissenschaftliche Hypothese, die Ursache und Wirkung zusammenbringt. Zur Veranschaulichung von Experiment und Hypothese dienen die beiden folgenden Beispiele:

1. Am 21. Oktober 2019 warf John einen Ball vom Eiffelturm, der sich nach unten bewegte.
2. Ein Körper wird in einem Gravitationsfeld freigesetzt und bewegt sich entsprechend dem Kraftfeld.

Das derzeit spezifizierte RDF-Datenmodell von SciMesh ist enger gefasst als das oben skizzierte sehr allgemeine Konzept. Der Grund dafür ist sehr einfach: Unsere Arbeit befindet sich in einem frühen Stadium und wir müssen uns auf bestimmte Forschungsbereiche konzentrieren, um mit unseren Ressourcen optimal haushalten zu können. Da wir in der *Task Area Caden* von NFDI4ING arbeiten, ist das Forschungsgebiet unserer Wahl der stichprobenbasierte Arbeitsablauf. Wir hoffen, dass wir in naher Zukunft auch in der Lage sein werden, Richtlinien für die Erweiterung von SciMesh auf andere Bereiche zu geben.

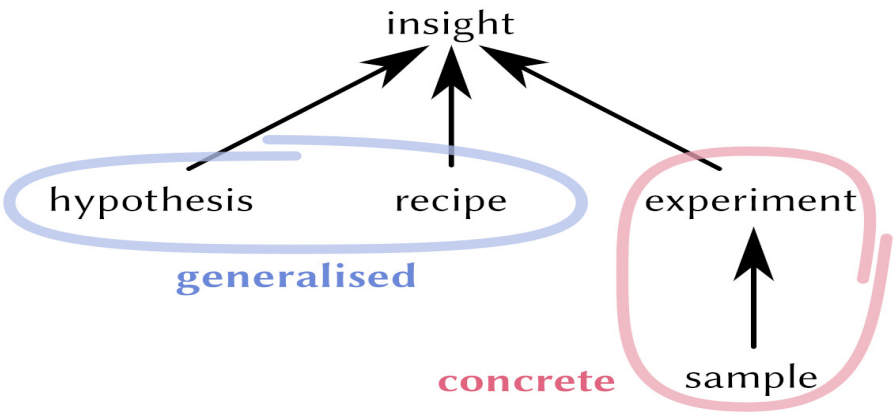


Abbildung 7.3.: Ist-ein Beziehungen von einsichtsähnlichen Entitätsklassen.

Das Datenmodell

Das RDF-Datenmodell basiert auf zwei Konzepten: Muster und Prozess. Beide sind im weitesten Sinne zu verstehen. Eine Probe kann sowohl den Zustand eines physischen Exemplars als auch einen Datensatz darstellen. Ein Prozess kann einen neuen Probenzustand erzeugen (d. h. die Probe verändern, z. B. einen Ätzprozess) oder neue Daten erzeugen (z. B. eine Messung an einer Probe) oder beides. Abbildung 7.4 gibt einen Überblick über die Anatomie eines Wissensgraphen in SciMesh. Es ist ein

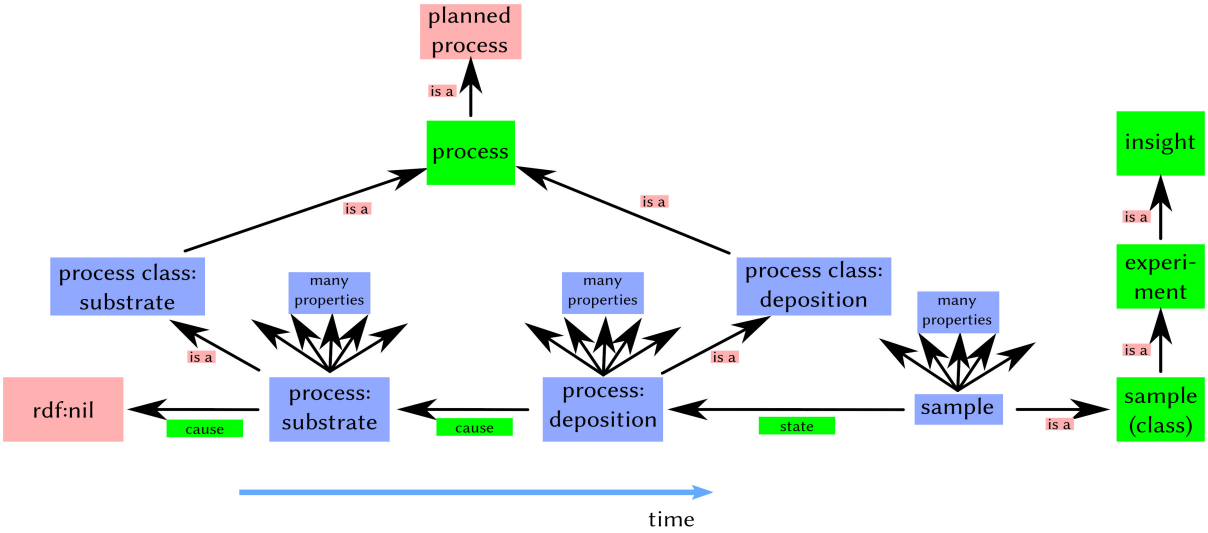


Abbildung 7.4.: Vereinfachte Beispieltopologie eines SciMesh-Wissensgraphen. Farben bezeichnen Namensräume: Blau ist der ELN-Namensraum, grün ist der SciMesh-Namensraum, und rot sind externe Namensräume (RDF, OWL, OBO usw.)..

sehr einfacher Graph, der jedoch die meisten grundlegenden Konzepte enthält. Im Kern besteht er aus einer Abfolge von Prozessen (unten), die an der Probe arbeiten. Der erste Prozess, *Substrat*, erstellt die Probe (die Probe beginnt ihr Leben als blankes Substrat). Seine RDF-Eigenschaften bestimmen die grundlegenden physikalischen Eigenschaften der Probe (z. B. Material und Größe). Anschließend werden im Abscheideverfahren weitere Materialschichten auf das Substrat aufgebracht. Gemeinsame Eigenschaften der meisten Prozesse sind Name, Methode, Zeitstempel, Bediener und Kommentare. Beachten Sie bitte, dass drei verschiedene Namensraumdomänen betroffen sind: 1. Die Domäne der ELN-Instanz: die Prozesse, die Prozesstypen, die Probe und ihre Zwischeninstanzen. Dies ist in blau dargestellt. 2. Die Domäne der ELN-Software: der Probentyp. Sie ist grün dargestellt. 3. Externe Domänen wie BFO/OBO und RDF. Dieser Bereich ist rot dargestellt. Eine ausführlichere Beschreibung des Datenmodells findet sich auf `https://scimesh.org`.

Ein Beispiel: JuliaBase als Prototyp

Das ELN JuliaBase haben wir bereits ausführlicher in Kapitel 5.3 kennengelernt. Im Folgenden soll die Anwendung von SciMesh an einem Beispiel aus JuliaBase demonstriert werden. JuliaBase ist ein Python/Django-Framework zur Erstellung von ELNs oder gleichen Datenbanken mit einem hohen Maß an Anpassungsfähigkeit. Da es einen prozess- bzw. probenbasierten Arbeitsablauf in einer hochstrukturierten Weise realisiert, ist es ein guter Kandidat für das Prototyping von SciMesh. Abbildung 7.5 zeigt ein einfaches Musterdatenblatt. In chronologischer Reihenfolge können Sie sehen, was mit der Probe gemacht wurde. In diesem Fall nur eines: Die *5-Kammer-Beschichtung* ist hier

das einzige Experiment. Es besteht aus drei Siliziumschichten, die auf dem Substrat abgeschieden wurden, jede mit ihrer eigenen Konfiguration (Temperatur, Gasflussraten). Obwohl dieses Beispiel-

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**	SiH_4: **4 sccm**
Layer type: **p**	H_2: **1 sccm**
Chamber: **p**	Silane conc.: **70.59 %**
Temperature: **158 / 165°C**	
Layer number: **002**	SiH_4: **3 sccm**
Layer type: **i**	H_2: **0 sccm**
Chamber: **i2**	Silane conc.: **100 %**
Temperature: **113 / 172°C**	
Layer number: **003**	SiH_4: **9 sccm**
Layer type: **n**	H_2: **11 sccm**
Chamber: **n**	Silane conc.: **32.93 %**
Temperature: **133 / 158°C**	

Abbildung 7.5.: Datenblatt der Probe *14S-005*, wie es der Browser in einer JuliaBase-Instanz zeigt..

datenblatt so einfach ist, ist seine RDF-Darstellung ziemlich komplex, siehe Abbildung 7.6. Diese RDF-Darstellung ist im Turtle-Format, welches für den Menschen lesbar ist (zumindest mit etwas Erfahrung). Nach den Namensraum-Präfixen (die Zeilen, die mit @prefix beginnen), die lediglich dazu dienen, gängige Präfixe mit sehr kurzen Namen abzukürzen, sehen Sie das Beispiel mit seinen Eigenschaften, die im jb-s-Namensraum leben. Die Eigenschaft cause verweist auf den letzten Prozess, der mit diesem Beispiel durchgeführt wurde. Dieser Prozess ist der einzige zur gleichen Zeit (sm:cause: ()). Es ist der Ablagerungsprozess. Er endet mit der Leerzeile. Die instanzspezifischen Entitäten (also die Dinge, die speziell für das jeweilige Institut sind, wie z.B. die Versuchsmethoden) befinden sich übrigens im Namensraum ns1. Was folgt, ist die erste Schicht mit ihren Daten. Sie ist über die Eigenschaft jb:isSubprocess mit ihrer Hinterlegung verknüpft. Die Daten der zweiten Schicht und der kompletten dritten Schicht werden hier der Übersichtlichkeit halber ausgelassen. Die prototypische Implementierung wird mit der SciMesh-Website synchronisiert. Sie wird gegen die JuliaBase-Software in ihrem Graphen-Zweig erstellt. Es gibt auch eine kurze Anleitung, wie man RDF-Daten aus einer JuliaBase-Testinstanz erhält.

7.2. Datennachverfolgung

Datenprozesse wandeln Eingangsdaten in Ausgangsdaten um. Sie können Simulationen, aber auch Datenkonvertierung, Auswertung, Aggregation und Visualisierung beinhalten. Sie sollten atomar sein, d.h. nicht aus Unterprozessen bestehen, aber das ist keine Voraussetzung. In SciMesh sind Datenprozesse von der Klasse *Process*, genau wie experimentelle Prozesse. Sie können mit *cause*-Relationen

```
@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> ;
    …
```

Abbildung 7.6.: Turtle-Darstellung der Probe *14S-005*.

verkettet werden, d.h. ein Datenprozess hat als potentiellen Input alle Daten, die von seinen Vorgängern produziert wurden.

Massendaten: Mit *Massendaten* meinen wir einen undurchsichtigen Oktettstrom von Daten unter einer bestimmten URL. Um in einem SciMesh-Graphen referenziert werden zu können, muss die Antwort des Webservers einen korrekten Inhaltstyp enthalten. Außerdem muss die URL die Prüfsumme der Daten enthalten. Wenn das Protokollschema selbst dies nicht vorsieht (z. B. bei IPFS-URLs), muss das URL-Fragment (der Teil hinter dem #) einen Hash unter Verwendung von Multiformats enthalten. Das Format ist im Einzelnen:

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

Mit anderen Worten, der binäre <multihash> wird durch die Funktion *base()* (z. B. base32) kodiert, und das Zeichen <base>, das diese Funktion bezeichnet (*b* im Fall von base32), wird vorangestellt. <version> ist immer das Byte 0x01.

Dateneingabe - Input: Um zu sehen, welche Daten genau verwendet werden, muss man tiefer in den

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

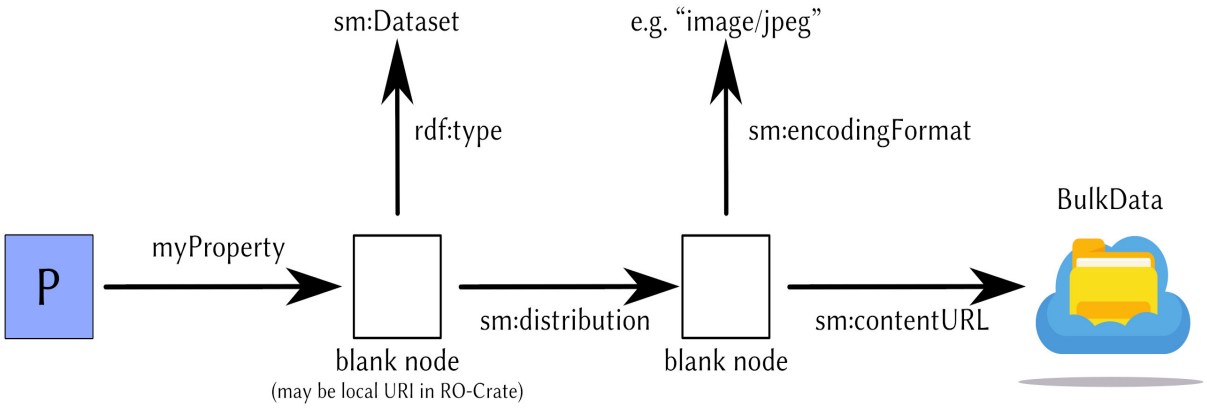


Abbildung 7.7.: Darstellung von Massenausgabedaten in SciMesh. Hier ist *sm* der Namespace *http://schema.org/*.

Prozess eindringen (z.B. durch Einsichtnahme in die Eingabedaten im Verarbeitungsprogramm). In SciMesh sind die URLs zu Bulk-Input-Daten nicht explizit. (Natürlich können Sie sie mit Ihrem eigenen Vokabular explizit machen.) Analog zu physikalischen Proben ist die Eingabe der gesamte Graph der Prozesse (und insbesondere deren Datenausgaben), die zu diesem Prozess geführt haben. Technisch gesehen kann das Programm, das die Datenverarbeitung durchführt, zwar alle Eingabedaten herunterladen, aber ein gültiger SciMesh-Graph stellt sicher, dass alle diese Daten von einem vorangegangenen Prozess ausgegeben wurden. Ein Verstoß gegen diese Vorschrift bedeutet, dass nicht alle Parameter, die die Probe beeinflussen, in einen physikalischen Prozess einbezogen werden. In einigen Fällen kann das bedeuten, dass man einen vorhergehenden Prozess erstellen muss, nur um ihn mit Massenausgabe-URLs zu verbinden. Tun Sie dies einfach, es ist in Ordnung.

Datenausgabe - Output: Alle ausgegebenen Daten werden durch URIs dargestellt, die in abrufbare URLs mit diesen Daten aufgelöst werden, die mit dem Prozess unter Verwendung eines benutzerdefinierten Vokabulars verknüpft sind (wie es bei Messdaten für experimentelle Prozesse der Fall ist). Der Prozess muss das Subjekt solcher Tripel sein (siehe Abbildung 7.7).

7.3. Implementierung von SciMesh

Im Folgenden werden bewährte Praktiken und normative Anforderungen für die Implementierung von SciMesh in einer Einrichtung, insbesondere in einem ELN oder einer Sample-Datenbank, beschrieben.

Der Workflow: Abbildung 7.8 zeigt das Bouncing von Proben und Daten zwischen zwei Instituten, die zusammenarbeiten. Es geht um die experimentellen Aktivitäten mit der Probe Nr. 1, die in Institut A erstellt, aber auch in Institut B untersucht wird. Dabei ist es notwendig, alle Daten in beiden Instituten einsehen zu können. Die textlastige Abbildung enthält die meisten Details. Wir werden daher im Folgenden nur einige Anmerkungen machen. Beide Institute haben ihr eigenes ELN. Bei diesen beiden ELNs handelt es sich nicht nur um unterschiedliche Instanzen auf unterschiedlichen Rechnern, sondern möglicherweise auch um unterschiedliche Software. Es ist wichtig zu sehen, dass Institut A die *eigentliche* Heimat der Probe ist, da die URI der Probe eigentlich eine URL in der Domäne von Institut A ist. Es gibt jedoch eine URL (nicht URI) für die Probe Nr. 1 bei Institut B. Jeder, der alle Daten der Probe Nr. 1 sammeln möchte, muss beide URLs abfragen. Es ist möglich, dass Instanzen Daten liefern, die auch bei anderen Instanzen gefunden wurden, aber das ist nicht garantiert. Institut A führt eine Liste mit allen URLs, die ebenfalls Daten der Probe Nr. 1 haben, und exportiert sie als Teil des SciMesh-Graphen der Probe Nr. 1. Zwei Aspekte werden in dieser Grafik überhaupt nicht behandelt: Caching und Berechtigungen. Während es sich bei ersterem um eine optionale, aber wichtige Optimierung handelt, ist der zweite Aspekt wesentlich.

7.4. Den Graphen erhalten

Eine große Herausforderung in SciMesh ist die Tatsache, dass die Daten einer Stichprobe oder einer Erkenntnis im Allgemeinen über viele Instanzen verstreut sind, möglicherweise in verschiedenen Institutionen und Ländern. Die zwei wichtigsten Dinge, die hier zu beachten sind, sind:

1. Alle URIs von Proben und Prozessen sind konstant. Sie ändern sich nie.

2. Alle diese URIs sind gleichzeitig auch URLs.

Wenn also ein ELN alle Daten für eine bestimmte Probe anzeigen möchte, führt es zunächst einen HTTP-GET mit der Proben-URI durch. Dadurch werden die Entität der Probe und möglicherweise einige Prozessentitäten ermittelt. Das ELN durchläuft dann den Prozessgraphen in der Zeit zurück. Immer wenn er auf einen Prozess mit fehlender Ursache stößt, führt er einen HTTP GET gegen dessen Prozess-URI durch. Daraus ergibt sich ein neuer Graph, der mit dem bestehenden zusammengeführt wird. Dann wird die Traversierung fortgesetzt. Irgendwann gibt es keine Ursachen mehr, nach denen gesucht werden kann (alle verbleibenden Ursachenfelder enthalten rdf:nil), oder ihre URLs können nicht abgerufen werden (weil die Server nicht antworten oder wir nicht die erforderlichen Berechtigungen haben). Anschließend wird das Diagramm dem Benutzer angezeigt.

Anforderungen an ELN's: Die teilnehmenden Datenbanken oder ELNs müssen Folgendes implementieren: 1. Jeder Prozess und seine Prozesshistorie (d.h. der Graph zurück in der Zeit) muss die Antwort auf einen HTTP GET zu dieser Prozess-URI sein. Externe Prozesse (d. h. mit URIs unter der Kontrolle anderer Systeme) müssen nicht einbezogen werden. 2. Ein HTTP GET auf den Muster-URI muss die Musterentität, die Prozesse, auf die sie in den *state*-Eigenschaften verweist, und den gesamten Prozessgraphen zurückgeben. Auch hier brauchen externe Prozesse nicht einbezogen zu werden. 3. Ein HTTP POST an den Beispiel-URI mit einer JSON-Nutzlast der Form:

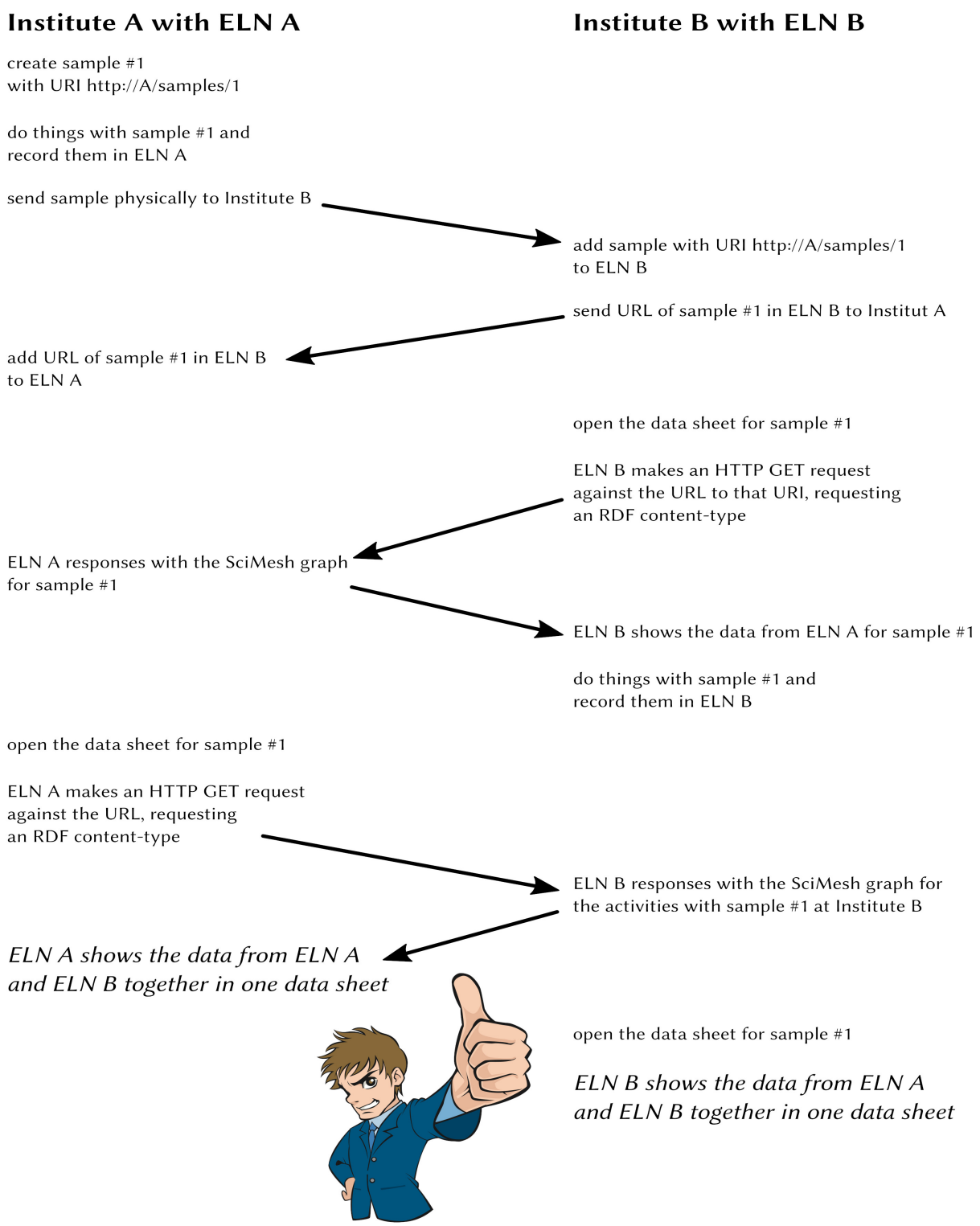


Abbildung 7.8.: Möglicher Arbeitsablauf für zwei Institute, die mit denselben Proben zusammenarbeiten.

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

fügt dem Beispiel die Prozess-URIs hinzu, d. h. es werden *Status*-Eigenschaften mit diesen URIs als Objekte hinzugefügt. Beachten Sie dabei, dass alle diese Anfragen - einschließlich der POST - mit HTTP 30x-Codes beantwortet werden können und mit der neuen URL wiederholt werden müssen. Weitergehende Ausführungen bezüglich der Visualisierung von Prozessen und Graphen, über gute URI's sowie über Authentifizierungen entnehmen Sie bitte der Website von SciMesh.

7.5. MetaData4Ing

Eine Alternative zu dem o.g. System SciMesh kann unter gewissen Umständen MetaData4Ing sein, zumindest was die Datennachverfolgung betrifft. MetaData4Ing ist eine Ontologie zur Beschreibung der Erzeugung von Forschungsdaten im Rahmen einer wissenschaftlichen Tätigkeit. Die Zielgruppe von MetaData4Ing (m4i) sind weniger Anwender im Forschungsdatenmanagement, sondern vielmehr IT-Experten wie zum Beispiel Anwendungsentwickler und Software-Ingenieure. Daher möchte wir an dieser Stelle, aus Gründen der Vollständigkeit, auch nur einen kurzen Überblick über das Projekt geben. Weitergehende Informationen finden sich unter `https://nfdi4ing.pages.rwth-aachen.de/metadata4ing/metadata4ing/`.

Die Ontologie m4i bietet einen Rahmen für die semantische Beschreibung von Forschungsdaten und des gesamten Datenerzeugungsprozesses, der den Untersuchungsgegenstand, alle Proben- und Datenbearbeitungsmethoden und -werkzeuge, die Datenbestände selbst sowie die Rollen von Personen und Institutionen umfasst. Der Aufbau und die Anwendung der Ontologie beruhen auf den Prinzipien der Modularität und der Vererbung. m4i ist eine Sammlung von Begriffen (Klassen und Eigenschaften), die verwendet werden können, um einen Forschungsdatensatz mit semantischen und maschinenlesbaren Metadaten anzureichern, um ihn zu annotieren oder in eine Datenbank oder einen größeren Wissensgraphen zu integrieren. Die Metadaten können in Formaten wie JSON-LD, YAML-LD oder Turtle serialisiert werden.

Warum sollte man MetaData4Ing zur Beschreibung von Forschungsdaten verwenden?

Metadaten enthalten strukturierte Informationen für eine kontextbezogene Beschreibung von Daten und sind sozusagen Daten über Daten. Metadaten werden benötigt, um Daten zu finden, zu verwalten und zu nutzen, nicht nur bei der Veröffentlichung von Daten, sondern auch im Forschungsalltag. In diesem Zusammenhang ist es wichtig, dass alle Informationen, die zum Auffinden und Verstehen der Daten erforderlich sind, in einer gemeinsamen und einheitlichen Sprache ausgedrückt werden, die aus eindeutigen, gut dokumentierten Begriffen besteht. Dieser Ansatz ist eine Voraussetzung für FAIRe (Meta-)Daten, insbesondere für deren Interoperabilität. m4i bietet ein allgemeines prozessbasiertes Modell, das eine flexible Beschreibung von Forschungsaktivitäten und deren Ergebnissen ermöglicht, wobei der Schwerpunkt auf der Provenienz sowohl von Daten als auch von materiellen Objekten liegt. m4i bietet eine Auswahl allgemeiner Konzepte wie Verarbeitungsschritte, In- und Output, verwendete Methoden und Werkzeuge, die es ermöglichen, Informationen über Forschungsprozesse und -ergebnisse in einer strukturierten, konsistenten und maschinenverarbeitbaren Weise zu modellieren. Einer der Hauptvorteile der Verwendung von m4i ist, dass die resultierende Beschreibung in hohem Maße interoperabel ist und die Integration von Daten aus sehr unterschiedlichen wissenschaftlichen Disziplinen in einen einzigen Wissensgraphen ermöglicht. Darüber hinaus verwendet m4i in hohem

Maße Konzepte aus bekannten allgemeinen oder Top-Level-Ontologien, z. B. Basic Formal Ontology (BFO), Data Catalog Vocabulary (DCAT) oder PROV Ontology (PROV-O), wodurch sich die in m4i modellierten Informationen nahtlos in größere Zusammenhänge einbetten lassen. Indem Sie Ihre Forschungsdaten mit m4i dokumentieren, erfüllen Sie nicht nur die Anforderungen guter wissenschaftlicher Praxis, sondern können auch konsistente Metadaten bei der Suche, Analyse oder anderweitigen Verwendung Ihrer Daten nutzen und auch bei der gemeinsamen Arbeit profitieren. Sie können RDF-Metadaten als JSON-LD speichern. Dieses Format bietet semantisch angereicherte Informationen, die für Menschen und Maschinen verständlich sind. Die Verfügbarkeit einer maschinenlesbaren Dokumentation Ihrer Daten erleichtert zudem die Veröffentlichung oder Archivierung Ihrer Daten in Daten-Repositories in zitierfähiger Form.

Das allgemeine Prozessmodell

Eines der Hauptziele der Metadata4Ing-Ontologie ist es, Forschern die Möglichkeit zu geben, die Herkunft von Daten und materiellen Objekten zu dokumentieren, die im Rahmen von Forschungsprozessen erstellt oder verändert werden. Metadata4Ing erreicht dies mit Hilfe eines verallgemeinerten Prozessmodells, in dessen Mittelpunkt die Klasse Verarbeitungsschritt steht. Die oben erwähnten Daten und Materialobjekte werden als Output des Verarbeitungsschrittes beschrieben. Andere relevante Informationen, wie z.B. die in einem Forschungsprozess verwendeten Methoden oder Werkzeuge, werden in separaten Klassen beschrieben, die mit dem Verarbeitungsschritt verknüpft werden können. Mit einer Reihe von Verarbeitungsschritten lassen sich komplexe Forschungsprozesse abbilden. Metadata4Ing kann daher als ein System von Bausteinen verstanden werden, die sich auf Verarbeitungsschritte beziehen und in ihrer Gesamtheit eine vollständige Beschreibung der Herkunft eines Datensatzes oder eines materiellen Objekts ermöglichen. Abbildung 7.9 veranschaulicht das allgemeine Prozessmodell.

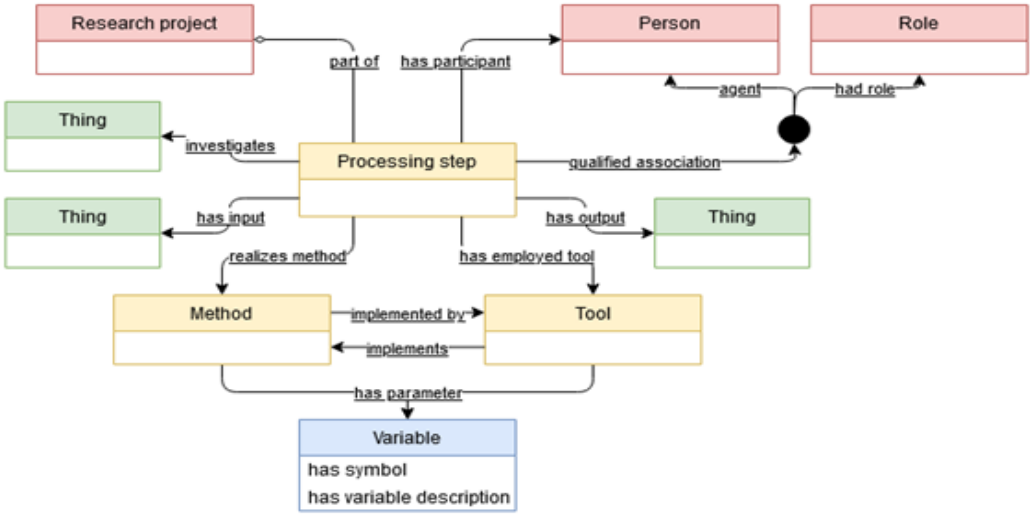


Abbildung 7.9.: Das allgemeine Prozessmodell von MetaData4Ing.

8. Datenpublikation

Das Publizieren von Daten als Supplement zu einem Artikel ist schon lange bekannt. In den letzten Jahren haben sich jedoch neue Publikationswege entwickelt. Daten können über Repositorien als eigenständige Datenpublikation veröffentlicht werden. Forschungsdaten werden außerdem relativ häufig informell im Kollegenkreis ausgetauscht. Sie können auch eigenständig veröffentlicht werden, sofern dem keine Argumente wie Datenschutz oder andere Nutzungskonzepte entgegenstehen (siehe u.a. rechtliche Rahmenbedingungen in Kapitel 3). Die offizielle Veröffentlichung von Daten hat sowohl für die Forscher, die sie zur Verfügung stellen, als auch für die Nutzenden Vorteile. Diese sind unter anderem:

- Es ist kein extra Aufwand nötig, wenn die Forschungsdaten in Zukunft angefordert werden.
- Daten können wie Textpublikationen eindeutig zitiert werden, was die Reputation erhöht.
- Die Veröffentlichung von Daten erhöht die Transparenz und Sichtbarkeit der eigenen Forschung.
- Die Erstellung der Daten wird als eigenständige wissenschaftliche Arbeit anerkannt.
- Die Veröffentlichung verhindert eine doppelte Datenerhebung und damit unnötigen Zeit- und Geldaufwand.
- Vorgaben, z.B. von Forschungsförderern, Verlagen und Richtlinien werden erfüllt.

8.1. Die Veröffentlichung von Datensätzen

Datensätze können prinzipiell in Datenrepositorien oder in Datenjournalen veröffentlicht werden. Aber auch wenn Forschungsdaten in einem Datenjournal veröffentlicht werden sollen, werden sie in der Regel zusätzlich in Datenrepositorium archiviert. Datenrepositorien sind Online-Dienste, in denen digitale Objekte archiviert, dokumentiert und veröffentlicht werden können. Sie erfüllen folgende Funktionen:

- Daten sicher und langfristig archivieren.
- Daten und Metadaten zusammenhalten.
- Daten teilen (entweder öffentlich oder nur mit einem begrenzten Benutzerkreis).
- Eigene Datensuche (Abfrage- und Suchoptionen).
- Einbeziehung von Daten anderer Autoren in Ihre eigene Arbeit.
- Daten lassen sich dauerhaft referenzieren mithilfe von persistenten Identifikatoren wie DOIs.

Allgemein ist ein Datenrepositorium ein Internetdienst, in den Daten (d.h. digitale Objekte) hochgeladen und mit einem dauerhaften Identifikator (PID) sowie mit Metadaten versehen werden. Die Metadaten beschreiben den Inhalt der Daten, wie sie entstanden sind, die verwendete Software und Methoden, rechtliche Aspekte und Nutzungsbedingungen. Die Datensätze können in der Regel auch mit einer zugehörigen Textpublikation verknüpft werden. Über Suchfunktionen können die Daten gefunden, eingesehen und bei entsprechender Berechtigung auch heruntergeladen werden. Die Registry of Research Data Repositories (*re3data*; `https://www.re3data.org/`) bietet einen sehr umfangreichen und gut filterbaren Überblick über Datenrepositorien weltweit. *re3data* ist sehr hilfreich, um sich einen Überblick darüber zu verschaffen, welche Repositorien für Ihre Disziplin oder Arbeitsgruppe in Frage kommen. *re3data* referenziert über 2500 Datenrepositorien weltweit und umfasst sowohl disziplinspezifische als auch generische Repositorien. Es bietet gute Filtermöglichkeiten, z.B. nach Fachgebiet, Datentypen, Nutzungsbedingungen oder fachspezifischen Metadatenstandards, und ermöglicht den Nutzenden eine grobe Einschätzung der Qualität der einzelnen Repositorien.

Beispiele für fachspezifische Repositorien sind die *TIB Hannover* (Medienspezifisches Datenrepositorium; `https://www.tib.eu/de/`), *NoMaD* (Fachspezifisches Datenrepositorium; Novel Materials Discovery; kostenfreies Datenrepositorium für Materialdaten; `https://nomad-lab.eu/nomad-lab/`) und *Zenodo* (Generisches Datenrepositorium; am CERN (`https://home.cern/`) gehostetes Repositorium für Datensätze bis 50 GB Datenvolumen; `https://zenodo.org/`) sowie *Jülich DATA* (Institutionelles Datenrepositorium; Forschungszentrum Jülich; `https://data.fz-juelich.de/`).

Fachspezifische Datenrepositorien sind der zentrale Ort, an dem Forscher nach Daten aus einem bestimmten Fachgebiet suchen können, was bedeutet, dass diese in der Fachgemeinschaft eine gute Sichtbarkeit erhalten. Die Repositorien enthalten in der Regel fachspezifische Metadatenstandards zur Beschreibung der Daten und sind oft mit speziellen Diensten ausgestattet, wie z.B. Tools für die Suche, Analyse und Visualisierung, d.h. wenn es für Ihre Fachrichtung ein derartiges Repositorium gibt, empfiehlt es sich, es zu nutzen. Generische Datenrepositorien sind für alle Forschungsbereiche und alle Arten von Forschungsdaten offen. Ihre Metadatenschemata sind in der Regel universell verwendbar. Institutionelle Datenrepositorien sind in der Regel für verschiedene Fächer offen, ähnlich wie allgemeine Datenrepositorien. Jülich DATA ist das Datenrepositorium des Forschungszentrums Jülich und dient als (bibliographisches) Referenzsystem für die Datenausgabe. Die Metadaten werden mit einer weltweit eindeutigen ID versehen und können durchsucht und heruntergeladen werden. Die Datensätze müssen nicht veröffentlicht werden, können aber der Öffentlichkeit zugänglich gemacht werden, so dass sie leicht zitierbar sind und einen DOI haben.

Für die Recherche nach Repositorien und Forschungsdaten stehen neben **re3data** noch weitere Quellen zur Verfügung. Mit **RIsources** (RI = Research Infrastructure) bietet die DFG ein Informationsportal zur Recherche nach Forschungsinfrastrukturen an. Im Katalog können Sie nach Repositorien und weiteren infrastrukturellen Angeboten suchen (`https://risources.dfg.de/home_de.html`). **Dataset Search**, eine Suchmaschine von Google, die Forscher bei der Suche nach frei zugänglichen Daten unterstützt (`https://datasetsearch.research.google.com/?hl=de`). Sie stellt eine Ergänzung zu Google Scholar dar, dem Suchdienst für akademische Studien und Berichte. **Mendeley Data**, eine Suchmaschine für Forschungsdaten des Elsevier-Verlages (`https://data.mendeley.com/`). DataCite Commons (**DataCite**), eine weltweite Suche über alle Publikationen, für die ein sogenannter DOI vergeben wurde. Bei dem Konsortium DataCite handelt es sich um einen internationalen Zusammenschluss von größtenteils öffentlichen Institutionen, die den Zugang zu Forschungsdaten fördern und diese zur Verfügung stellen wollen (`https://commons.`

datacite.org/). **B2FIND** (EUDAT), eine Suchmaschine für Forschungsdaten, zur Verfügung gestellt von der Europäischen Union im Rahmen des europäischen Netzwerks **EUDAT** (https://eudat.eu/service-catalogue/b2find).

8.2. Beispiel: Publikation eines Datensatzes bei Zenodo

Um einen Forschungsdatensatz bei Zenodo zu publizieren, müssen Sie sich zunächst registrieren und anmelden oder sich mit Ihrem GitHub-Konto anmelden. Anschließend können Sie Ihre Daten hochladen, Metadaten wie Titel, Autoren und Beschreibung hinzufügen und den Datensatz veröffentlichen. Zenodo vergibt automatisch einen DOI (Digital Object Identifier) für Ihren Datensatz, der ihn dauerhaft identifizierbar macht.

1. Registrierung und Anmeldung: Besuchen Sie die Zenodo-Website und registrieren Sie sich mit Ihrer E-Mail-Adresse oder melden Sie sich mit Ihrem GitHub-Konto an. Verbinden Sie nach Möglichkeit Ihr ORCID-Konto mit Zenodo, um Ihre Identität zu bestätigen und die Auffindbarkeit Ihrer Forschung zu verbessern.

2. Neuen Upload erstellen: Klicken Sie auf das Pluszeichen in der Kopfzeile und wählen Sie *Neuer Upload*. Laden Sie Ihre Forschungsdaten hoch, entweder durch Drag & Drop oder durch Auswahl der Dateien in Ihrem Dateiverzeichnis. Zenodo unterstützt das Hochladen von Dateien, nicht von Ordnern. Wenn Sie eine Ordnerstruktur haben, komprimieren Sie diese in eine ZIP-Datei und laden Sie die ZIP-Datei hoch.

3. Metadaten hinzufügen: Geben Sie grundlegende Informationen wie Titel, Autoren, Beschreibung und Ressourcentyp an. Fügen Sie alle relevanten Informationen hinzu, die für die Auffindbarkeit und Nachnutzung Ihrer Daten wichtig sind. Wählen Sie den entsprechenden Ressourcentyp (z.B. Dataset, Publication, Image). Fügen Sie bei Bedarf einen bereits vorhandenen DOI hinzu oder lassen Sie Zenodo einen neuen zuweisen.

4. Veröffentlichung: Überprüfen Sie Ihre Daten und Metadaten sorgfältig. Sie können die Sichtbarkeit des Datensatzes einstellen (öffentlich oder eingeschränkt) und ein Embargo setzen, falls erforderlich. Klicken Sie auf die grüne Schaltfläche *Veröffentlichen*, um Ihren Datensatz zu veröffentlichen. Beachten Sie, dass ein veröffentlichter Datensatz nicht mehr geändert oder gelöscht werden kann.

5. Community: Sie können Ihren Datensatz einer oder mehreren Zenodo-Communities hinzufügen, um ihn einer bestimmten Fachrichtung oder einem Forschungsprojekt zuzuordnen.

6. DOI: Zenodo vergibt für jeden hochgeladenen Datensatz einen DOI (Digital Object Identifier), der eine eindeutige und dauerhafte Kennzeichnung darstellt. Dieser DOI ermöglicht die einfache Zitierung und Referenzierung Ihrer Daten in anderen Publikationen.

7. Testen: Sie können die Testplattform (https://sandbox.zenodo.org) nutzen, um sich mit Zenodo vertraut zu machen, bevor Sie Ihre tatsächlichen Forschungsdaten hochladen.

   Weitere, hilfreiche Anleitungen für die Publikation von Forschungsdaten, können Sie direkt bei Zenodo finden, zum Beispiel die Publikation von F. Schmitt (https://doi.org/10.5281/zenodo.10868941).

8.3. Die Nachnutzung von Forschungsdaten

Forschungsdaten können auf verschiedene Weisen nachgenutzt werden, indem sie öffentlich zugänglich gemacht und in passenden Repositorien archiviert werden. Dabei ist eine gute Dokumentation der Daten und ihrer Herkunft entscheidend für die Nachnutzung durch andere. Durch die Vergabe eindeutiger Identifikatoren (z.B. DOIs) und die Beschreibung der Daten mit Metadaten wird eine einfache Auffindbarkeit und Zitierbarkeit sichergestellt. Dabei sollten sowohl die FAIR-Prinzipien sowie immer auch rechtliche Aspekte dringend berücksichtigt werden (siehe Kapitel 2 und 3). Es ist wichtig, die Urheberrechte und Nutzungsrechte an den Forschungsdaten zu klären. Dabei regeln Lizenzbedingungen wie die Daten nachgenutzt werden dürfen. Es ist ebenfalls ratsam, frühzeitig mit einem Forschungsdatenzentrum oder Repositorium Kontakt aufzunehmen, um ggf. offene Fragen zu klären. Die Nachnutzung von Forschungsdaten kann die wissenschaftliche Arbeit anderer Forscher unterstützen und zu neuen Erkenntnissen führen. Durch die Veröffentlichung und Archivierung von Daten wird die Transparenz und Reproduzierbarkeit von Forschungsergebnissen gefördert (Stichwort: **Gute wissenschaftliche Praxis**). Die Nachnutzung kann auch dazu beitragen, dass Daten nicht ungenutzt bleiben und so ihr volles Potenzial ausgeschöpft werden kann. Bei der Betrachtung von Nachnutzung ist diese von anderen Begriffen zu differenzieren. Unter **Nachnutzung** verstehen wir die Nutzung von Forschungsdaten/Publikationen durch andere als die ursprünglichen Ersteller, oft für neue Zwecke. **Wiederverwendung** bedeutet eine mehrmalige Nutzung eigener Daten im Rahmen eigener Projekte. Als **Recycling** wird die Verwendung alter Daten in modifizierter oder aufbereiteter Form definiert. **Sekundärnutzung** bezeichnet die Nutzung bestehender Daten durch andere Forschungsteams, z. B. im Rahmen von Meta-Studien oder zur Überprüfung und Replikation. Wir wollen uns im Folgenden nur auf die Nachnutzung i.e.S. fokussieren.

Das Ende eines Projekts ist ein wichtiger Zeitpunkt für die Durchführung von Datenmanagementaktivitäten, um sich für die zukünftige Wiederverwendung von Daten zu rüsten. Das liegt daran, dass Sie sich noch an alle wichtigen Details zu Ihren Daten erinnern und gute Entscheidungen über die Vorbereitung der Daten für die Zukunft treffen können.
Daten werden oft in einem Dateityp gespeichert, der nur mit einer bestimmten, teuren Software geöffnet werden kann - dies wird als *proprietärer Dateityp* bezeichnet. Sie können erkennen, dass Ihre Daten in einem proprietären Dateityp gespeichert sind, wenn Sie den Zugriff auf die Daten verlieren, wenn Sie den Zugriff auf die Software verlieren. Wenn Daten in einem proprietären Dateityp vorliegen, ist es immer eine gute Idee, die Daten als Backup in einen gebräuchlicheren, offenen Dateityp zu kopieren; Sie verlieren vielleicht ein wenig Funktionalität, aber es ist besser, ein Backup zu haben, als überhaupt keine Daten zu haben.
Um sich in Zukunft die Zeit zu ersparen, die Sie mit dem Durchsuchen aller Ihrer Forschungsdateien verbringen, legen Sie die wichtigsten Dateien in einem separaten *Archiv*-Ordner ab. Tun Sie dies am Ende des Projekts, solange Sie noch wissen, welche Dateien wichtig sind und wo sie sich befinden. Der Ordner *Archiv* sollte nur eine kleine Teilmenge der wichtigsten Dokumente enthalten, die wahrscheinlich wiederverwendet werden. Sie müssen zwar immer noch alle Ihre Dateien durchsehen, aber in den meisten Fällen werden Sie Zeit sparen, wenn Sie das, was Sie brauchen, einfach im Ordner *Archiv* finden. Sollten Sie ein elektronisches Laborbuch verwenden, so wie in diesem Handbuch mehrfach beschrieben, dann können Sie sich diese Mühe sparen, denn Sie haben bereits alle Daten strukturiert zur Verfügung und können Ihre Daten in die verschiedensten Dateiformate exportieren (siehe Kapitel 5).
Forscher verlassen regelmäßig Forschungseinrichtungen, um an einem anderen Ort eine neue Stelle anzunehmen. Da dies häufig vorkommt, stellt es einen kritischen Übergang dar, bei dem Daten

verloren gehen können. Die Verwendung eines elektronischen Laborbuches, inklusive verknüpfter Datenbanken, verhindert einen solchen Datenverlust (siehe Kapitel 5 und 9).

Möchte man jetzt die Daten einer anderen Wissenschaftlerin oder eines anderen Wissenschaftlers nutzen, so steht man vor einer Reihe von Herausforderungen. Als Erstes müssen die rechtlichen Fragestellungen geklärt werden (Kapitel 3). Anschließend muss die Qualität der Daten und deren Aufbereitung für eine sinnvolle Nachnutzung geklärt werden [**Briney**]. Große Herausforderungen bestehen häufig in einer mangelnden Dokumentation und in fehlerhaften Daten. Eine adäquate Dokumentation ist einer der wichtigsten Aspekte für die Nachnutzung, da man die Details eines Datensatzes kennen und verstehen muss, um diesen nutzen zu können. Wenn die Namen von Variablen z.B. unbekannt sind, so ist eine sinnvolle Nutzung nicht möglich. Wurde parallel ein Artikel publiziert, so kann man mit dessen Hilfe versuchen, die notwendigen Informationen zu erhalten. Die ultima ratio kann auch die direkte Kontaktaufnahme zum Autor bzw. Erzeuger des Datensatzes sein.
Fehler im Datensatz können ebenfalls ein ernsthaftes Problem sein (Kapitel 6). Das können Inkonsistenzen, ungültige Werte, fehlende oder falsche Werte sein. Selbst wenn man in einer ersten Durchsicht der Daten keine Fehler entdecken kann, so sollte man zunächst ein paar einfache Tests mit den Daten durchführen, bevor man sie produktiv einsetzen will. Eine grafische Analyse wie in Kapitel 6.2 gezeigt, kann hier schnelle Ergebnisse liefern. Gleichzeitig entwickelt man ein besseres Verständnis für den Datensatz. Deshalb empfehlen wir als Anfang einer Nachnutzung von Forschungsdaten aus anderen Quellen ein *Herumspielen* mit den Daten, um sich sicher zu werden, ob man die Daten tatsächlich nutzen kann und will.

Zusammenfassend lässt sich sagen, dass eine gute Vorbereitung und eine transparente Handhabung von Forschungsdaten die Nachnutzung erheblich erleichtern und somit den wissenschaftlichen Fortschritt fördern.

8.4. Forschungsdaten für Maschinelles Lernen (KI)

Eine sehr interessante Form der Nachnutzung von Forschungsdaten, deren Bedeutung mit exponentieller Geschwindigkeit zunimmt, ist die Verarbeitung mittels maschinellem Lernen (ML) bzw. allgemein durch den Einsatz künstlicher Intelligenz (KI). Die existierende Literatur zum Thema künstliche Intelligenz und maschinelles Lernen füllt mittlerweile ganze Bibliotheken und die Thematik ist inzwischen durch die verstärkte Nutzung von großen Sprachmodellen (LLM) wie ChatGPT, Grok, Gemini, Mistral, DeepSeek, Claude u.v.m. omnipräsent. Für eine verständliche Einführung empfehlen wir die beiden Bücher von J. Frochte (Maschinelles Lernen (2021); [**Frochte**]) und O. Zeigermann & C.N. Nguyen (Machine Learning kurz & gut (2024); [**Zeigermann**]). Daher werden wir an dieser Stelle nicht auf die verschiedenen Verfahren und ML- bzw. KI-Algorithmen näher eingehen, sondern die wesentlichen Merkmale beschreiben, die Forschungsdaten aufweisen müssen, um (automatisiert) durch eine KI genutzt werden zu können.

Das einfachste Dateiformat stellen Tabellen in Form von **CSV-Dateien** dar, d.h. einfache Tabellen von Zahlenwerten, welche durch Kommata getrennt sind und bei denen die erste Zeile aus Kurzbeschreibungen der jeweiligen Spalten besteht. Ein einfaches Beispiel sehen Sie in Abbildung 6.2. In diesem Fall sprechen wir von **strukturierten Daten**. Ebenfalls geeignet sind auch Daten im **JSON-Format**, da dieses Format im allgemeinen direkt maschinenlesbar ist. Die elektronischen Laborbücher JuliaBase, eLabFTW und Kadi4Mat zum Beispiel (Kapitel 5) können Daten direkt in das JSON-Format exportieren. JSON (JavaScript Object Notation) ist ein leichtgewichtiges Daten-

austauschformat, das für Menschen leicht zu lesen und zu schreiben und für Maschinen leicht zu analysieren und zu erzeugen ist. Es basiert auf einer Teilmenge der JavaScript-Programmiersprache und wird üblicherweise für die Übertragung von Daten zwischen einem Server und Webanwendungen verwendet. Der folgende Code zeigt ein einfaches Beispiel für die Beschreibung einer 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
}
```

Eine weitere Kategorie sind **unstrukturierte Daten**. Unstrukturierte Daten sind Informationen, die nicht in einem vordefinierten, relationalen Format gespeichert sind. Sie liegen oft in Form von Text, Bildern, Videos, Audiodateien oder anderen Datensätzen vor, die keine klare Struktur haben. Im Gegensatz zu strukturierten Daten, die in Datenbanktabellen mit festen Spalten und Zeilen gespeichert werden, können unstrukturierte Daten in verschiedenen Formaten und ohne eine festgelegte Anordnung vorliegen.

Eine weitere Möglichkeit sind sog. **Big Data**. Damit sind Datenbestände gemeint, die bzgl. ihrer Menge, Komplexität, schwachen Strukturierung und/oder Schnelllebigkeit ein Problem für die herkömmliche Datenverarbeitung bzw. Datenanalyse sind [**Frochte**]. Big Data sind gekennzeichnet durch ein großes Datenvolumen, eine große Geschwindigkeit, in der neue Daten generiert werden sowie eine große Bandbreite der Datentypen und -quellen. Große Datenvolumina stellen im Allgemeinen kein Problem dar, eine große Geschwindigkeit und Bandbreite hingegen schon. Die meisten KI-Algorithmen sind darauf angewiesen, dass die Merkmale der Daten konstant bleiben. Wird zum

Beispiel eine neue Datenquelle eingeführt, ein neuer Sensor zur Messung der Temperatur oder Luftfeuchtigkeit, und dieser Sensor liefert Daten, die vorher nicht zur Verfügung standen, dann steht man zunächst vor einer Herausforderung. Der Umgang mit solchen unstrukturierten Daten, die auch für jeden Einzelfall separat betrachtet werden müssen, erfordern Expertenwissen und eine entsprechende Darstellung würde weit über den Anspruch dieses Handbuches hinausgehen. Wir möchten an dieser Stelle daher lediglich das entsprechende Problembewusstsein schaffen. Für Einsteiger empfehlen wir zunächst den vermeintlich deutlich einfacheren Umgang mit strukturierten Daten.

9. Dauerhafte Datenspeicherung

Um ungewollten Datenverlust zu vermeiden, bedarf es guter und etablierter Strategien für die sichere Speicherung, Sicherung (Backup), Übertragung und Entsorgung von Daten. Bei Gemeinschaftsprojekten treten zusätzliche Herausforderungen bzgl. der gemeinsamen Speicherung und des Zugangs zu Daten auf. Daher sollten die grundsätzlichen Strukturen und Strategien bereits am Anfang eines neuen Projektes im Datenmanagementplan geregelt werden (Kapitel 4). Die Beantwortung der folgenden, fundamentalen Fragen hilft bei der Planung [**Corti**]:

- Wieviel Speicherplatz wird für das Projekt benötigt?
- Wer benötigt einen Datenzugang?
- Welche Sicherheitsvorkehrungen müssen gegen Datenverlust getroffen werden?
- Werden personenbezogene Daten verarbeitet bzw. gespeichert?

Für die dauerhafte Speicherung von Forschungsdaten sollten die folgenden Grundsätze beachtet werden:

1. Daten sollten in nicht-komprimierten und nicht-kommerziellen (proprietären) Formaten oder in Dateiformaten mit offenen Standards gespeichert werden, um eine langfristige Lesbarkeit zu gewährleisten.
2. Kopieren oder migrieren Sie Dateien alle zwei bis fünf Jahre auf neue Medien, da sowohl optische als auch magnetische Speichermedien altern.
3. Setzen Sie eine Speicherstrategie ein, auch bei kürzeren Projekten, mit mindestens zwei verschiedenen Speicherformen, wie zum Beispiel Festplatten und optischen Speichermedien (CD, DVD) parallel.
4. Prüfen Sie regelmäßig die Datenintegrität gespeicherter Daten, zum Beispiel mittels Prüfsummen.
5. Organisieren und kennzeichnen Sie gespeicherte Daten eindeutig mit einem vorher festgelegten Namensschema, um Daten später (leichter) wiederfinden zu können.
6. Stellen Sie sicher, dass die Räume, in denen sich die Speichermedien befinden, gültigen Sicherheitsanforderungen entsprechen, d.h. auch gegen Brände und Überflutungen gesichert sind. Beugen Sie auch unautorisiertem Zugang vor.
7. Erzeugen sie für eine langfristige Sicherung digitale Kopien von Papierdaten im PDF/A-Format.

Sollten Sie ein elektronisches Laborbuch für Ihre Arbeit einsetzen, und einen DMP geschrieben bzw. zur Verfügung haben, dann sind die meisten der genannten Aufgaben bereits automatisch für Sie erledigt. Die meisten ELN's verwenden eine eigene Datenbank und wenn Ihr ELN von einem Systemadministrator betreut wird, dann sollten auch regelmäßig Backups aller Daten auf zusätzliche Speichermedien erfolgen. Bei den drei in Kapitel 5 beschriebenen ELN's ist genau das der Fall, zumal diese drei ELN's von den Autoren dieses Handbuchs selbst betrieben werden.

Verschlüsselung

Eine weitere Option für die auch langfristig sichere Handhabung von Forschungsdaten ist die Verschlüsselung von Daten, zum Beispiel für Backups und den Datentransfer. Es können nicht nur einzelne Dateien verschlüsselt werden, sondern auch ganze Speichermedien oder Container für viele Dateien, letzteres zum Beispiel mit der Open-Source Software *VeraCrypt* (`https://veracrypt.io/en/Downloads.html`). Verschlüsselungssoftware verwendet spezielle Algorithmen für die Verschlüsselung von Daten und für die anschließende Entschlüsselung wird ein Schlüssel in Form eines Passwortes oder einer Passphrase benötigt. Je größer die Schlüsselgröße ist (z.B. 256 Bit), desto schwieriger wird ein unerlaubter Zugriff auf die Daten. Allerdings wird es in Zukunft durch quantenkryptografische Verfahren wieder neue Strategien geben müssen, um eine sichere Verschlüsselung zu gewährleisten. Aktuell ist das aber weitestgehend noch Zukunftsmusik.
Es gibt auf dem Markt eine Vielzahl von Verschlüsselungsprogrammen. Ein Standard ist Pretty Good Privacy (PGP), was in der Form von *GnuPG* als Open-Source Version frei verfügbar ist. Für die Anwendung müssen ein öffentlicher und ein privater Schlüssel generiert werden sowie eine Passphrase. Diese Komponenten werden für eine digitale Signatur verwendet, die es dem Empfänger eines Datensatzes erlaubt die Identität des Absenders festzustellen. Der öffentliche Schlüssel des Empfängers muss dafür beim Absender vorhanden sein, damit dieser autorisiert ist, Daten für diesen bestimmten Empfänger zu verschlüsseln. Hier folgt eine Kurzanleitung für die Verschlüsselung von Daten mit PGP:

1. Nur einmal müssen die folgenden Schritte durchgeführt werden:

- Installieren Sie die Verschlüsselungssoftware, z.B *GnuPG*.
- Erzeugen Sie einen Schlüsselpaar mit einem öffentlichen und einem privaten Schlüssel sowie eine Passphrase.
- Laden Sie den öffentlichen Schlüssel einer Institution herunter, mit der Sie Daten austauschen möchten.
- Importieren Sie diesen öffentlichen Schlüssel in Ihre PGP-Software.

2. Die folgenden Schritte sind dann für jeden Verschlüsselungsvorgang durchzuführen, zum Beispiel bei dem Versenden verschlüsselter Daten an einen bestimmten Empfänger:

- Wählen Sie die zu verschlüsselnden Dateien aus.
- Wählen Sie den passenden, öffentlichen Schlüssel des Empfängers aus.
- Signieren Sie die zu verschlüsselnden Dateien mit ihrem privaten Schlüssel und der Passphrase.
- Verschlüsseln Sie anschließend die Dateien mit dem öffentlichen Schlüssel des Empfängers.

- Senden Sie die Daten an den Empfänger über ein sicheres Dateitransferprotokoll (z.B. FTPS, HTTPS, SFPT, u.a.) oder per Post auf einem externen Datenträger.

Zusammenfassung:

Bevor wir noch auf die Unterschiede zwischen Datenbanken, Repositorien und deren Verwendung eingehen, hier eine kurze Zusammenfassung der wichtigsten Punkte, die bei der sicheren und langfristigen Datenspeicherung beachtet werden sollten. Forscher sollten für ihre Projekte eine wohl durchdachte Strategie für das Speichern, die Sicherung, die Weiterleitung und die Entsorgung von Forschungsdaten entwickeln. Auf diese Weise können Forschungsdaten vor Angriffen von außen geschützt sowie Verluste verhindert werden. Eine solche Strategie unterstützt auch den Umgang mit Forschungsdaten entsprechend den FAIR-Prinzipen (siehe Kapitel 2) und sollte Folgendes leisten:

- Identifizieren und wählen Sie den besten Platz für Ihre Datenspeicherung aus.
- Seien Sie sich der Risiken und Vorteile einer Datenspeicherung in einer Cloud bewusst und wägen Sie sorgsam ab.
- Entwickeln Sie eine persönliche Strategie für die Verschlüsselung Ihrer Daten.
- Überlegen Sie, wie oft und wo Backups Ihrer Daten gespeichert werden sollen.
- Überlegen Sie sich auch, wie effektiv und sicher nicht mehr benötigte Daten an den verschiedensten Stellen gelöscht bzw. entsorgt werden können, wenn Sie am Ende eines Projektes alle Daten gesichert in einer Datenbank oder einem Repositorium abgelegt haben.

9.1. Datenbanken

Eine Datenbank ist eine organisierte Sammlung von Daten, die leicht zugänglich sind, verwaltet und aktuell gehalten werden können. Auch wenn zu Beginn eines neuen Projekts der Einsatz einer Datenbank nicht explizit vorgesehen ist (was ein guter DMP verhindern möge!), so werden Ihre Daten im Laufe der Verarbeitung zwangsweise mit einer oder mehreren Datenbanken in Berührung kommen. Sollten Sie ein ELN verwenden, dann wird es auf jeden Fall geschehen. Datenbanken liefern viele Vorteile. Zum einen können Sie mit Hilfe von Skripten Daten zwischen verschiedenen Anwendungen austauschen und weiterhin helfen Ihnen Datenbanksprachen wie SQL bei der Organisation von Daten und beantworten Fragen zu den Daten. Zusätzlich erlauben viele Datenbank-Systeme die Integration und Ausführung von eigenem Programmcode (z.B. Python), um die Leistung und Modularität einer Anwendung zu verbessern.

Es gibt zwei Kategorien von Datenbanken: **Relationale und nicht-relationale Datenbanken** (NoSQL). Relationale Datenbanken besitzen feste Schemata wie Daten gespeichert werden. Dieser Ansatz soll die Datenintegrität, -konsistenz und Genauigkeit der Daten gewährleisten. Der große Nachteil von relationalen Datenbanken ist allerdings, dass sie mit steigendem Datenvolumen nicht einfach skalierbar sind. Im Gegensatz dazu gibt es bei NoSQL-Datenbanken keine Einschränkungen bzgl. der Datenstrukturen und sie ermöglichen so mehr Flexibilität, Anpassbarkeit und Skalierbarkeit. Wer tiefer in Datenbanktechniken einsteigen möchte und sich auch nicht vor der Entwicklung von SQL-Skripten und der Programmiersprache Python scheut, der findet in dem Buch von Y. Vasiliev [**Vasiliev**] einen guten und einfachen Einstieg.

Am gebräuchlichsten sind derzeit **relationale Datenbanken** und sie liefern eine strukturierte Art und

Weise um Daten zu speichern. Um mit einer solchen Datenbank arbeiten zu können, muss zuerst ein Datenschema festgelegt werden, also zum Beispiel für Bücher die Felder Buchtitel, Autor, Verlag, Jahr, usw. Die Daten müssen in der Datenbank in diesem vordefinierten Schema abgelegt werden. Wenn man mit einer relationalen Datenbank arbeitet, beginnt man zunächst mit der Definition eines solchen Schemas. Man definiert eine Sammlung von Tabellen, jede bestehend aus einem Satz von Feldern oder Spalten, und man legt fest, welche Art von Daten die verschiedenen Felder speichern sollen. Außerdem etabliert man Beziehungen zwischen den Tabellen. So kann man Daten in einer relationalen Datenbank speichern, Daten aus ihr erhalten oder Daten aktualisieren. Bekannte relationale Datenbankmanagementsysteme sind **MySQL**, **MariaDB** und **PostgreSQL**.

NoSQL-Datenbanken benötigen dagegen kein vordefiniertes Organisationsschema für die zu speichernden Daten und sie unterstützen auch keine Standard-Datenbankoperationen wie *join*. Als SQL-JOIN (auf Deutsch: Verbund) bezeichnet man eine Operation in relationalen Datenbanken, die Abfragen über mehrere Datenbanktabellen ermöglicht. JOINs führen Daten zusammen, die in unterschiedlichen Tabellen gespeichert sind, und geben diese in gefilterter Form in einer Ergebnistabelle aus. Stattdessen ermöglichen NoSQL-Datenbanken das Speichern großer Datenvolumina in flexibler Form, wodurch die Handhabung großer bis sehr großer Datenmengen vereinfacht wird. Das Speichern sog. Schlüsselwerte ermöglicht das Speichern und die Ausgabe von Daten als Paare von Schlüsselwerten. Das Speichern von verschiedenen Informationen über ein spezifisches Objekt in Form mehrerer Tabellen ist damit nicht länger notwendig. Unter anderem ist die Speicherung und Verarbeitung von Dokumenten im JSON-Format möglich. Diese sind dann prinzipiell auch maschinenlesbar (siehe Kapitel 8.4). NoSQL-Datenbanken sind besonders für **Echtzeitanwendungen** und für **Big Data** Projekte geeignet, und so setzt Google sie bei seinem Email-Dienst Gmail ein oder auch die Business-Plattform LinkedIn. **MongoDB** ist ein für solche Zwecke bekanntes NoSQL-Datenbankmanagementsystem.

Zu Beginn eines Projektes sollten Sie daher überlegen, welche Art von Datenbank zu ihrem Projekt passt und dieses auch im DMP festhalten. Für die praktische Benutzung und Programmierung spezifischer Datenbanken möchten wir sie an die in Massen verfügbare Literatur verweisen oder empfehlen im Zweifelsfall entsprechende Spezialisten in Ihrer wissenschaftlichen Einrichtung zu kontaktieren.

9.2. Repositorien

Über Repositorien haben wir in Kapitel 8 bereits im Zusammenhang mit der Publikation von Datensätzen einiges geschrieben. In diesem Kapitel möchten wir diese Informationen noch um die mögliche Anwendung von Repositorien zur Datenspeicherung ergänzen, insbesondere zur Speicherung von Metadaten (Kapitel 2.1). Ein Repositorium ist ein Speicherort für digitale Objekte, der diese einer Öffentlichkeit oder einem beschränkten Nutzerkreis zur Verfügung stellt. Es handelt sich um eine Art Datenbank oder elektronisches Archiv, das zur Speicherung, Verwaltung und Bereitstellung von Daten, Publikationen oder anderen digitalen Ressourcen dient. Ein Repositorium bietet Funktionen zur Organisation, Klassifizierung und Verwaltung der gespeicherten Daten und ermöglicht den Zugriff und die Nutzung der gespeicherten Ressourcen durch autorisierte Personen oder die Öffentlichkeit. Im wissenschaftlichen Bereich werden Repositorien häufig zur Speicherung von Forschungsdaten, Publikationen (z.B. Dissertationen, Artikel) oder Open Educational Resources (OER) genutzt. Obwohl Repositorien oft mit Archiven verglichen werden, gibt es einen Unterschied. Während Archive primär der langfristigen Aufbewahrung historischer Dokumente dienen, können Repositorien auch für die kurz- oder mittelfristige Speicherung und Nutzung aktueller Daten genutzt werden und dabei unterschiedliche Anforderungen und Nutzungsbedingungen berücksichtigen.

Ein in der wissenschaftlichen Community beliebtes Repositorium haben wir in diesem Handbuch schon mehrfach kennengelernt: Zenodo. Forschungsdaten-Repositorien sind spezialisierte Archive, die Forschungsdaten dauerhaft speichern, organisieren und zugänglich machen. Sie dienen der Sicherung und Nachnutzung von Forschungsdaten und können sowohl fachspezifisch als auch institutionell oder generisch sein. Die Auswahl eines geeigneten Repositoriums sollte sich an den Gepflogenheiten der jeweiligen Fachdisziplin oder den Vorgaben von Förderinstitutionen orientieren. Beispiele für solche Repositorien sind **Zenodo** - Ein generisches Repositorium, das von CERN betrieben wird (siehe oben), **DRYAD** - Ein Repositorium für Forschungsdaten aus den Lebenswissenschaften, **Figshare** - Ein weiteres generisches Repositorium für Forschungsdaten, **DaKS** - Das Datenrepositorium der Universität Kassel, **ResearchData** - das Repositorium der Heinrich-Heine-Universität Düsseldorf, **GFZ Data Services** - Ein Datenrepositorium für die Geowissenschaften und **PANGAEA** - Ein weiteres Datenrepositorium für die Geowissenschaften. Die Website `https://open-access.network/informieren/publizieren/repositorien` listet inzwischen mehr als 5700 Repositorien auf, von denen auch einige mehr als die oben genannten Beispiele Daten speichern.

Wie Daten in einem solchen Repositorium veröffentlicht und auch gespeichert werden können, haben wir bereits in Kapitel 8.2 am Beispiel von Zenodo prinzipiell demonstriert. Die Ablage von Forschungsdaten in einem Repositorium kann auch intern erfolgen, ohne dass die Daten veröffentlicht werden. Viele Repositorien bieten die Möglichkeit, Daten mit einem Embargo zu versehen, um sie erst zu einem späteren Zeitpunkt öffentlich zugänglich zu machen. Es gibt auch Datenjournale, die Forschungsdaten mit einem Peer-Review-Prozess veröffentlichen, ähnlich wie klassische wissenschaftliche Zeitschriften.

9.3. Coscine

Coscine (**Co**llaborative **Sc**ientific **In**tegration **E**nvironment) ist weder eine reine Datenbank, noch ein reines Repositorium, sondern versteht sich vielmehr als eine **Plattform für Forschungs-Datenmanagement**. Die Plattform wird an der RWTH Aachen (`https://www.itc.rwth-aachen.de/cms/it-center/services/forschung/~smhwy/coscine/`) gehostet und bietet Speicherplatz (Zugang zu kostenlosem Speicherplatz auf dem Research Data Storage), Integration (Zugriff auf projektbezogene Datenquellen, zum Beispiel Forschungsdatenspeicher, verlinkte Dateien, archivierte Daten), Kollaboration (Zugriff für alle Projektmitglieder), Metadaten (Automatische Verknüpfung mit Projektdaten), Individualität (Erstellen projektspezifischer Metadaten als Applikationsprofile) und eine Archivierung (Forschungsdaten an Ort und Stelle archivieren).

Für Forscher, die kein ELN nutzen können oder wollen, kann Coscine unter gewissen Umständen eine Alternative sein. Dabei sollte man aber immer im Hinterkopf haben, dass es sich hierbei um eine Cloud-basierte Lösung handelt, mit allen Vor- und Nachteilen, insbesondere im Bereich der Datensicherheit. Im Vergleich zu einem ELN, welches einem vor Ort zur Verfügung steht, bestehen gewisse Einschränkungen, insbesondere bei der Anbindung von Experimenten, die sich in einem Institut in einem IT-mäßig abgesicherten Bereich befinden. Coscine ist Open-Source und wird auf GitLab entwickelt (`https://git.rwth-aachen.de/coscine`).

A. Anhang

Für die Korrektheit und die Sicherheit der in diesem Handbuch aufgeführten Weblinks (URL's) sowie auch für die Inhalte der verlinkten Webseiten übernehmen wir ausdrücklich keinerlei Verantwortung.

A.1. Forschungsdatenorganisationen in Deutschland

- **Allianz der deutschen Wissenschaftsorganisationen** (https://www.allianz-der-wissenschaftsorganisationen.de/)
- **Datenkompetenzzentren** an Hochschulen und Fachhochschulen (https://www.bmbf.de/DE/Forschung/Wissenschaftssystem/Forschungsdaten/DatenkompetenzenInDerWissenschaft/datenkompetenzeninderwissenschaft_node.html)
- Deutsche Forschungsgemeinschaft **DFG** (https://www.dfg.de/de)
- Landesinitiative **fdm.nrw** für Forschungsdatenmanagement NRW (https://www.fdm.nrw/), als ein Beispiel für eine Landesinitiative
- Informationsportal **forschungsdaten.info** (https://forschungsdaten.info/)
- **NFDI** Nationale Forschungsdateninfrastruktur (https://www.nfdi.de)

A.2. Weiterführende Informationen im 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

forschungsdaten.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. Abkürzungsverzeichnis

API Application Programming Interfaces (Programmierschnittstelle)

ASCII ASCII-Format von Textdateien (American Standard Code for Information Interchange)

B2FIND Suchmaschine für Forschungsdaten, zur Verfügung gestellt von der Europäischen Union

BDSG Bundesdatenschutzgesetz

BFO Basic Formal Ontology

BMFTE Bundesministerium für Forschung, Technologie und Raumfahrt

BMWE Bundesministerium für Wirtschaft und Energie

BPersVG Bundespersonalvertretungsgesetz

BVerwG Bundesverwaltungsgericht

CC-BY Creative Commons Lizenzmodell

CCO Creative Commons Lizenzmodell

CERN Europäische Organisation für Kernforschung (Conseil Européen pour la Recherche Nucléaire)

CIF Crystallographic Information File (Standard-Textdateiformat zur Darstellung kristallographischer Informationen)

COD Crystallography Open Database (kostenfrei zugängliche Datenbank, in der Kristallstrukturen aus wissenschaftlichen Veröffentlichungen erfasst werden)

Coscine Collaborative Scientific Integration Environment

CSV Comma-separated values (Dateiformat)

DaKS Datenrepositorium der Universität Kassel

DataCite Ein internationales Konsortium, das sich zum Ziel gemacht hat, einen einfachen Zugang zu wissenschaftlichen Forschungsdaten zu ermöglichen.

DCAT Data Catalog Vocabulary

DFG Deutsche Forschungsgemeinschaft

DMP Datenmanagementplan

DOI Digital Object Identifier (Ein eindeutiger und dauerhaft gültiger Identifikator für Publikationen, Forschungsdaten, Videos und weitere wissenschaftliche Ressourcen im Internet, ähnlich einer ISBN oder ISSN für Bücher bzw. Zeitschriften.)

DRYAD Open Data Publishing Platform

DSGVO Datenschutzgrundverordnung

ELN Electronic Lab Notebook (elektronisches Laborbuch)

EU Europäische Union

EXIF Exchangeable Image File Format

EUDAT European Data Initiative

FAIR Findability, Accessibility, Interoperability, and Reusability

FDM Forschungsdaten-Management

figshare Provider of Open Research Repository Infrastructure

GG Grundgesetz der Bundesrepublik Deutschland

GNU GNU's Not Unix (alternatives, freies Betriebssystem)

GnuPG GNU Privacy Guard

HGF Helmholtz-Gemeinschaft Deutscher Forschungszentren

HMC Helmholtz Metadata Collaboration

HTTP Hypertext Transfer Protocol

IPFS InterPlanetary File System

JSON JavaScript Object Notation

KI Künstliche Intelligenz

LLM Large Language Model

m4i Metadata4Ing

ML Maschinelles Lernen

NFDI Nationale Forschungsdateninfrastruktur

NFDI4Ing Das Konsortium NFDI4Ing für Ingenieure, als Teil von NFDI

NMR Nuclear Magnetic Resonance (Kernspinresonanzspektroskopie)

ODC Open Data Commons - Lizenzmodell

OER Open Educational Resources

PANGAEA Repositorium für Geowissenschaften

PGP Pretty Good Privacy

PID Persistent and unique identifier (Dauerhafte, digitale Kennung)

PROV-O The PROV Ontology of W3C

RDF Ressource Description Framework

RDMO Research Data Management Organiser

re3data Verzeichnis von Forschungsdatenrepositorien

RIsources Portal für Forschungsinfrastrukturen

UrhG Urheberrechtsgesetz

URI Uniform Resource Identifier (Identifikator; besteht aus einer Zeichenfolge, die zur Identifizierung einer abstrakten oder physischen Ressource dient)

URL Uniform Resource Locator (identifiziert und lokalisiert eine Ressource, z.B. eine Webseite)

Literatur

[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 zum Text und Data Mining für Forschungszwecke in Deutschland*; NFDI4ing und TIB - Leibniz-Informationszentrum Technik und Naturwissenschaften, Universitätsbibliothek: Hannover, Germany, 2022.

[Bremecker] Bremecker, D., *Mitbestimmung/Mitwirkung / 2.4.17 Einführung und Anwendung technischer Kontrolleinrichtungen*; 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: Exexter, 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] Deutsche Forschungsgemeinschaft *Handlungsempfehlung zum Umgang mit Forschungsdaten*; DFG: Bonn, Germany, 2023.

[EU] European Commision, *H2O2O Programme AGA - Annotated Model Grant Agreement Version 5.2*, EU: Brüssel, Belgien, 2019.

[Frochte] Frochte, J. *Maschinelles Lernen*; 3. Auflage; Hanser-Verlag, München, 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. *Das Recht des Forschers auf Datenschutz*; Springer-Verlag, Datenschutz und Datensicherheit - DuD, 11, 817–822, 2012.

[Lauber] Lauber-Rönsberg, A. *Rechtliche Aspekte des Forschungsdatenmanagements*; In: *Praxishandbuch Forschungsdatenmanagement*; Putnings, M.; Neuroth, H.; Neumann, J. (Eds.); De Gruyter: Berlin/Boston, 2023.

[Nield] Nield, T. *Mathe-Basics für Data Scientists*; O'Reilly, Heidelberg, 2024.

[Papula] Papula, L. *Mathematik für Ingenieure und Naturwissenschaftler, Band 3*; Springer Vieweg, Wiesbaden, 8. Auflage, 2024.

[Putnings] Putnings, M.; Neuroth, H.; Neumann, J. (Eds.) *Praxishandbuch Forschungsdatenmanagement*; 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 (Hrsg.) *ELN Wegweiser: Elektronische Laborbücher im Kontext von Forschungsdatenmanagement und guter wissenschaftlicher Praxis - ein Wegweiser für die Lebenswissenschaften*; 2. Auflage; Publisso: Köln, Germany, 2020.

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

Index

Admins, 35
Algebra, 61
API, 48
Archiv, 82, 90
Archivierung, 19
ASCII, 7
Authentifizierung, 49

B2FIND, 81
Backup, 19, 82, 89
Beobachtungsdaten, 8
Berechtigung, 53
Best-Practice, 35
Big Data, 84, 90
Bilddaten, 8
Blockchain, 40
BMP, 10
Boolean, 53
Browser, 35

C, 7
CERN, 8, 10, 80, 91
CIF-Format, 8
Cloud, 36, 89
COD, 8
Computer-Simulationen, 5
Coscine, 91
Credentials, 49
CSV, 18, 83
CSV-Dateiformat, 10

DaKS, 91
Data Mining, 13
DataCite, 18, 20, 80
Dataset Search, 80
Dateiformat, 35, 54
Datenanalyse, 61, 66, 84
Datenaustausch, 67
Datenbank, 10, 24, 47, 83, 89
Datenformat, 18
Datenintegrität, 87
Datenjournal, 79, 91
Datenmanagementplan, 17
Datennachverfolgung, 11, 67
Datenprozess, 73
Datenpublikation, 79
Datenqualität, 11, 61
Datenrepositorien, 18
Datenrepositorium, 79
Datensatz, 53
Datenschutz, 79
Datenschutzgrundverordnung, 14
Datenschutzrecht, 14
Datenspeicherung, 87
Datenströme, 17
Datentyp, 84
Datentypen, 18
Datenverlust, 19, 83, 87
Datenvolumen, 18, 84
DFG, 80
Dictionary, 53
Digitale Kennung, 10
Digitale Signatur, 88
DMP, 17, 88, 90
Docker-Container, 36
DOI, 18, 81
Dokumentation, 9
Drittmittelgeber, 5, 17
DRYAD, 91
DSGVO, 14

Eigentum, 19
eLabFTW, 18, 20, 35, 83
Ellipsis-Menü, 40
ELN, 7, 48, 67, 88
Email, 45
Embargo, 91
Embargofrist, 11
Entität, 71
EUDAT, 81
Europäische Union, 81
EXIF, 8
Experiment, 39, 51
Experimentelle Daten, 8

FAIR, 5, 9, 18, 20, 76
FDM, 7
Fehler, 61, 83
Figshare, 91
Fileserver, 19
Float, 53
Formate, 18
Forschungsdaten, 5
Forschungszentrum Jülich, 80
Fotografie, 14

Ganzzahl, 53
GFZ Data Services, 91
GitHub, 81
GitLab, 91
GnuPG, 88
Gnuplot, 66
Google, 80, 90
Google Scholar, 80
Graph, 51, 71, 74
Gretl, 61
Gruppenrollen, 54

Heatmap, 65
HTTP-API, 57

Identifikator, 80
Infrastruktur, 19

JavaScript, 84
JOIN, 90
JPG, 7, 10

JSON, 36, 83, 90
JSON-Editor, 44
JSON-LD, 76
JuliaBase, 23, 70, 83
Jülich DATA, 80
JülichDATA, 18, 20

Kadi4Mat, 48, 83
Kernspinresonanzspektroskopie, 8
KI, 61, 66, 83
Klasse, 77
Klassen, 76
Klima, 5
Kodex, 19
Korrelationskoeffizient, 64
Korrelationsmatrix, 65
Kuratieren, 20
Künstliche Intelligenz, 5, 83

Laborbuch, 83
Labplot, 66
LDAP, 49
Lebenszyklus, 17
Leistungsschutzgesetze, 13
Lineare Regression, 64
Linux, 61
Linux-Server, 36
Liste, 53
Lizenz, 11
LLM, 83
Login, 49

m4i, 76
Mac, 36
Maple, 66
MariaDB, 90
Maschinelles Lernen, 5, 66, 83
Massendaten, 73
Mathematica, 66
Matlab, 66
Matplotlib, 66
Mendeley Data, 80
Messreihe, 62
Messwert, 8
MetaData4Ing, 76
Metadaten, 8, 11, 18, 51, 53, 56, 76
Metadatenschema, 80
Metadatenstandard, 10
ML, 83
Modell, 64
Molekül-Editor, 47
MongoDB, 90
MP3, 7, 10
MP4, 7, 10
MySQL, 90

Nachnutzung, 19, 82
NAS-Server, 36
Neuronales Netz, 66
NFDI4ING, 69
Nicht-relationale Datenbank, 89
NMR, 8
NoMaD, 80
NoSQL, 89
Nutzer, 35

OAuth2, 57
ODC, 15
OER, 90
Ontologie, 76, 77
Open Data Commons, 15
Open-Access, 20
Open-Source, 18, 61, 91
OpenAIRE, 20
OpenAPI, 57
OpenID, 49
ORCID, 81
Origin, 61
Outlier, 61, 63

PANGAEA, 91
Passphrase, 88
Passwort, 57, 88
Patent, 39
Patente, 13
Peer-Review, 91
PGP, 88
PID, 10, 80
PNG, 10
PostgreSQL, 90
Primärdaten, 8
Privater Schlüssel, 88
Programmiersprache, 18
Projektnetzlaufwerk, 19
Projektplanung, 17
proprietär, 82
Provenance Tracking, 67
Prozess, 68, 71
Prüfsumme, 73
Publikation, 58
Python, 7, 18, 66, 89
Python-Code, 24

QtiPlot, 61
Quellcode, 7

R, 61
RDF-Darstellung, 71
RDF-Datenmodell, 69
RDMO, 17
re3data, 80
Rechtemanagement, 42
rechtlich, 19
Recycling, 82
Registerkarte, 54
Regressionsgerade, 64
Relationale Datenbank, 89
Repositorium, 10, 80, 82, 90
Repository, 48
ResearchData, 91
responsiv, 35
Ressourcen, 20, 45, 51, 56
REST-API, 35
Revision, 54
RIsources, 80
Rohdaten, 8
Rollen, 42, 76
RWTH Aachen, 91

Sammlung, 51, 54, 55
Schlüssel, 88
Schlüsselwort, 10
Schnittstelle, 20, 35
SciMesh, 67
Sekundärdaten, 8
Sekundärnutzung, 82
Server, 17

Shibboleth, 49
Sicherheit, 19
Skript, 19
Software, 17
Speichermedien, 87
Speicherstrategie, 87
Sperrfrist, 19
Sprachmodell, 83
SQL, 89
Statistik, 61
Stichprobe, 68, 74
String, 53
Strukturierte Daten, 83
Suchfunktion, 45, 58
Sysadmins, 35
Systemadministrator, 88

Template, 35, 39, 56
TIB Hannover, 80
TIFF, 10
Token, 57
Tool, 47
Trainingsdaten, 5
Turtle, 71, 76

Unstrukturierte Daten, 84
Urheber, 19
Urheberrecht, 13, 19
URI, 73
URL, 73

VeraCrypt, 88
Vererbung, 76
Verschlüsselung, 88
Verschlüsselungsprogramm, 88
Vertrauensniveau, 62
Veröffentlichung, 19
Veusz, 66
Visualisierung, 66
Vorlagen, 35, 51, 56

Wahrscheinlichkeitsrechnung, 61
Wiederverwendung, 82
Windows, 36
Wissensgraph, 67, 70, 76

YAML-LD, 76

Zeitstempel, 26, 39
Zenodo, 10, 58, 80, 81, 91
ZIP, 81
Zugangstoken, 57
Zugriffsberechtigung, 54
Zugänglichkeit, 19

Öffentlicher Schlüssel, 88