HeFDI Data Talk

Date	Topic	Presenter(s)
06 October 2023	Kadi4Mat – Karlsruhe Data Infrastructure for Materials Science	Dr. Michael Selzer (Institute for Nanotechnology, KIT)

Abstract:
In research today, many digital results are produced from experiments, simulations or measurements. Often, however, only the most important data is stored in the long term, for example if it is part of a scientific publication. The Kadi4Mat software creates technical possibilities to store as much of the accumulating data as possible in such a way that a benefit can be drawn from it. On the one hand, it is particularly important to record the data simply and as automatically as possible, but on the other hand, it is also important to store the information correctly. Kadi4Mat is needed to store information in a meaningful way and to make it available in the long term. Only in this way the accumulated data can be efficiently searched, analysed, visualised and compared with each other in order to maximize the information gain from scientific work.

About the HeFDI Data Talks:

The HeFDI Data Talks are a bi-weekly open information and discussion event focused on data management in the context of science, in which relevant NFDI consortia as well as research data management services present themselves. The series discusses current topics and presents numerous – including local and regional – tools and services. The HeFDI Data Talks are an offer of the HeFDI Initiative (Landesinitiative HeFDI), which is funded by Hesse's Ministry for Science and Arts (HMWK).

DOI-Link: https://doi.org/10.5281/zenodo.8424794; Lizenzinformation: Creative Commons Attribution 4.0 International (CC BY 4.0)

gefördert durch Hessisches Ministerium für Wissenschaft und Kunst

Karlsruhe Institute of Technology

Kadi4Mat: General Overview and Sample Applications

Michael Selzer et al.

6 October 2023

Motivation

- In materials science, understanding new materials is becoming increasingly complex
- Without appropriate **data science methods**, it will no longer be possible to manage the ever-growing volumes of heterogeneous data from simulations and experiments
- An important aspect to be able to perform corresponding data analyses smoothly is the structured storage of research (meta)data with the help of a suitable **research data infrastructure**

Research Data Infrastructure

- The development of such an infrastructure is being pursued as part of several research projects, including **FestBatt 2** [1], **POLiS** [2], **AQua** [3], **NFDI4Ing** [4], and others

- Some of the most important objectives of the infrastructure are:
  - Structured **data storage** and web based **data exchange and publication** via different interfaces
  - Integrating heterogeneous and reproducible **workflows** to make use of application-specific tools in an automated and documentable manner
  - Allowing other institutions to set up their **own installations** with the possibility to **connect** multiple of such instances

[1] https://festbatt.net/
[2] https://www.postlithiumstorage.org/
[3] https://www.iam.kit.edu/mms/5562.php
[4] https://nfdi4ing.de/

Implementation

- The concept for the research data infrastructure is a modular and generic architecture that combines the two components **electronic lab notebook (ELN)** and **repository**

- There are already several **open source** solutions for both components, but none that combines the desired features of both components
- The goal is to combine **established technologies** and systems with **new concepts**

Kadi4Mat

Karlsruhe Data Infrastructure
for Materials Science

- **Kadi4Mat** [1] is the **Karlsruhe Data Infrastructure for Materials Science**, a generic open source [2] software under active development for managing research data
- The software as a whole can best be described as a **virtual research environment**

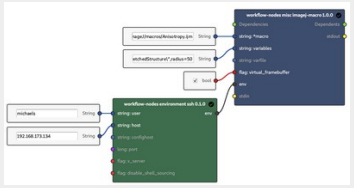
[1] Brandt et al., 2021. Kadi4Mat: A Research Data Infrastructure for Materials Science. Data Science Journal, 20(1), p.8. DOI: http://doi.org/10.5334/dsj-2021-008

[2] https://gitlab.com/iam-cms/kadi

The Kadi eco-system

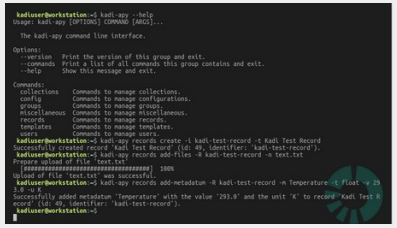
KadiStudio

Workflows



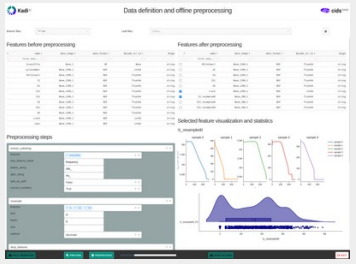
KadiAPY

(Python) API



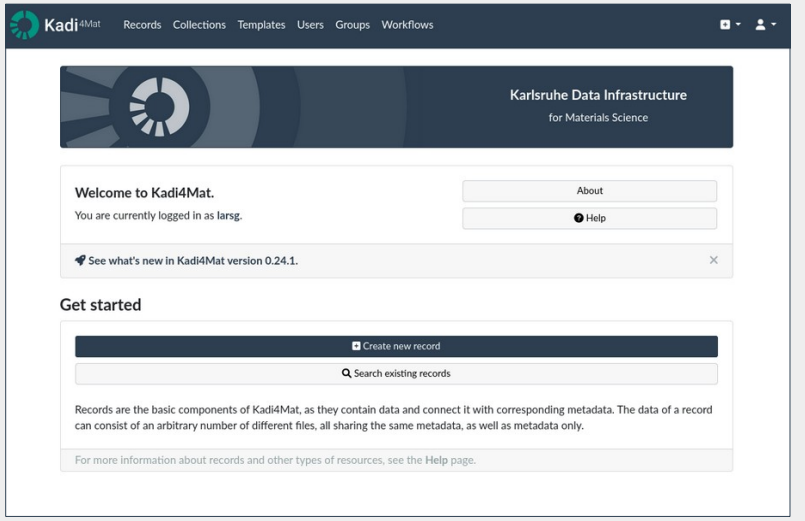
KadiAI

ML Interface



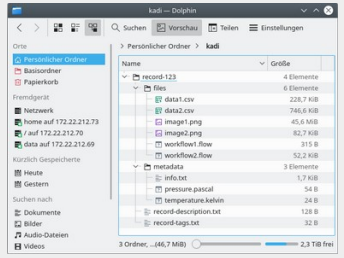
Streamline <u>your</u> research

KadiWeb



KadiFS

Filesystem



Kadi4Mat in the Research Data Lifecycle

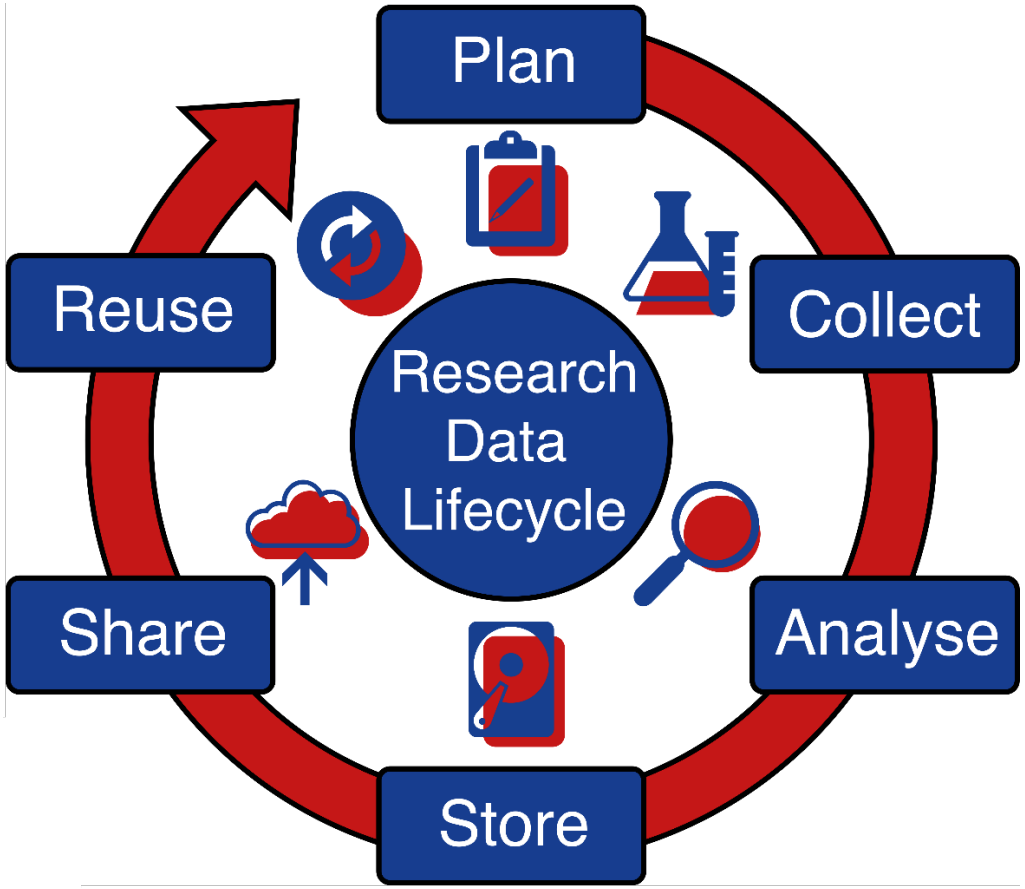


Image: Ephraim Schoof (HIU, Germany)

ELN → "ELN 2.0"

- Focused on automation and reproducibility

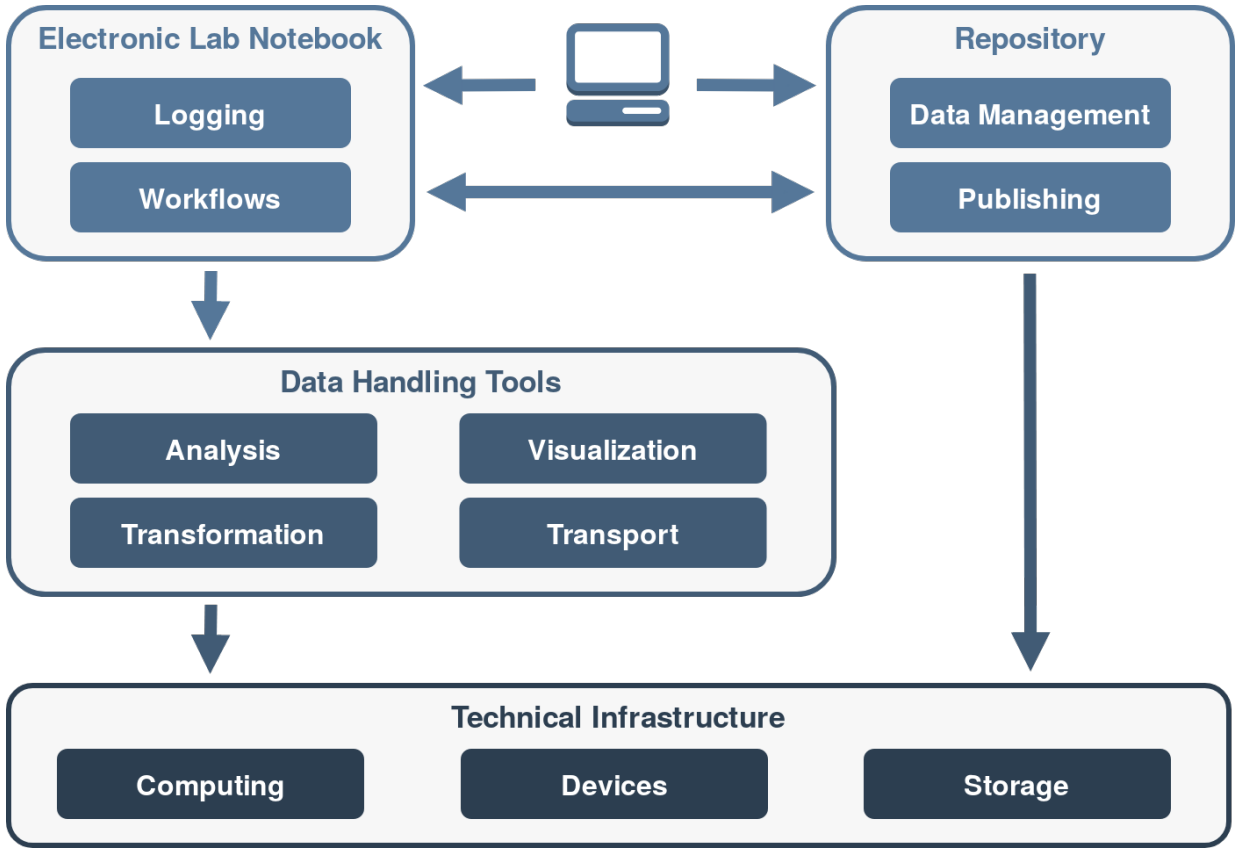
Repository → "Community repository"

- Focused on managing and sharing "warm data" within a community

- Other steps, like **publishing data** (e.g. via Zenodo), can be offered by integrating with existing software as much as possible

Image: https://zenodo.org; CC BY 4.0

Conceptual Overview of Kadi4Mat



Electronic Lab Notebook

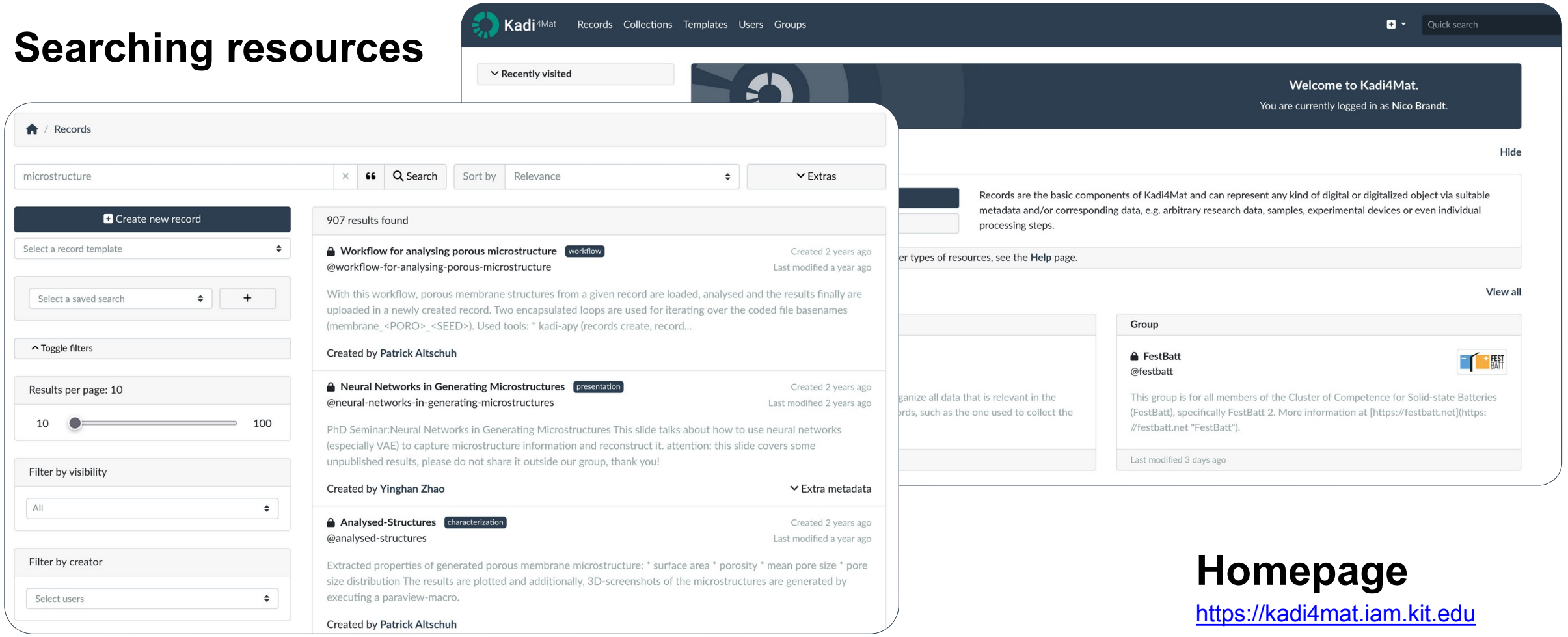
- Documentation and data provenance
- Heterogeneous and reproducible workflows using a compendium of data handling tools

Repository

- Structured data storage and exchange with user-defined access permissions
- Publishing data

Web-Interface of Kadi4Mat

Searching resources



Homepage

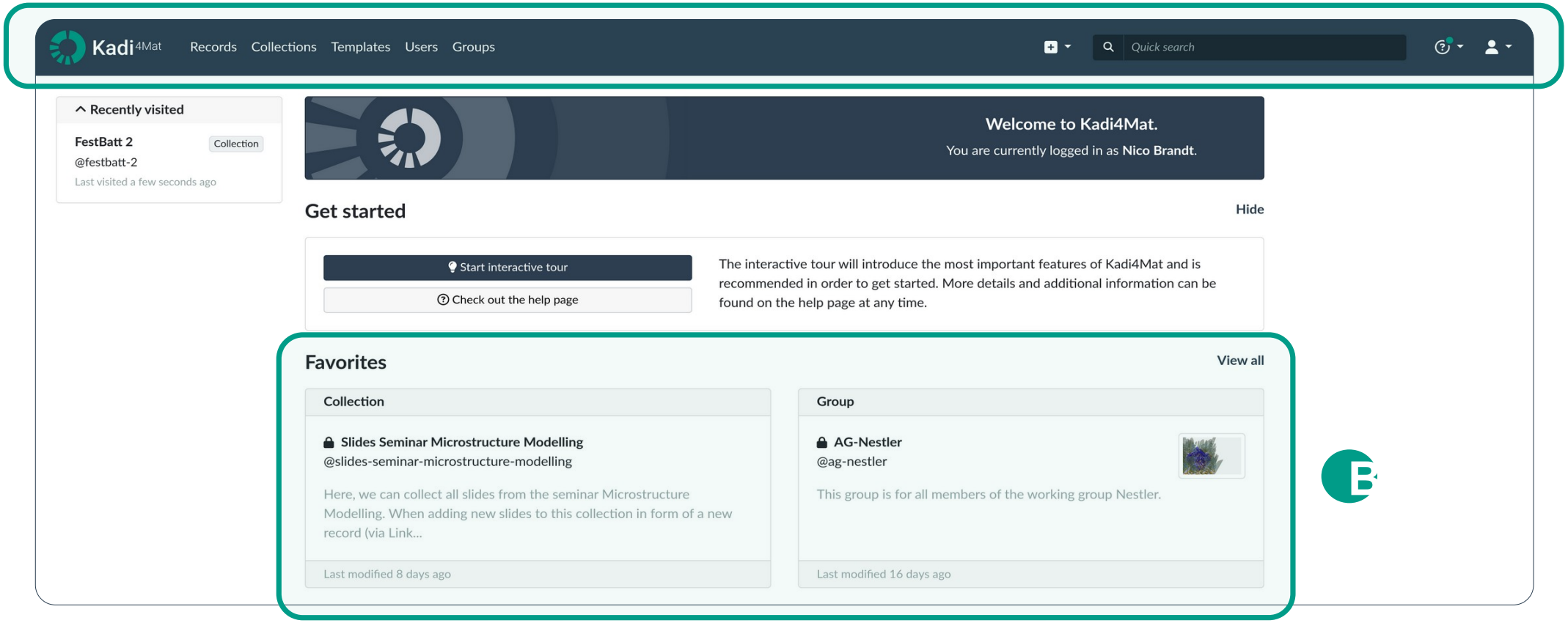
https://kadi4mat.iam.kit.edu

https://demo-kadi4mat.iam.kit.edu/

Basic Structure of Kadi4Mat

- **Records:** Combine data with generic, user-defined metadata
- **Collections:** Group multiple records and/or other collections
- **Templates:** Offer blueprints for records and generic metadata
- **Groups:** Group multiple users to ease access permission management

Basic Structure of Kadi4Mat

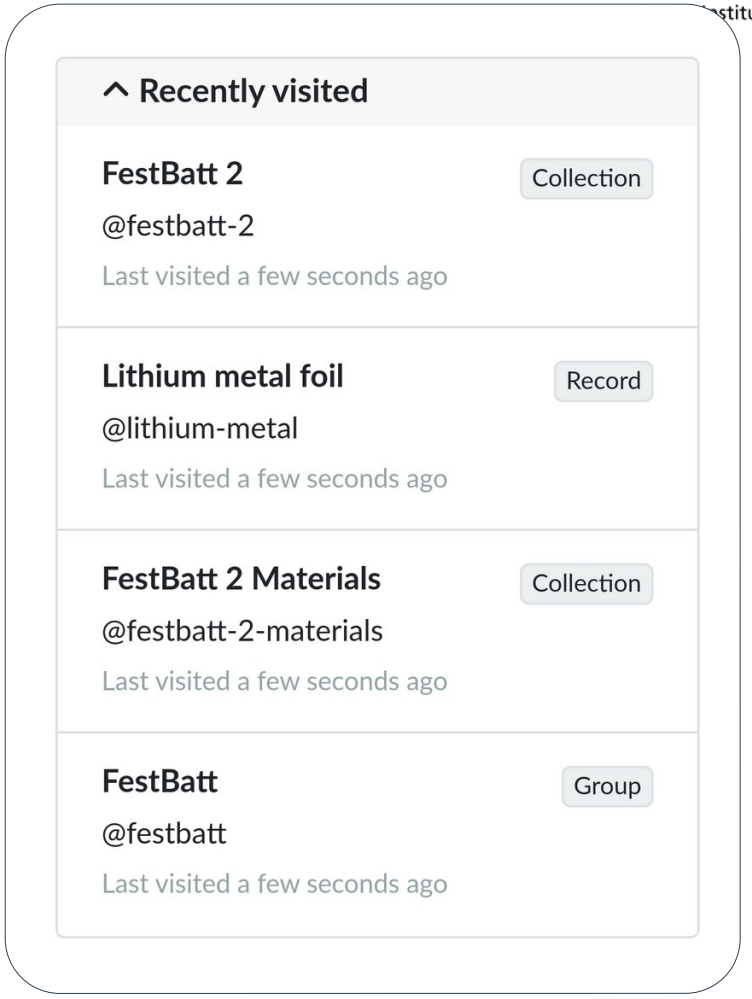


(A) Navigation bar with various links and menus

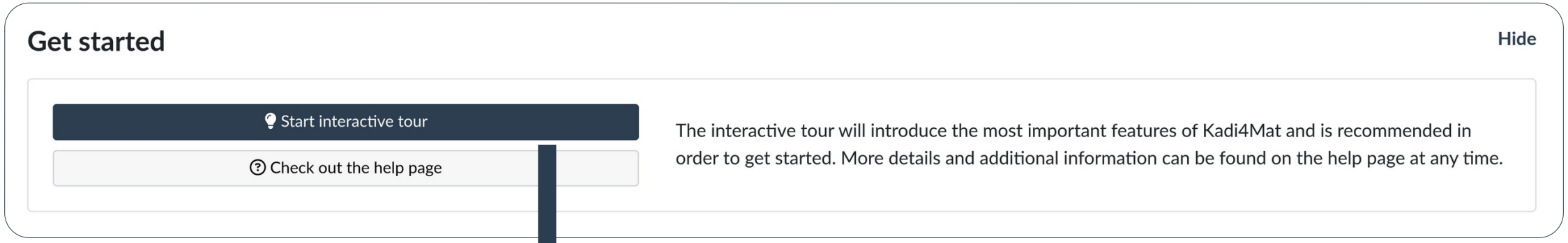
(B) Favorite resources and latest updates (customizable)

New Feature of Kadi4Mat: Recently Visited Resources

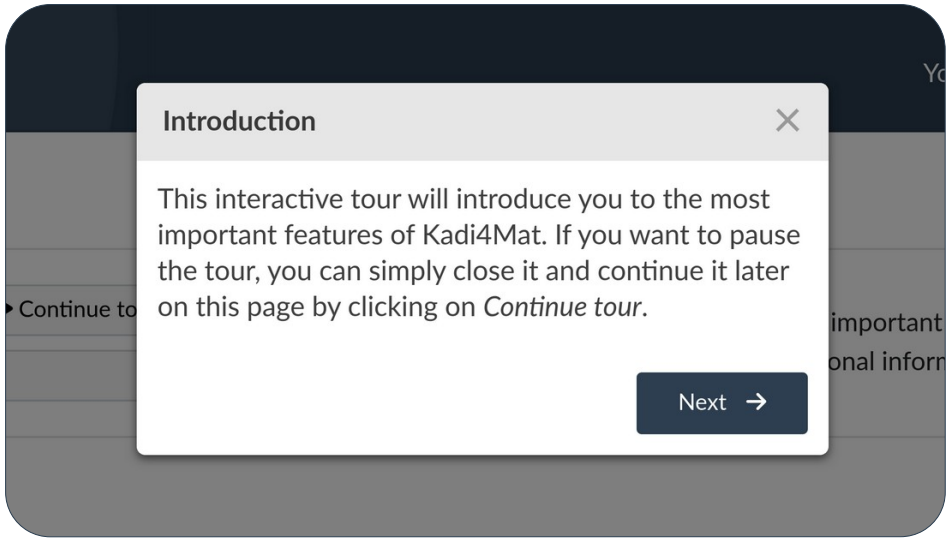
- The five most recently visited resources are shown in a new overview on the left next to the main content
- This should aid with e.g. navigating sub-collection hierarchies



New Feature of Kadi4Mat: Interactive Tour

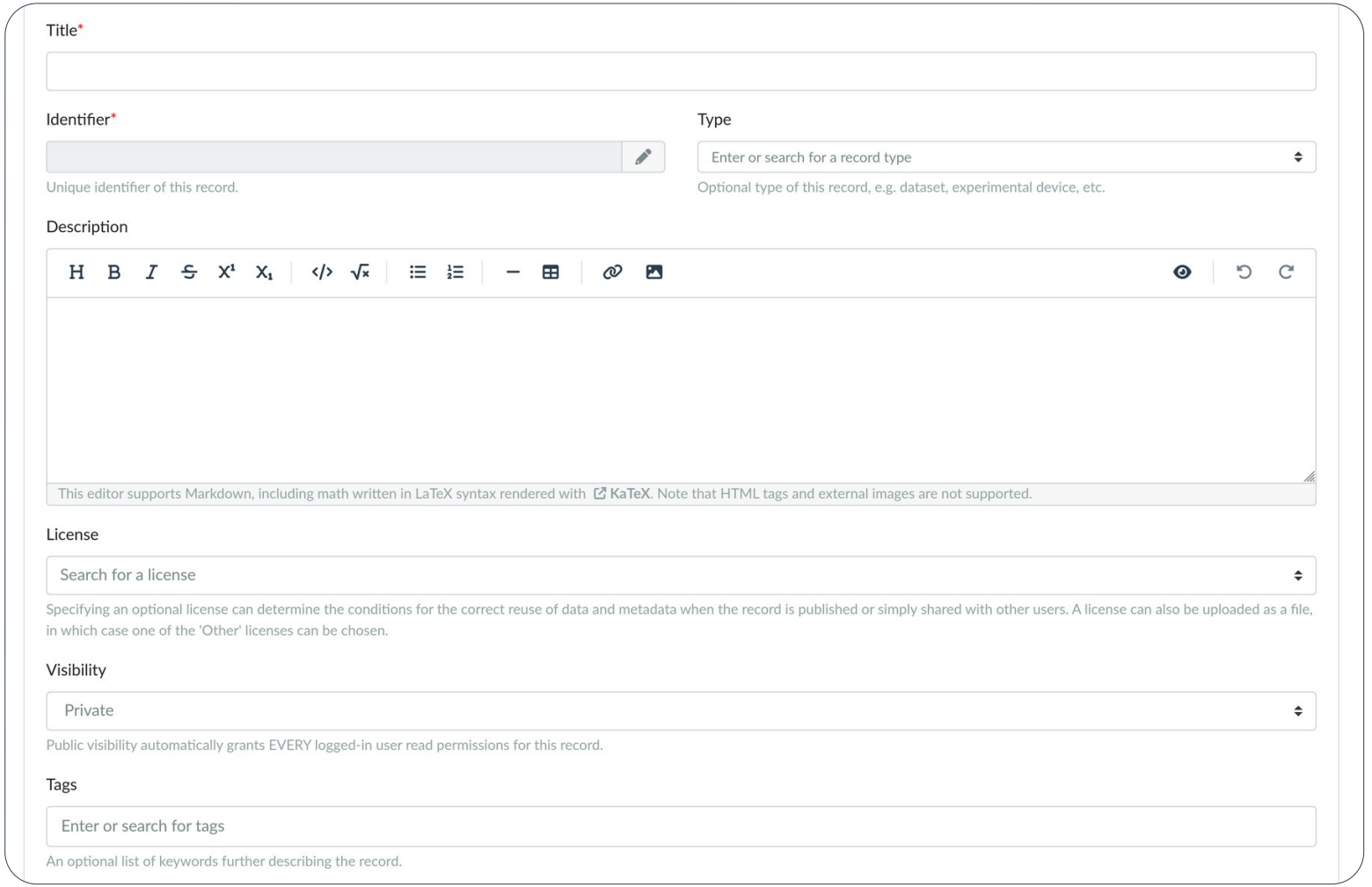


- The first interactive tour introduces the basic functionality of Kadi4Mat with focus on records



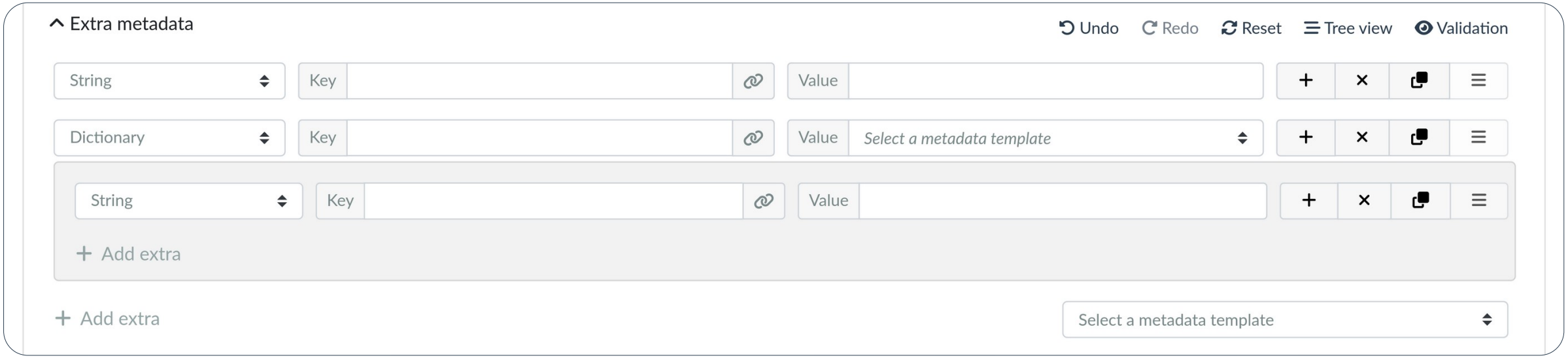
Basic Feature of Kadi4Mat: Creating new Records

- **Basic metadata schema for each record**
  - Title (required)
  - Unique identifier (required)
  - Type (e.g. *dataset*, *device*, ...)
  - Description (supports Markdown syntax and LaTeX formulas)
  - License
  - Visibility (*private* or *public*)
  - Tags (a.k.a. keywords)

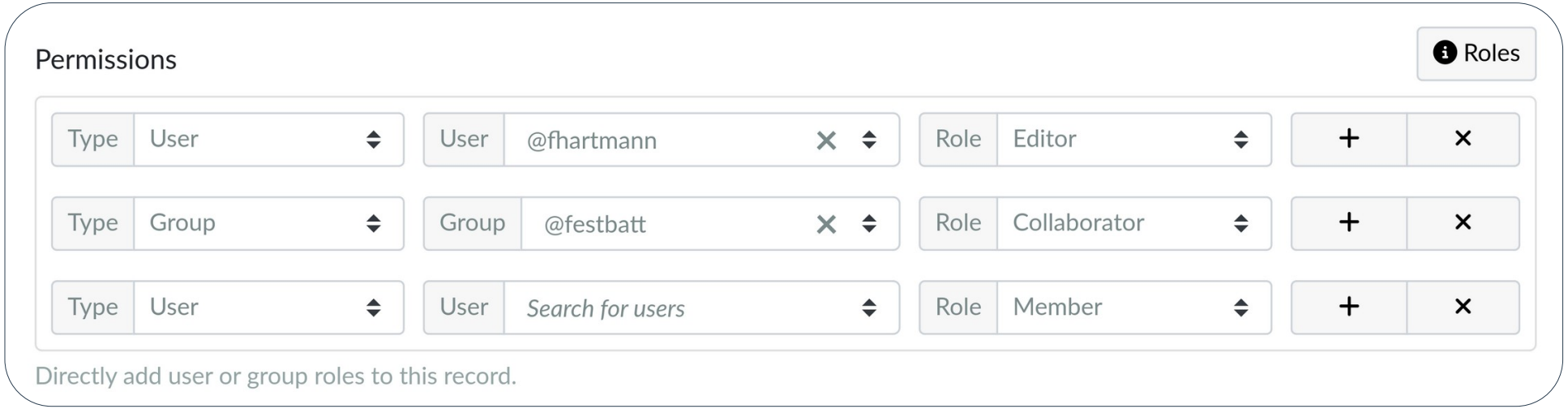


Basic Feature of Kadi4Mat: Creating new Records

- **Generic metadata (schema free)**
  - To specify user- and application-specific metadata, optionally based on existing schemes
  - Each metadatum consists of a key-value pair of a certain type, allowing for either literals or nested values
    - **Literal values:** String, Integer, Float, Boolean (*true* or *false*), Date
    - **Nested values:** Dictionary, List (uses *indices* instead of *keys*)

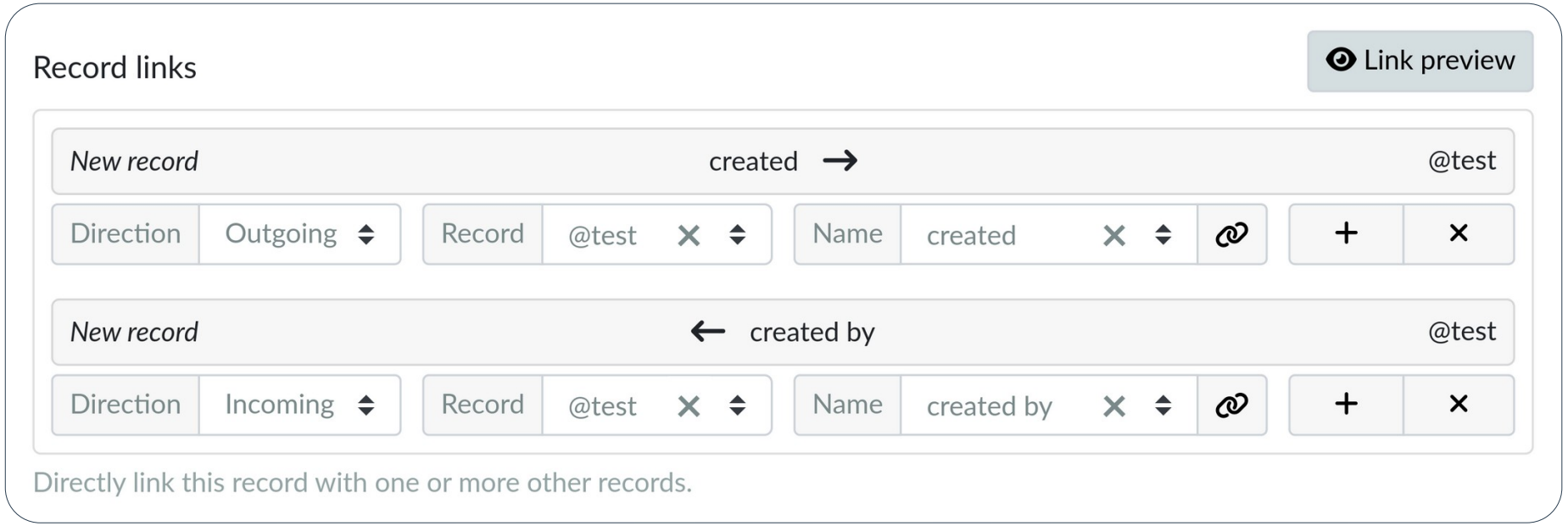


New Feature of Kadi4Mat: Permissions Editor



- Permissions for users or groups can now be specified directly when creating new resources
- The same is possible within record templates, so the editor can be prefilled when creating a corresponding record

New Feature of Kadi4Mat: Record Link Editor

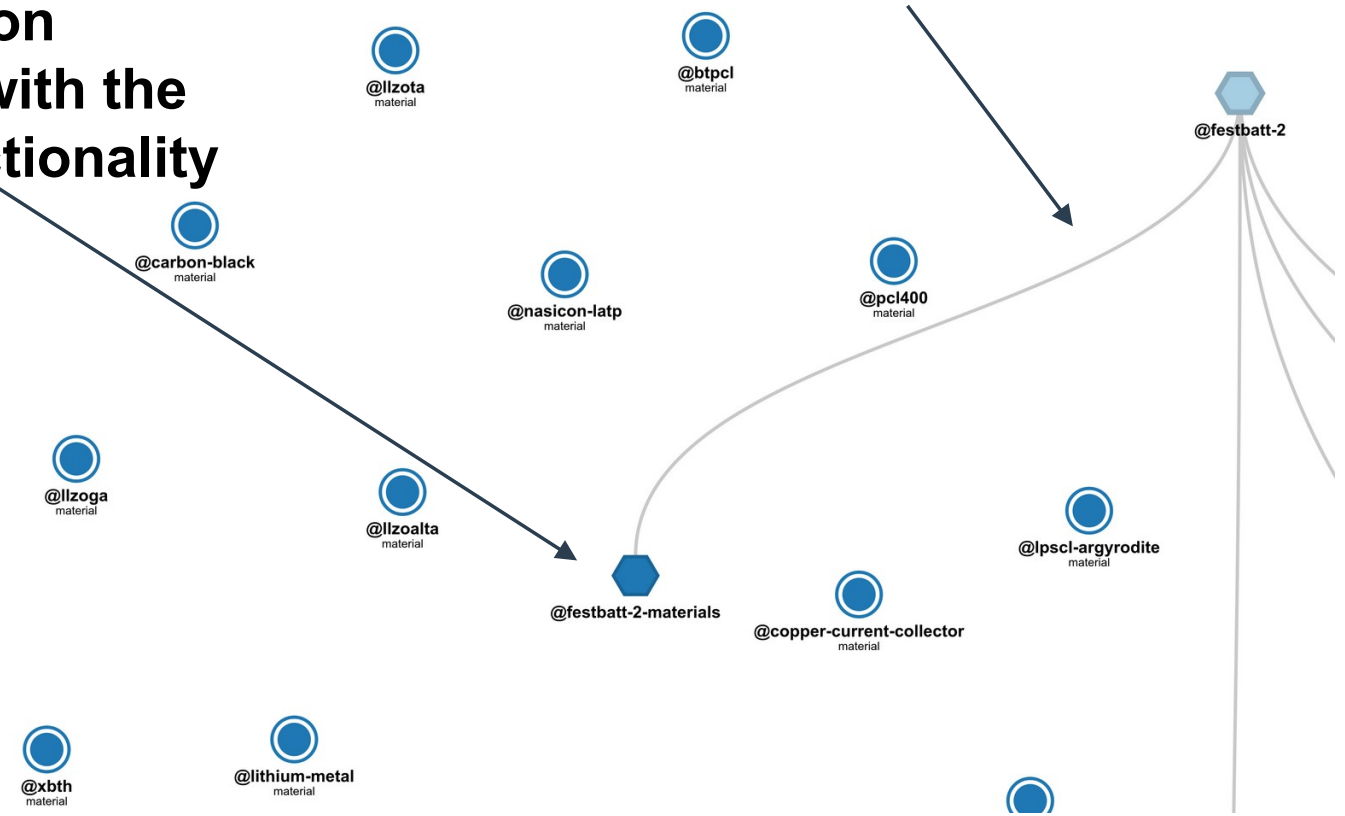


- Similarly, when creating a new record, it is now possible to directly link it with other records using a new editor, which can also show a small preview for each link
- The same is possible within record templates, so the editor can be prefilled when creating a corresponding record

New Feature of Kadi4Mat: Interactive Collection Visualization

Sub/Child collections (loadable on demand) with the same functionality

Records (and their links, if applicable) within a specific collection



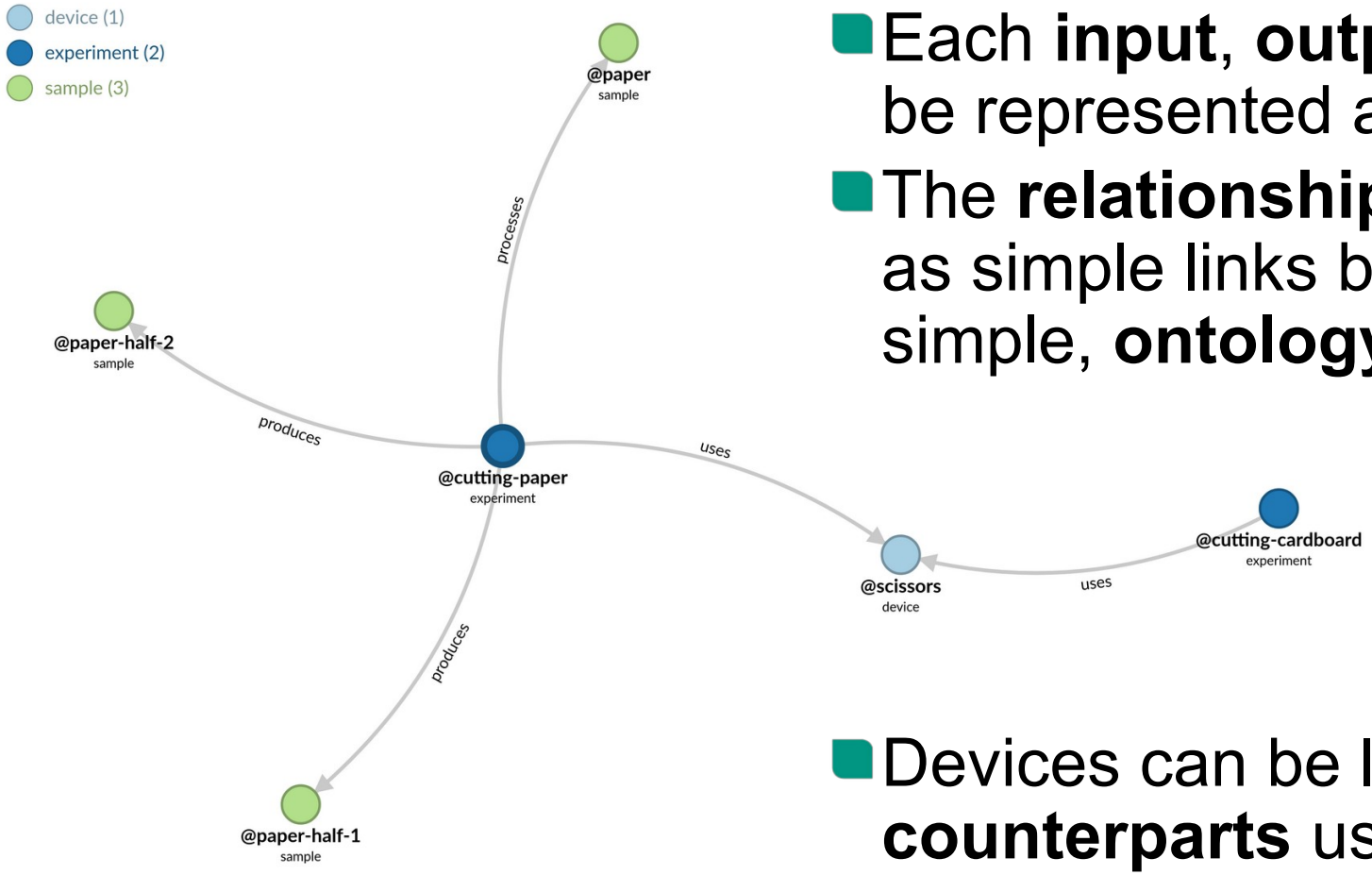
- An interactive visualization of collections has been added, similar to the existing one for record links
- Next to records, it also allows visualizing the collection hierarchy (and the respective records)

Example: Using Kadi4Mat in Experiments

- A simple **experiment** (cutting a piece of paper) is performed using a **device** (scissors) on some **sample** (paper)

Device **Sample** **Experiment** **Results (new samples)**

Kadi4Mat in Experiments



- Each **input**, **output** and **processing** step can be represented as a record in Kadi4Mat
- The **relationships** of the steps are represented as simple links between the records, creating a simple, **ontology-like** structure

- Devices can be linked to their **digital counterparts** using e.g. QR codes

Ontologies

*"An ontology is a **formal**, **explicit** specification of a **shared conceptualization**"*

machine-readable

concepts, properties, relations, functions, constrains, axioms are explicitly defined

consensual knowledge

abstract model and simplified view of a phenomenon

Ontologies

- Important benefits ontologies provide are
  - ➔ Semantic interoperability
  - ➔ Reasoning and inference – ontologies include logical axioms and rules that capture domain knowledge and relationships
  - ➔ Domain understanding and analysis – enables exploration of complex knowledge landscapes
  - ➔ Knowledge based systems and applications – enables the creation of intelligent agents and expert systems

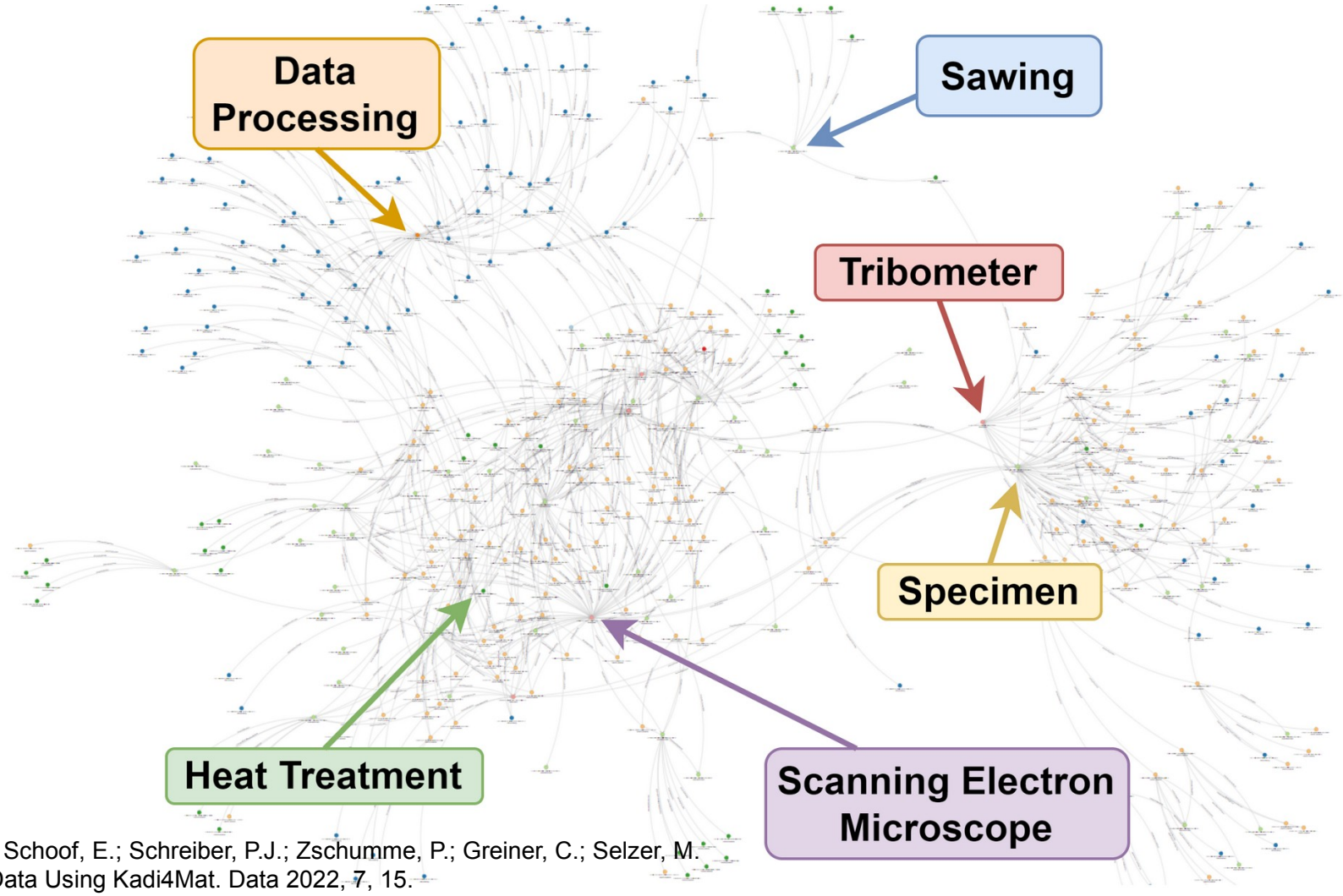
Ontologies

- Crucial for ensuring consistency, interoperability and resuability of data
- Implemented via
  - Metadata standards – metadata templates in Kadi4mat
  - Instrument/device templates

Ontologies

Brandt, N.; Garabedian, N.T.; Schoof, E.; Schreiber, P.J.; Zschumme, P.; Greiner, C.; Selzer, M.
Managing FAIR Tribological Data Using Kadi4Mat. Data 2022, 7, 15.
https://doi.org/10.3390/data7020015

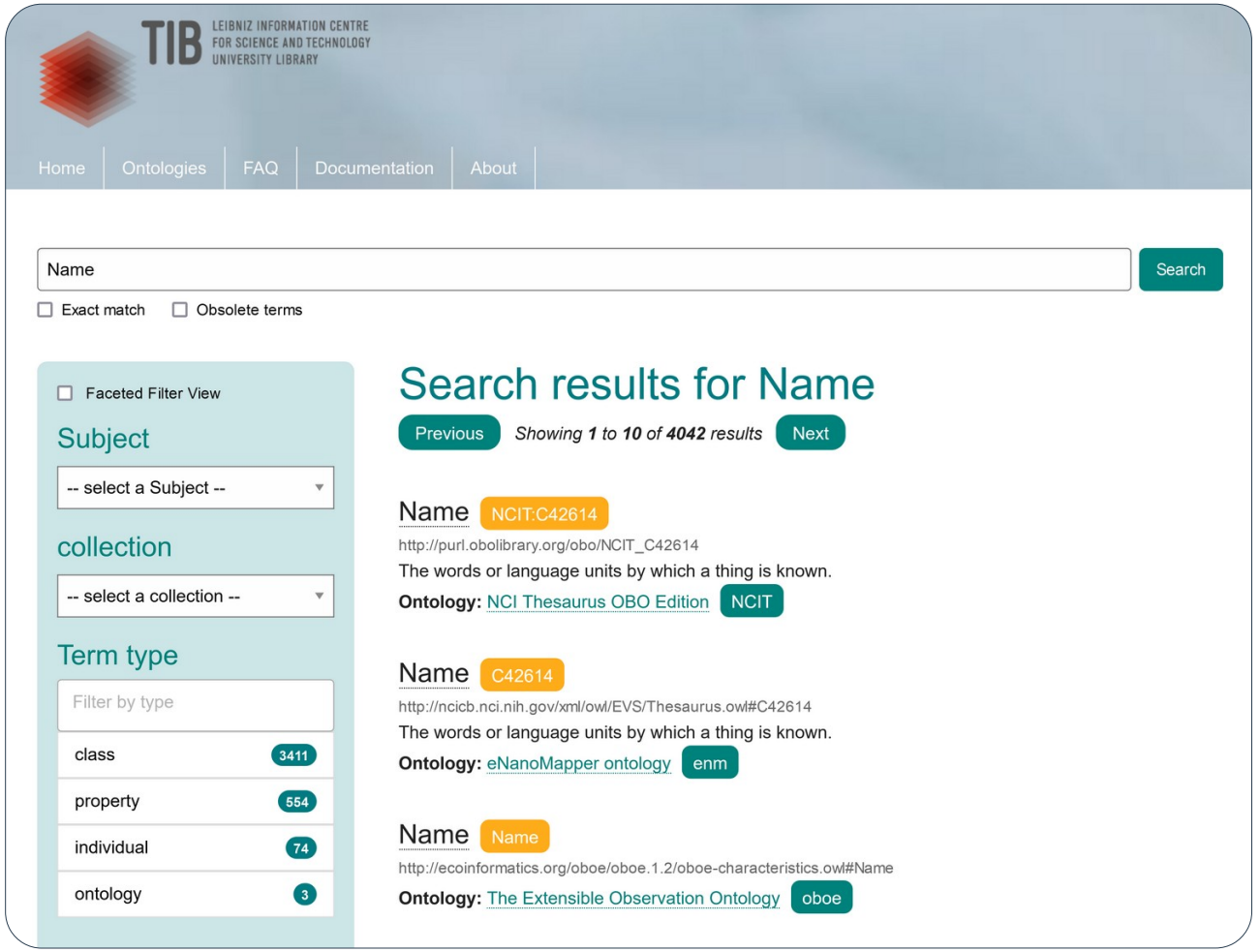
Ontologies



Brandt, N.; Garabedian, N.T.; Schoof, E.; Schreiber, P.J.; Zschumme, P.; Greiner, C.; Selzer, M. Managing FAIR Tribological Data Using Kadi4Mat. Data 2022, 7, 15. https://doi.org/10.3390/data7020015

New Feature of Kadi4Mat: Metadata Term Search

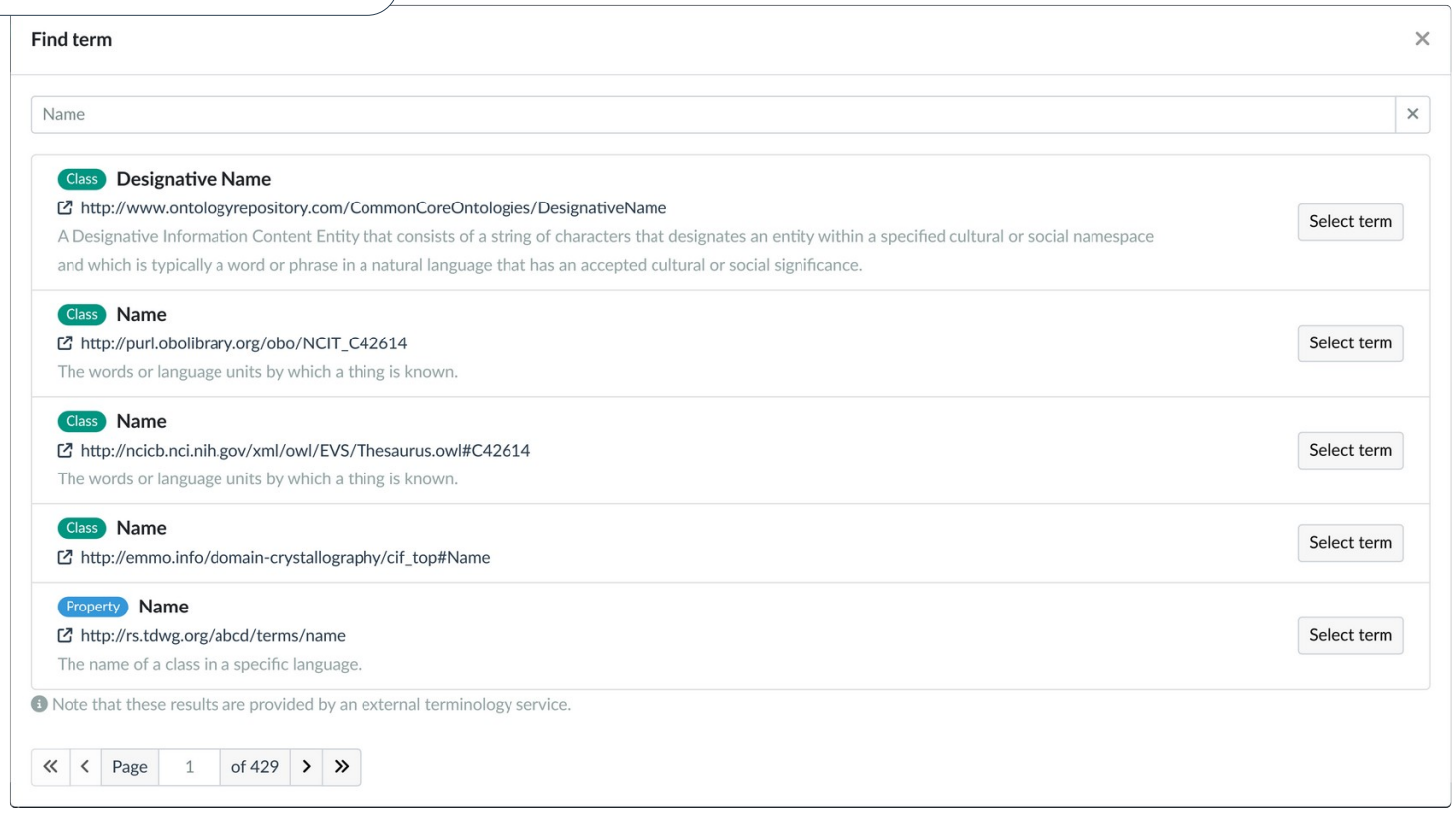
- Terminology services are one option to reuse existing **ontologies** (which define concepts and their relationship within a specific domain)



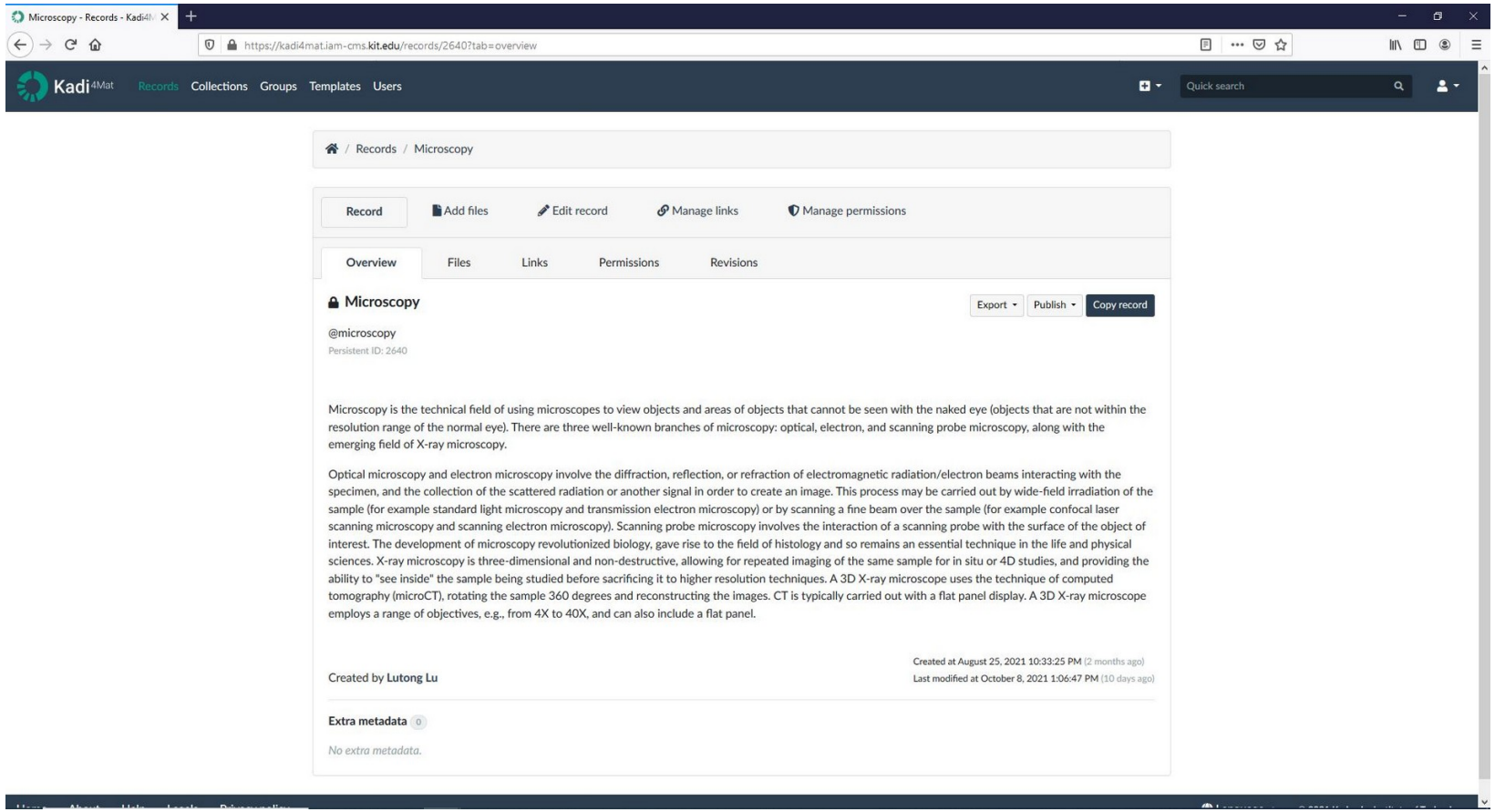
https://service.tib.eu/ts4tib

New Feature of Kadi4Mat: Metadata Term Search

- The term search is available in all places where term IRIs can be specified
- It is implemented using the **plugin infrastructure** of Kadi4Mat, so an integration of other services is easily possible



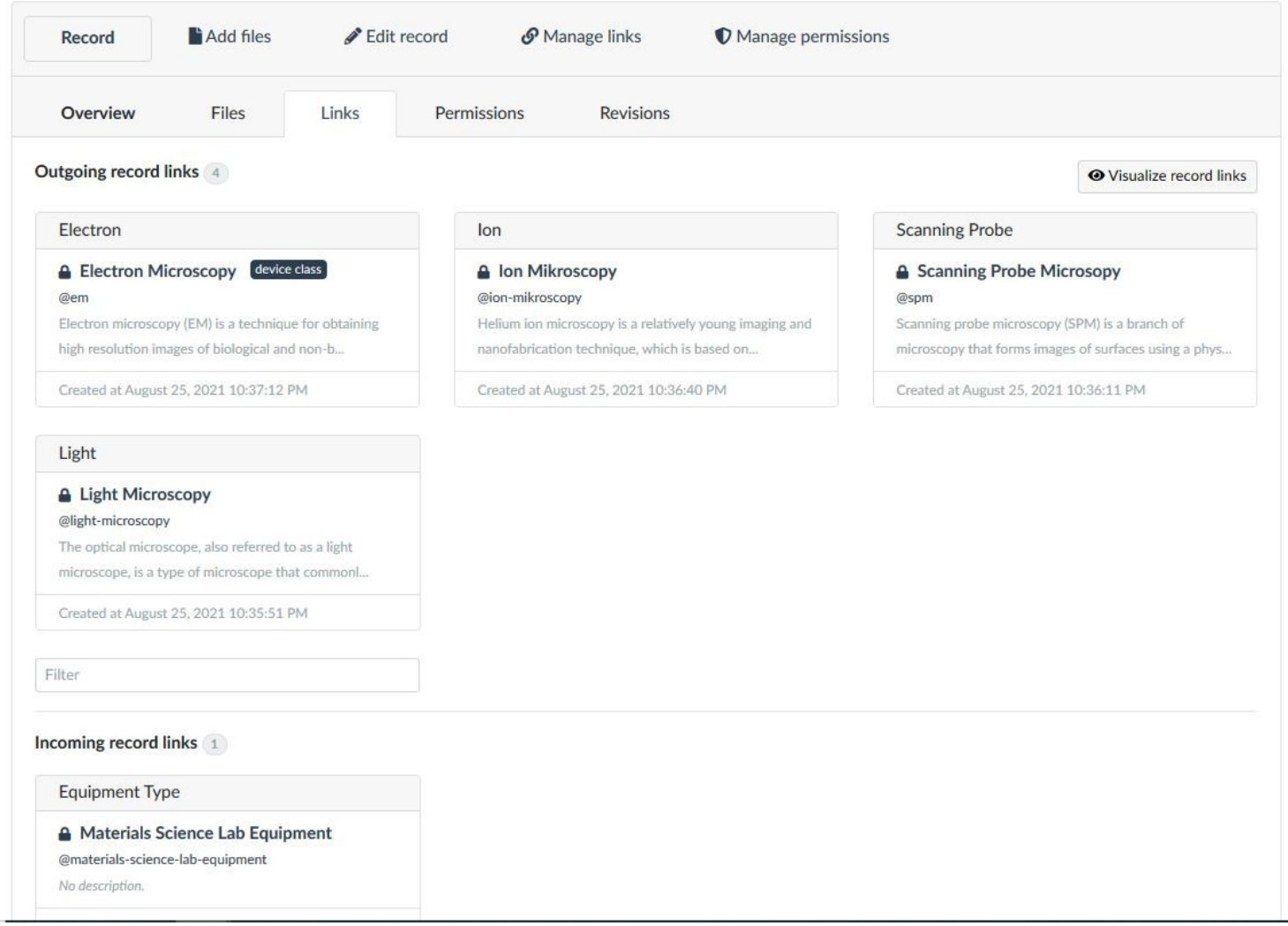
The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat



Kadi4Mat

Kadi4Mat is the **Karlsruhe Data Infrastructure for Materials Science**, a software for managing research data with the aim of combining new concepts with established technologies and existing solutions.
https://kadi4mat.iam-cms.kit.edu/about

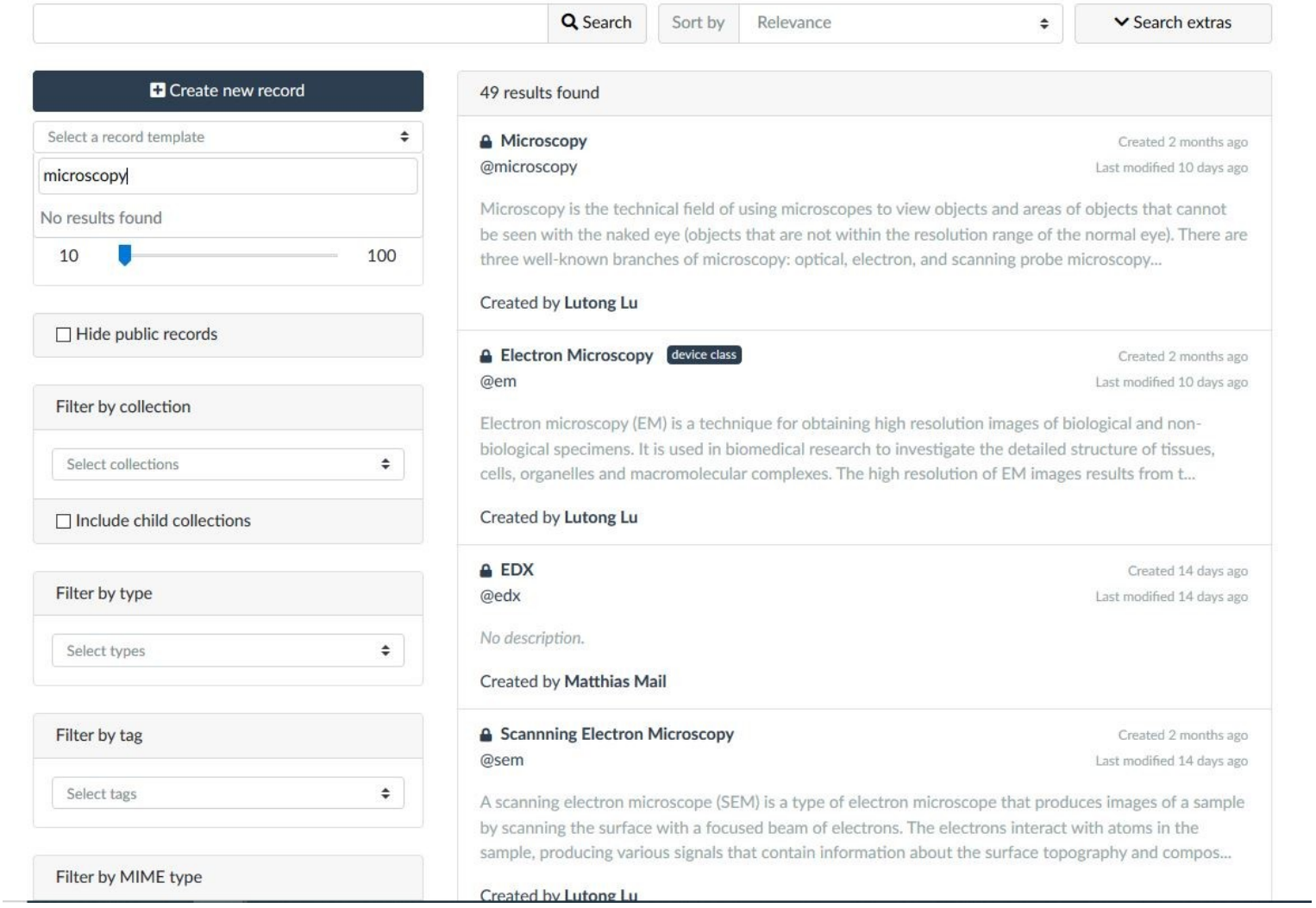
The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat - Overview



Kadi4Mat

Kadi4Mat is the **Karlsruhe Data Infrastructure for Materials Science**, a software for managing research data with the aim of combining new concepts with established technologies and existing solutions.
https://kadi4mat.iam-cms.kit.edu/about

The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat - Searching

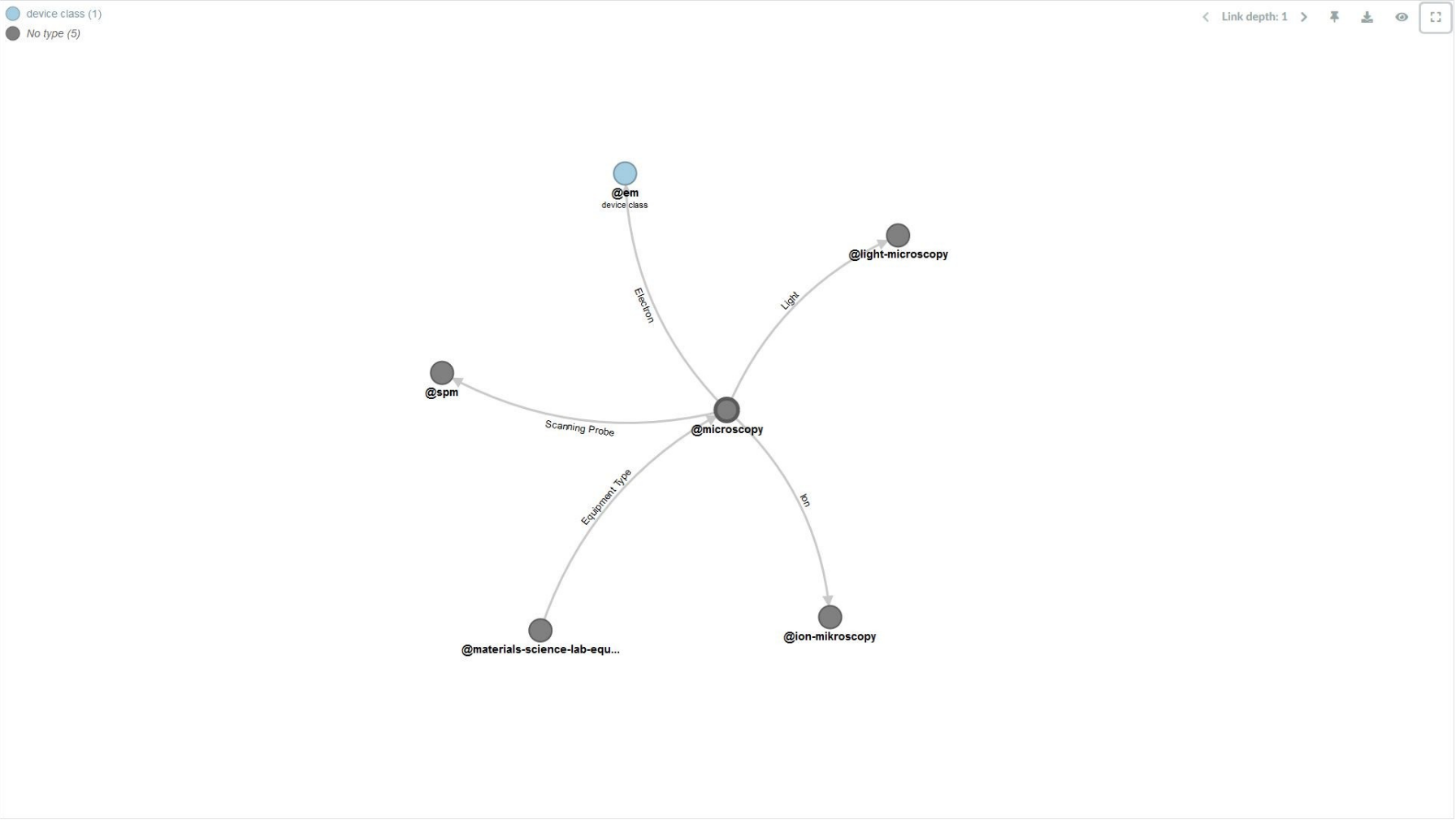


Karlsruhe Institute of Technology

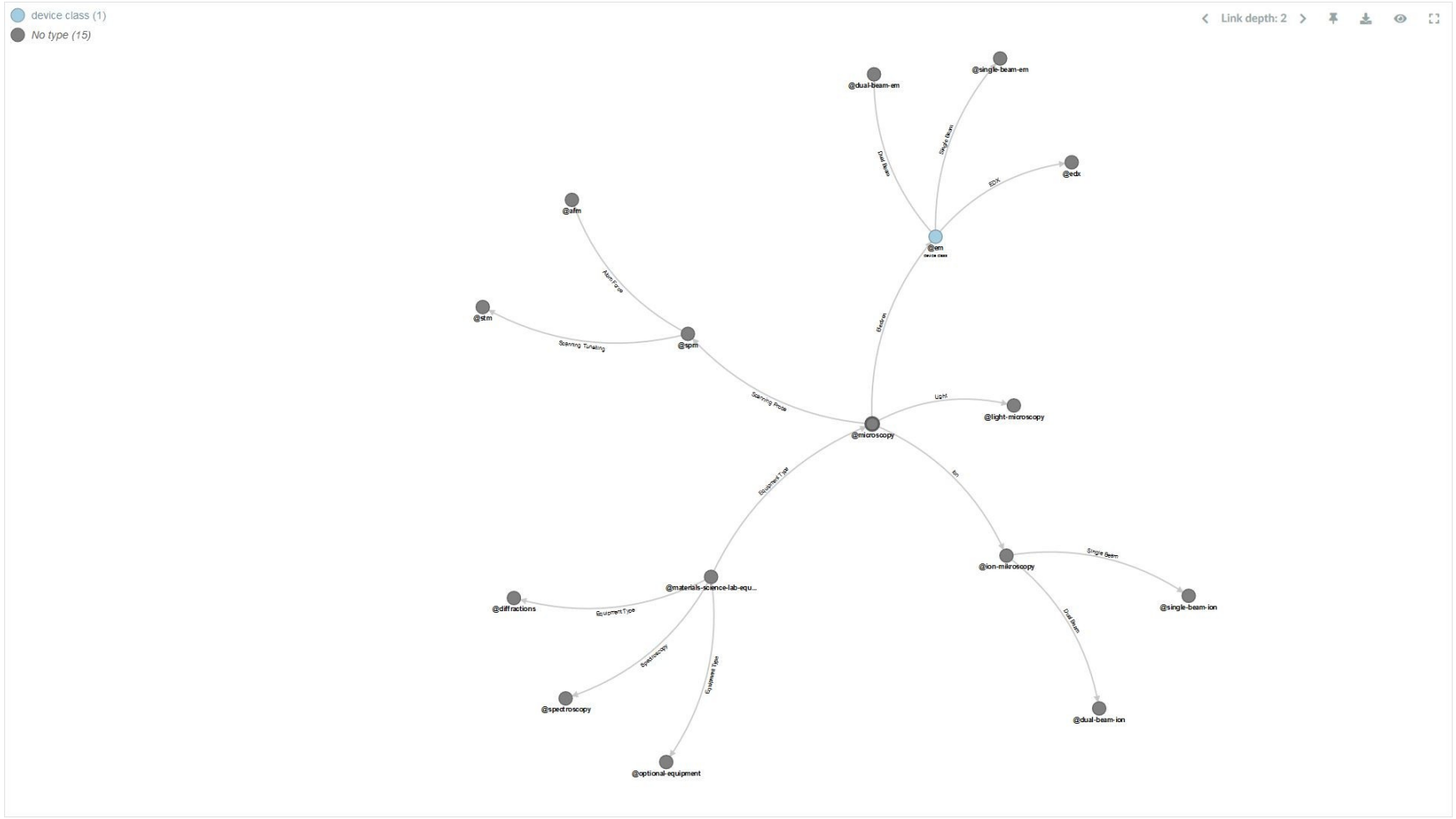
Kadi4Mat

Kadi4Mat is the **Karlsruhe Data Infrastructure for Materials Science**, a software for managing research data with the aim of combining new concepts with established technologies and existing solutions.
https://kadi4mat.iam-cms.kit.edu/about

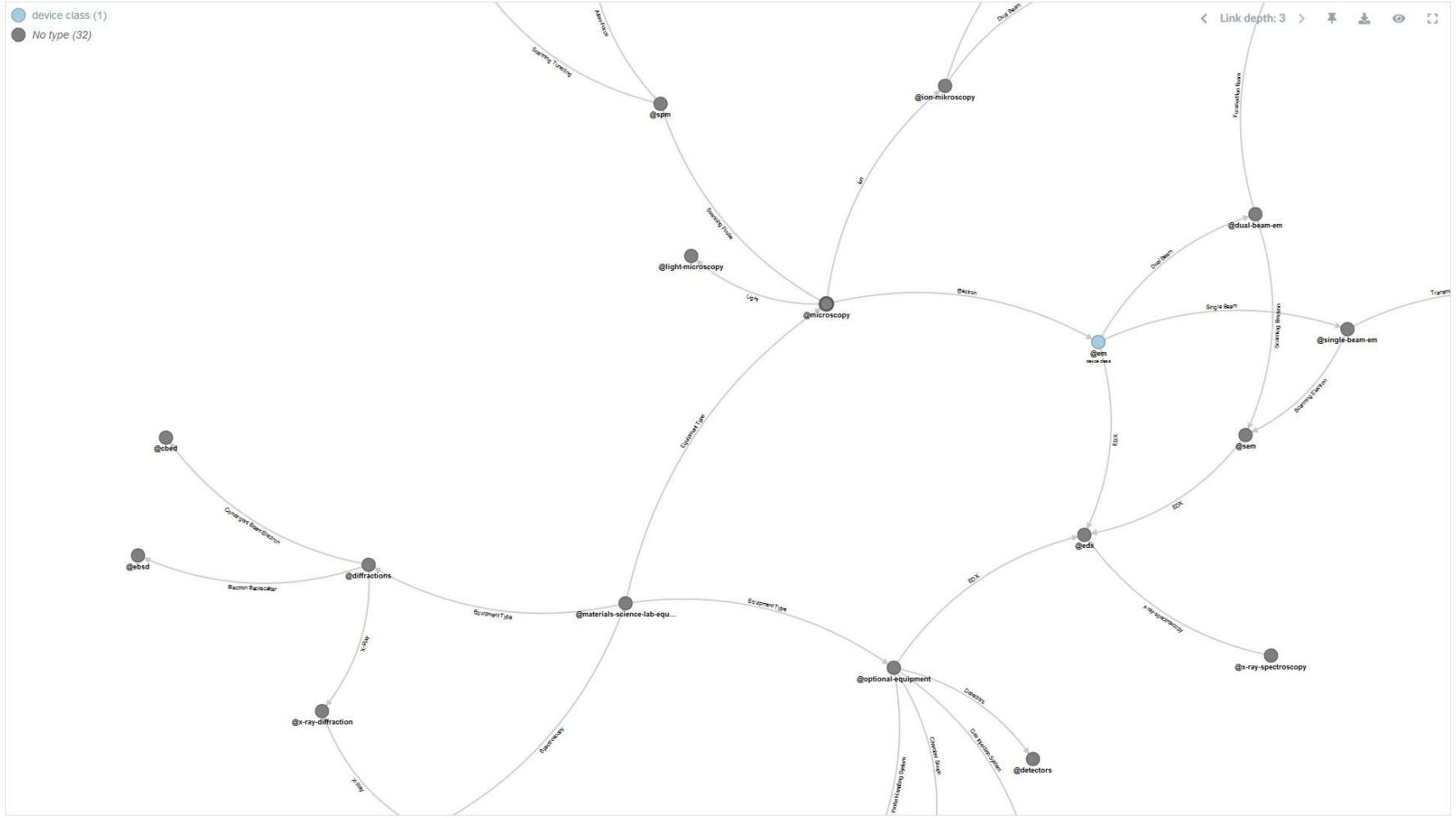
The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat - Searching



The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat - Searching

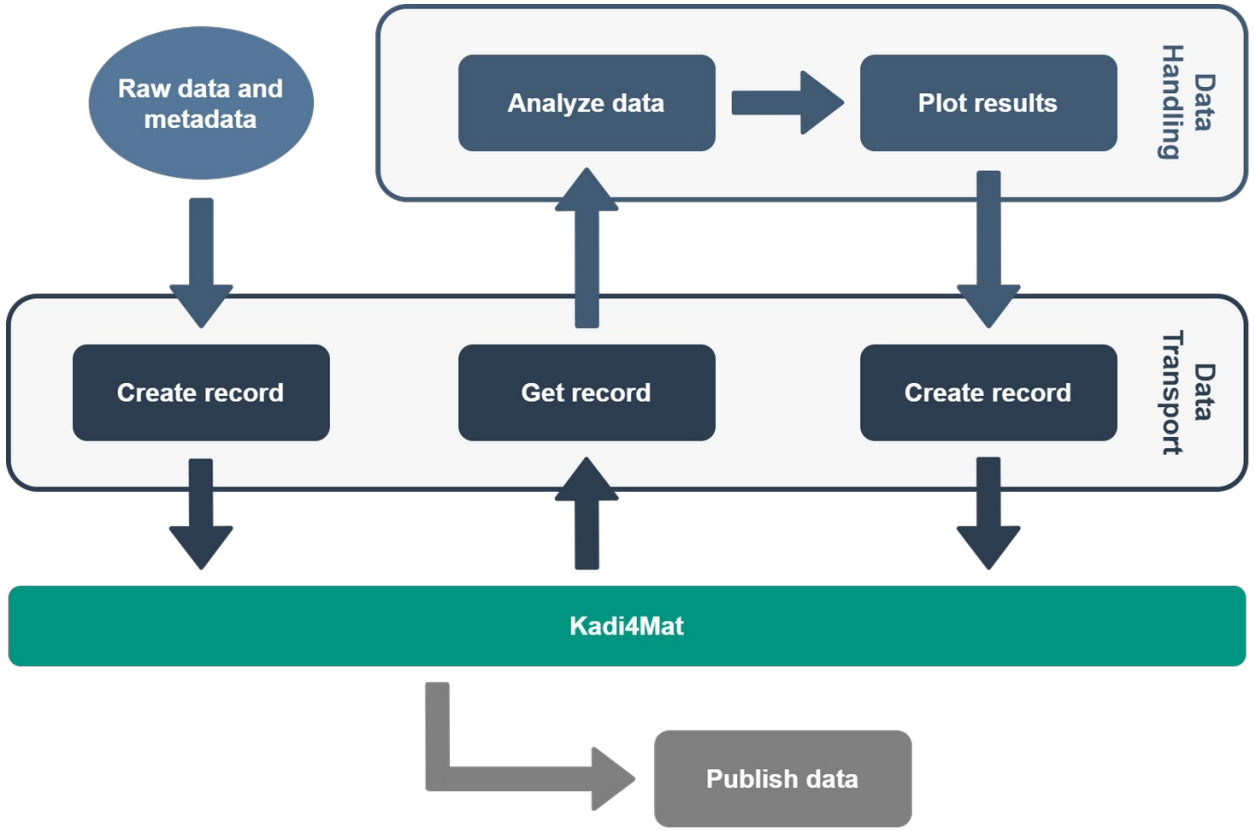


The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat - Searching



Establishing of workflows (ELN)

- A workflow is a well-defined sequence of sequential or parallel steps, which are processed as automatically as possible
- Workflows can be used to integrate existing software (e.g. Origin), devices, and **data science tools**
- A full integration of defining workflows graphically and running them via Kadi4Mat is in development



Establishing of workflows (ELN)

- A workflow is being established as a reference manual for users that will document the sequence of experiments conducted during TEM sample preparation
- The workflows have been designed for two specific applications: preparing TEM samples using FIB Lamella techniques and Ultramicrotomy
- A collection is created at the start of the workflow, to which all the records (all the experiments) are added sequentially.

Using the API via Python

- Sending HTTP request via Python is easiest done with the **requests** library [1]:

```
$ pip install requests
```

- The **GET** request shown previously, done via the browser, can be sent using this library e.g. via the interactive Python shell, using the token created before:

```
$ python
>>> import requests
>>> response = requests.get(
...     "https://kadi4mat.iam-cms.kit.edu/api",
...     headers={"Authorization": "Bearer <token>"},
... )
>>> response.json()
{...}
```

[1] https://requests.readthedocs.io/en/latest/

The Kadi-APY Library

- In parallel to the API itself that Kadi4Mat provides, a Python library called **kadi-apy** is available and under development by various people in our group and contributors

- The library is supposed to make using the API via Python as easy as possible, offering higher-level functionality to interact with Kadi4Mat as well as several quality-of-life features

- More information and documentation can be found at **https://gitlab.com/iam-cms/kadi-apy**

The Kadi-APY Library

- The library can be installed as any other library:

```
$ pip install kadi-apy
```

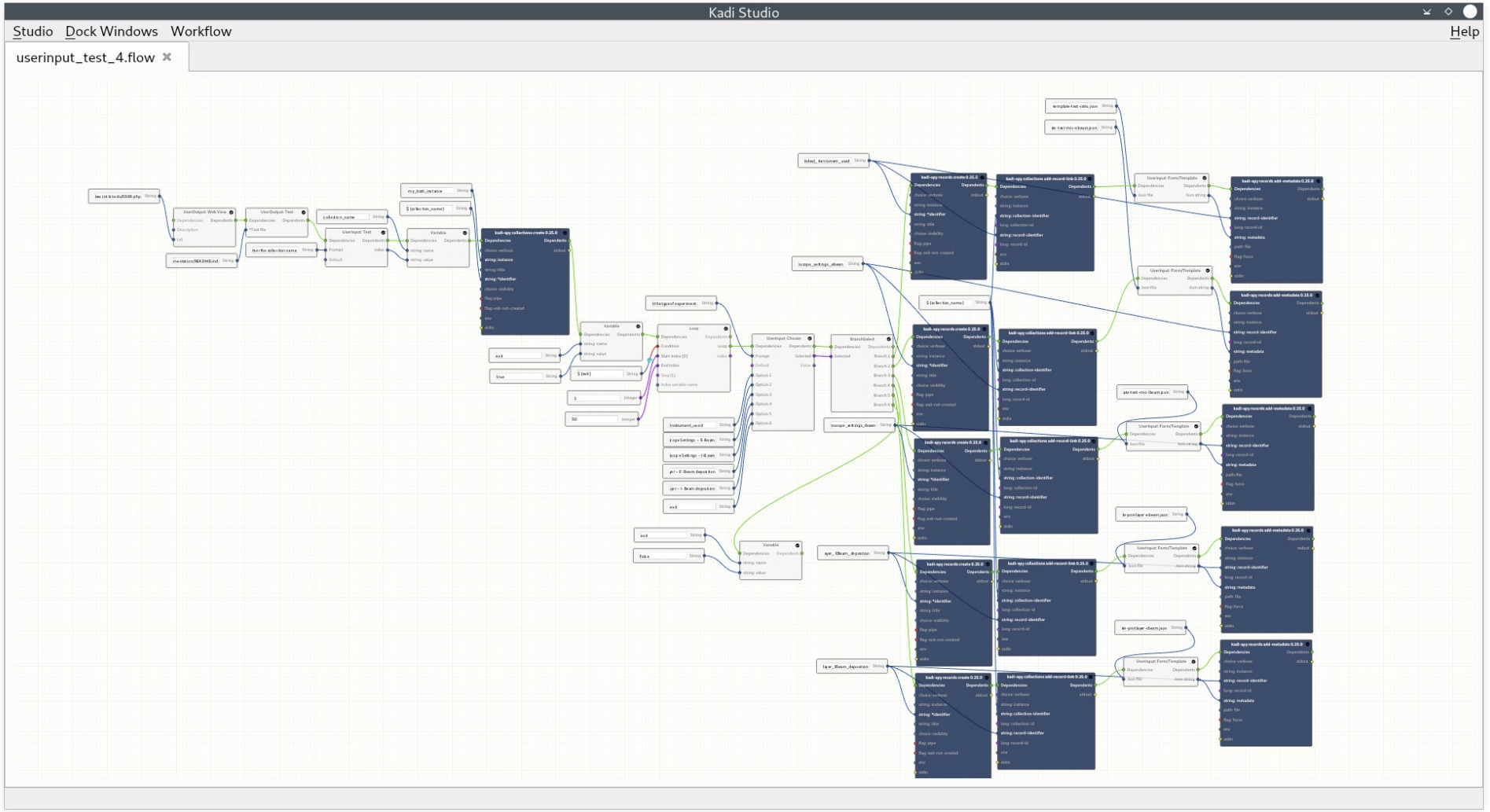
- Besides the **Python module** itself …

```
$ python
>>> import kadi_apy
>>> ...
```

- … it also includes a **command line interface (CLI)** using the **kadi-apy** command:

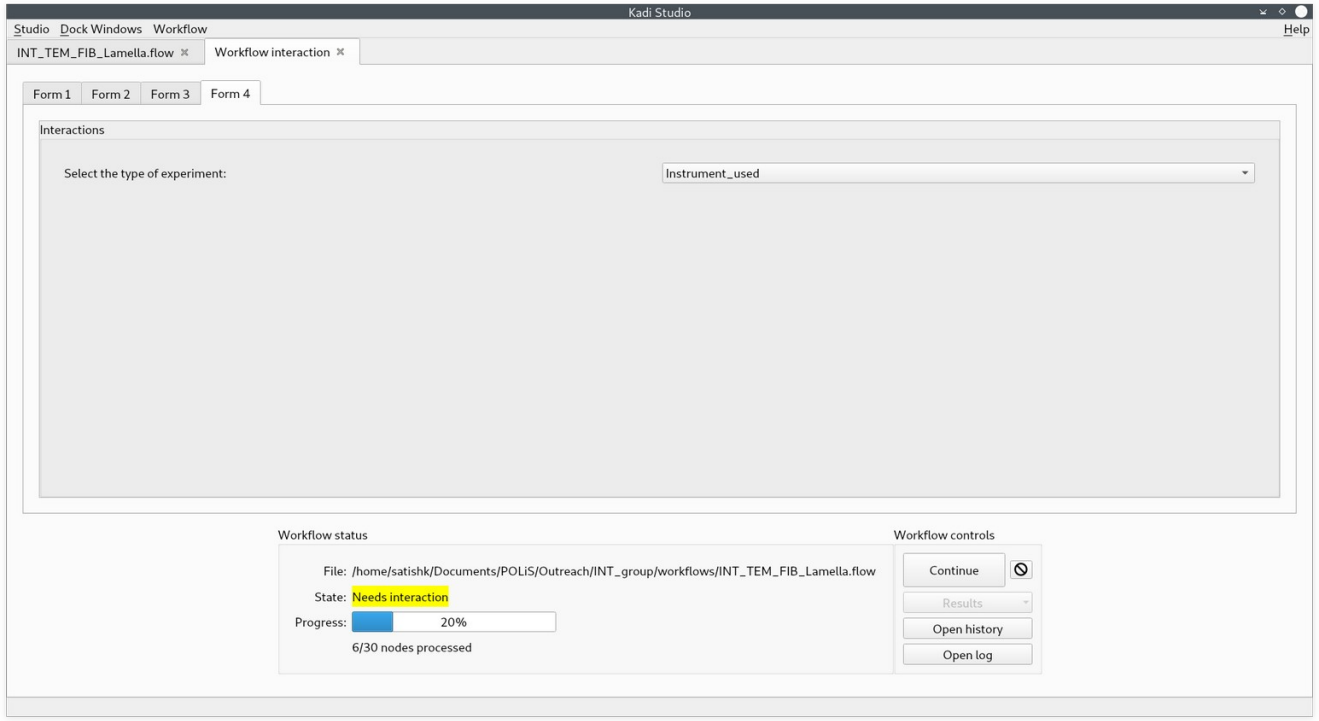
```
$ kadi-apy --help
...
```

Workflow


Kadi Studio
Studio
Dock Windows
Workflow
Help
userinput_test_4.flow

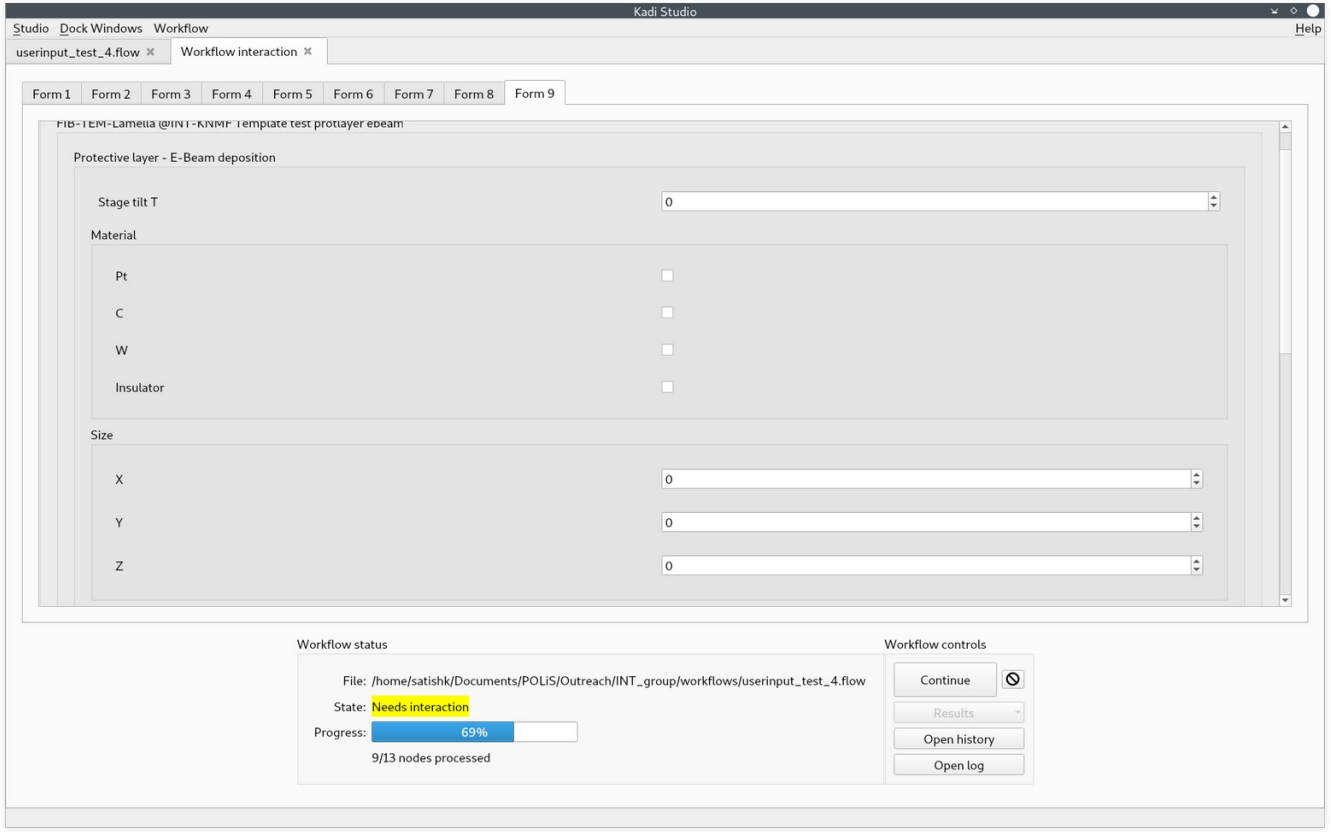
Workflow

- Used instrument can be selected
- User input form is displayed from the metadata templates exported from kadi4mat web application.
- Customized templates in json format
- can be imported into the workflow
- and displayed

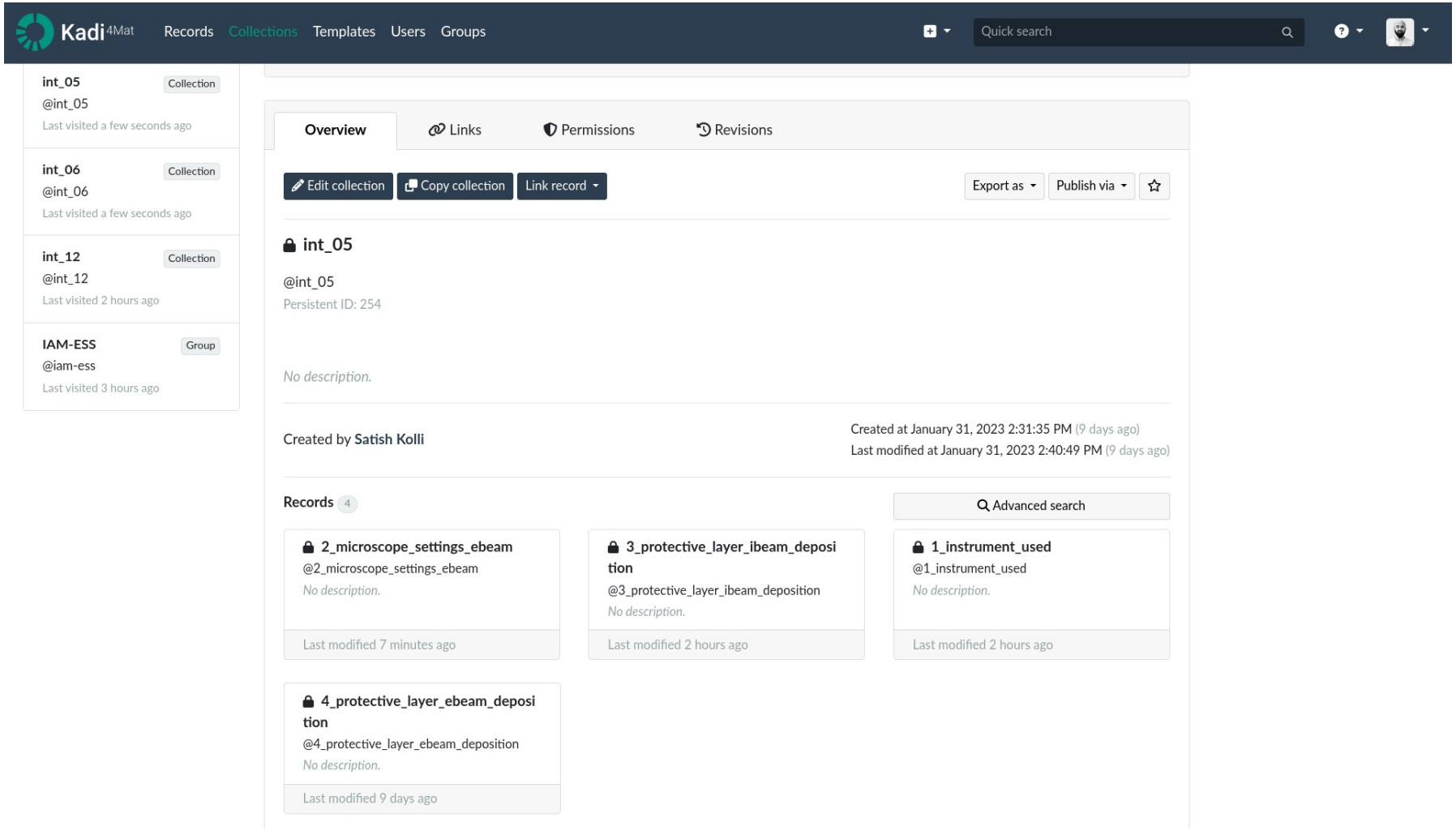


Workflow

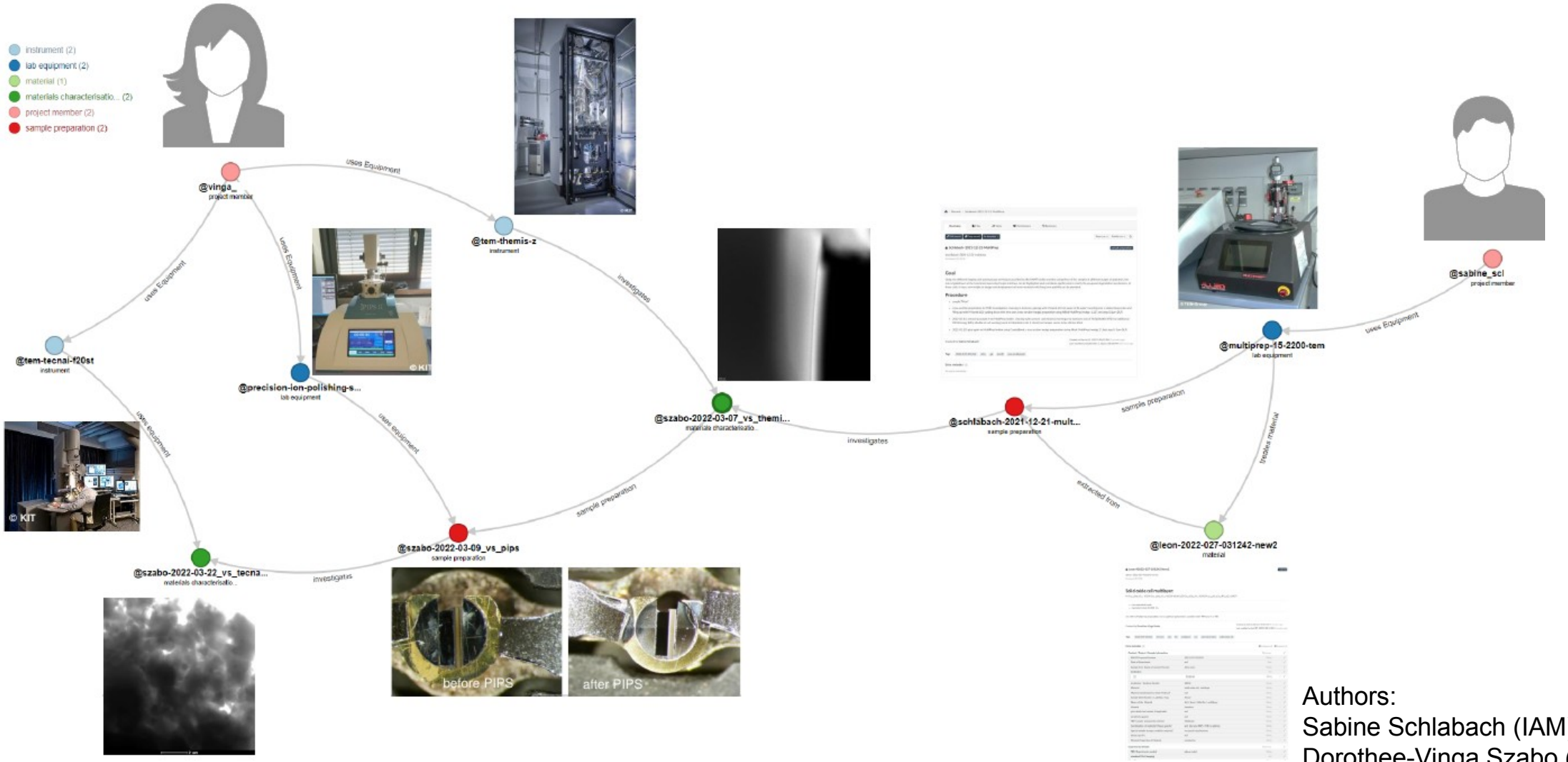
- Used instrument can be selected
- User input form is displayed from the metadata templates exported from kadi4mat web application.
- Customized templates in json format
- can be imported into the workflow
- and displayed



Collection created in Kadi4mat



Usage of Instrument descriptions in the research process



Authors:
Sabine Schlabach (IAM)
Dorothee-Vinga Szabo (IAM)

Programming Interface (API) of Kadi4Mat

- A REST-like **HTTP API** enables programmatic and automated usage of most features
- A **Python wrapper library** [1] exists on top of the API, which can also be used to integrate almost any tool via command line interfaces

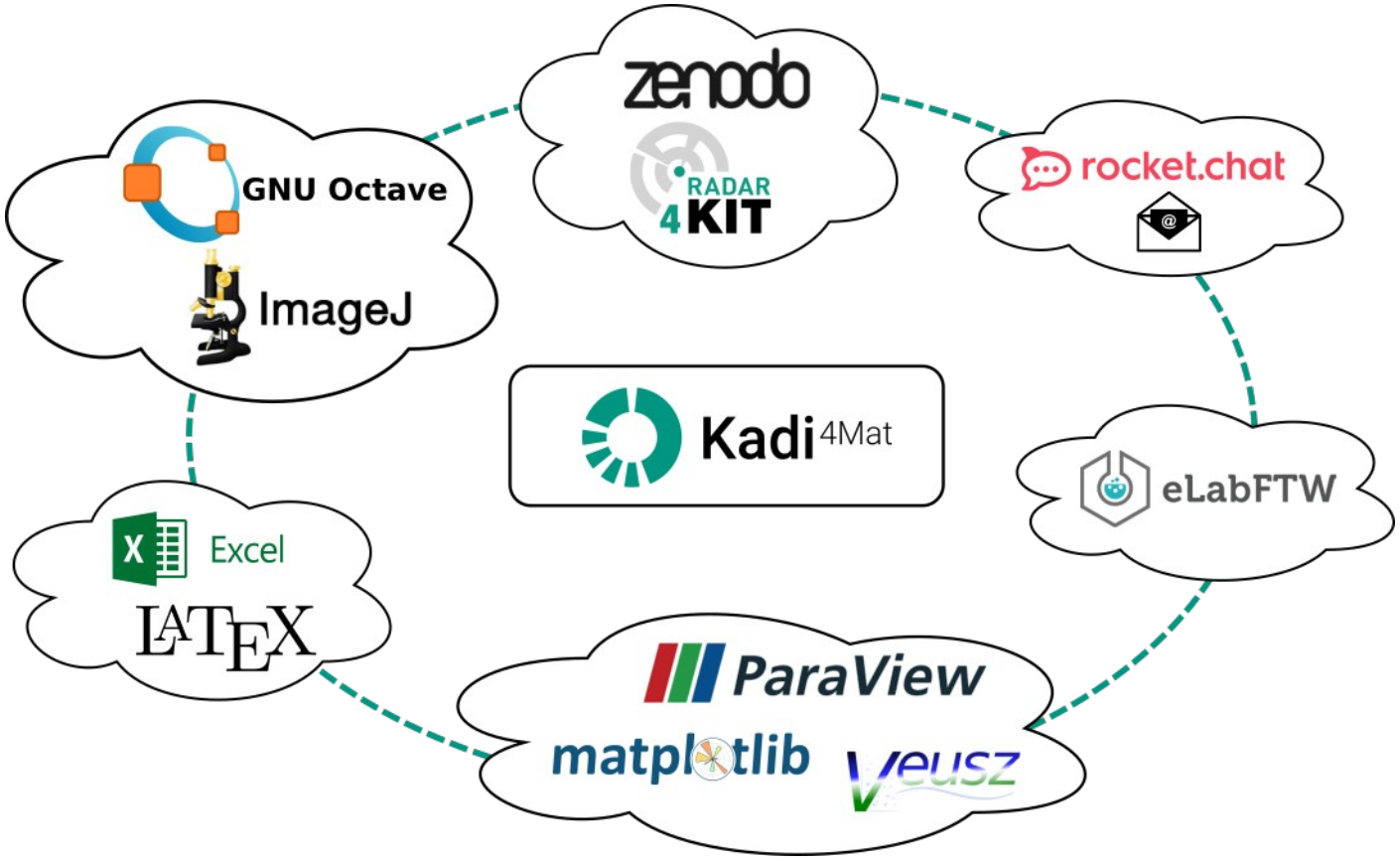
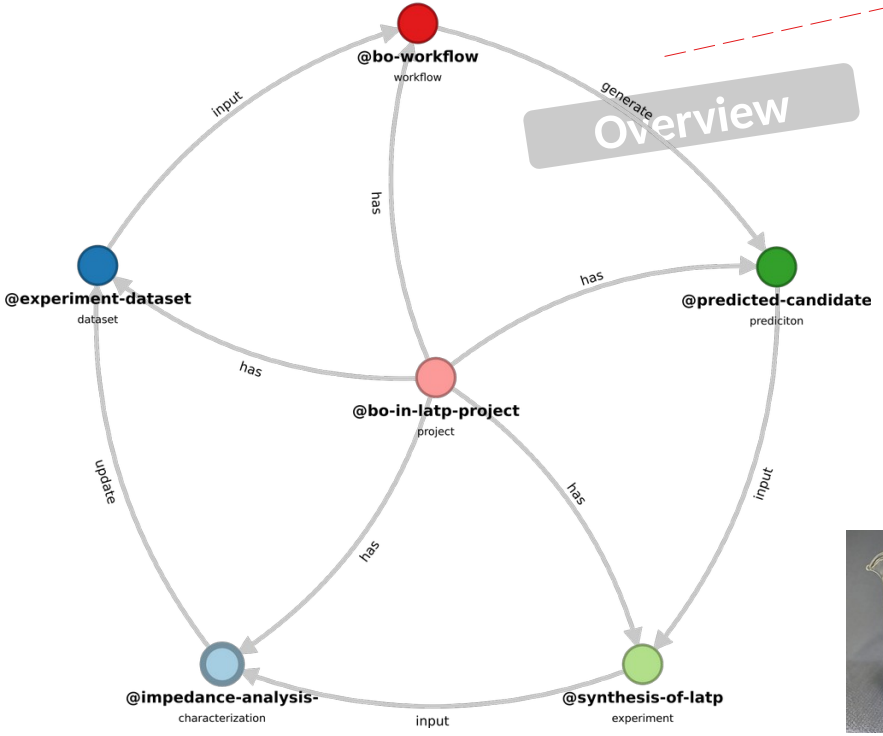


Image: Patrick Altschuh (KIT/HKA, Germany)

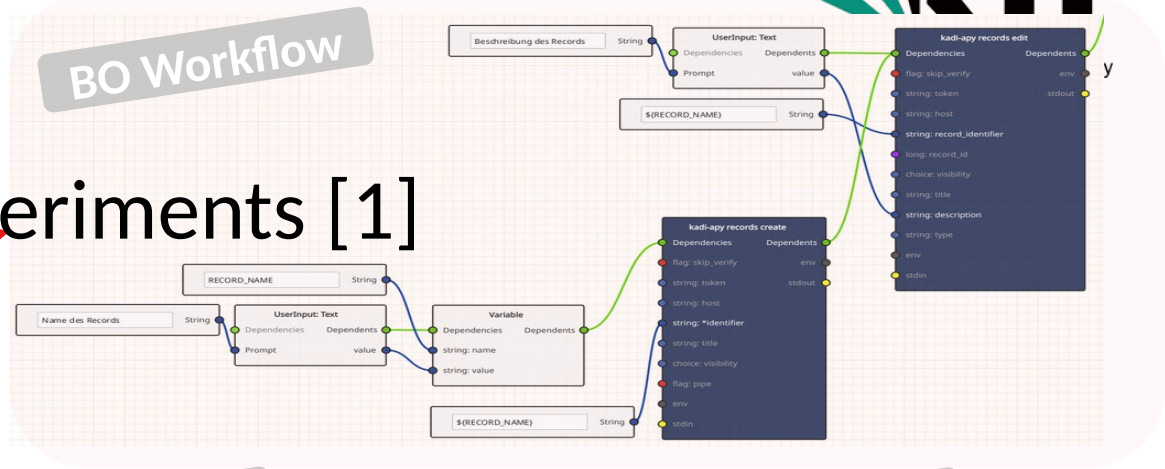
[1] https://gitlab.com/iam-cms/kadi-apy

Data Science: Application of Bayesian optimization methods

- Reducing required number of experiments [1]



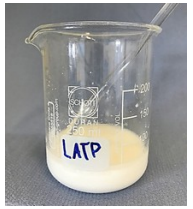
Overview



Input

Output

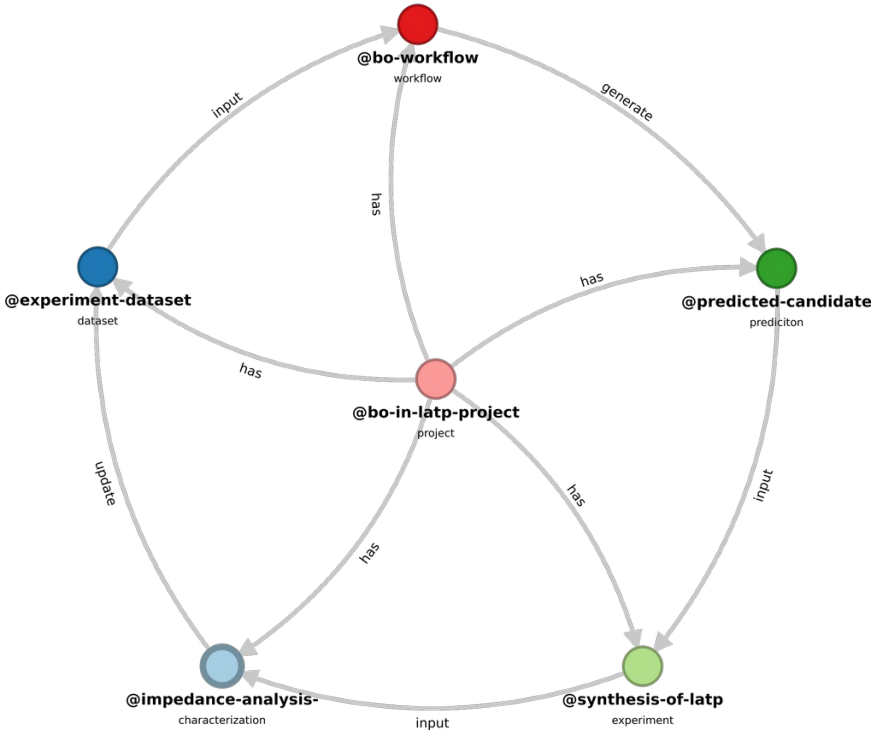
No.	Predicted experiment parameter: rel. H_3PO_4 wt%	Predicted experiment parameter: Sintering T °C	Predicted experiment parameter: Holding time min	Measured properties: Sintered density %	Measured properties: Ionic conductivity S/cm
1	-7.50	1000	40	87.99	5.67E-04
4	-7.50	1000	540	95.81	1.06E-03
13	-11.25	1000	450	97.52	1.09E-03
20	-15.00	1000	450	97.21	7.81E-04



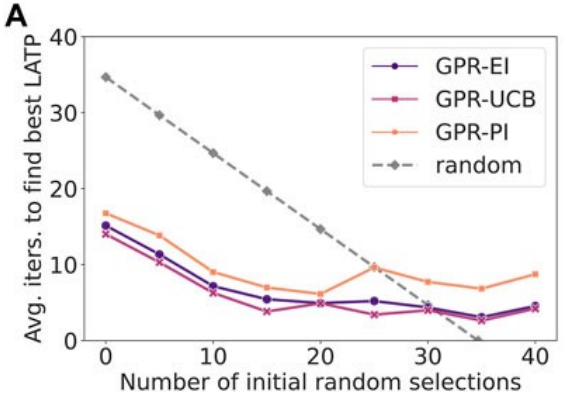
[1] Yinghan *et al.*, 2022, Frontiers in materials, w

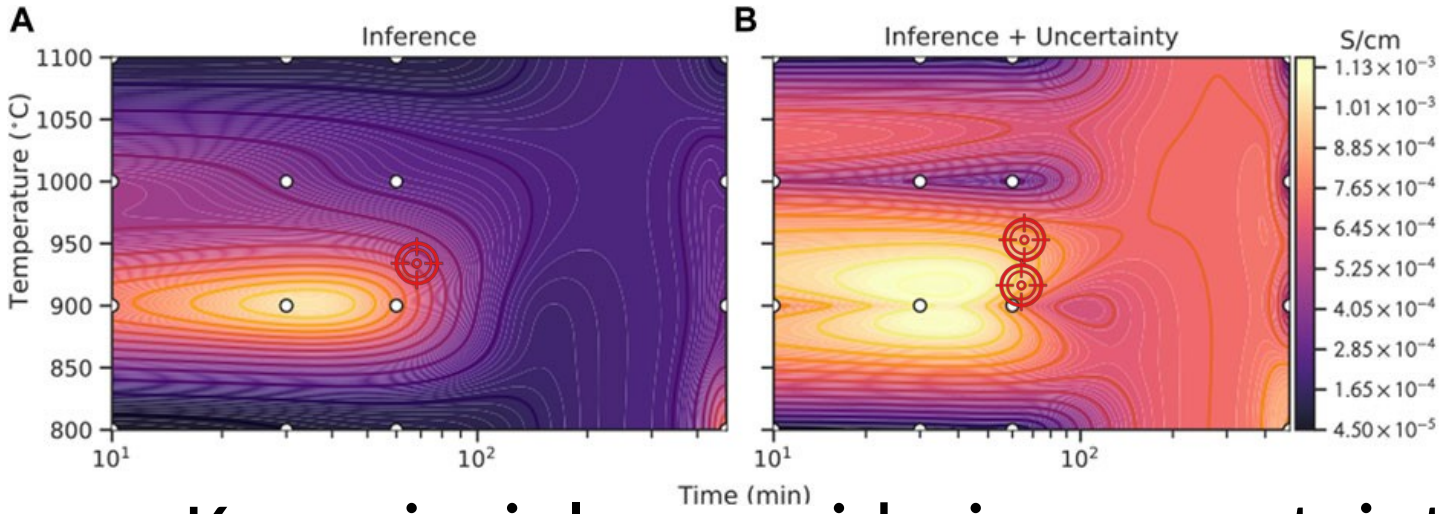
Data Science: Application of Bayesian optimization methods

- Reducing required number of experiments [1]



~ 1.5 – 2 times faster



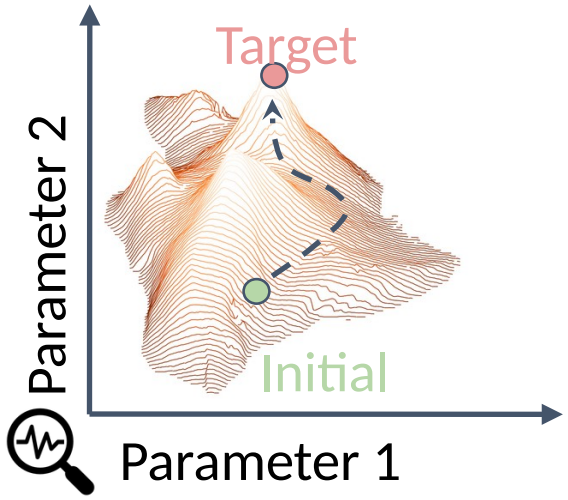


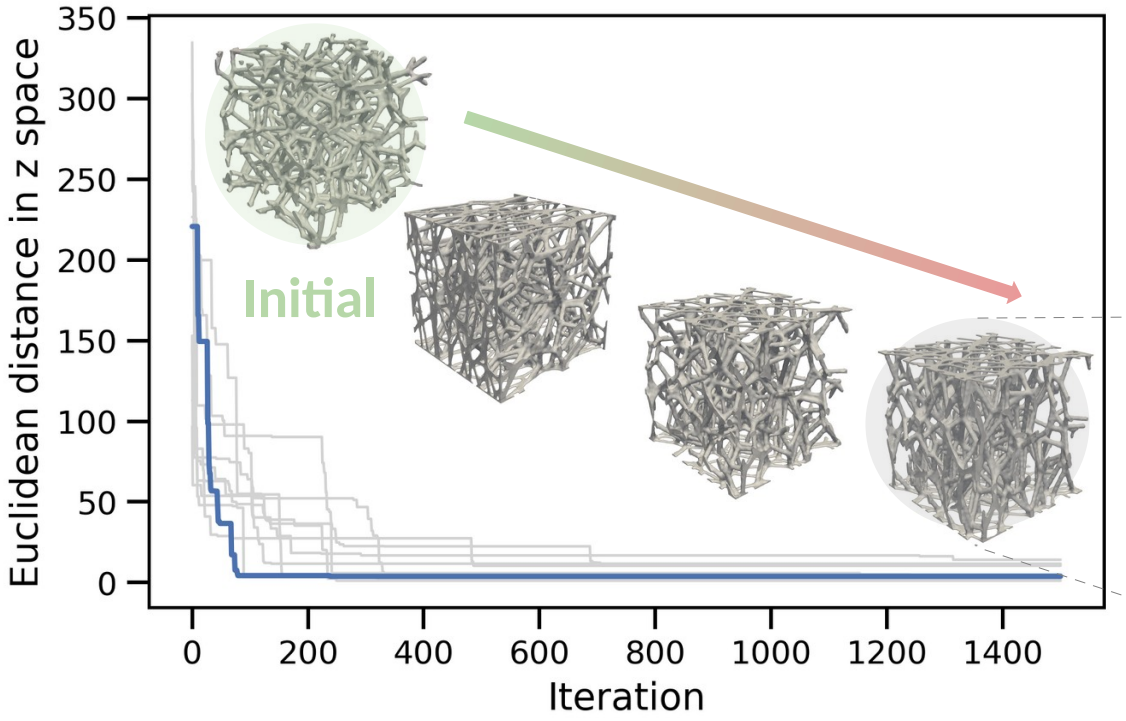
Key principle: considering uncertainty

[1] Yinghan *et al.*, 2022, Frontiers in materials, work with Nikolas Schiffmann *et al.*, KIT. BO applied in LATP synthesis

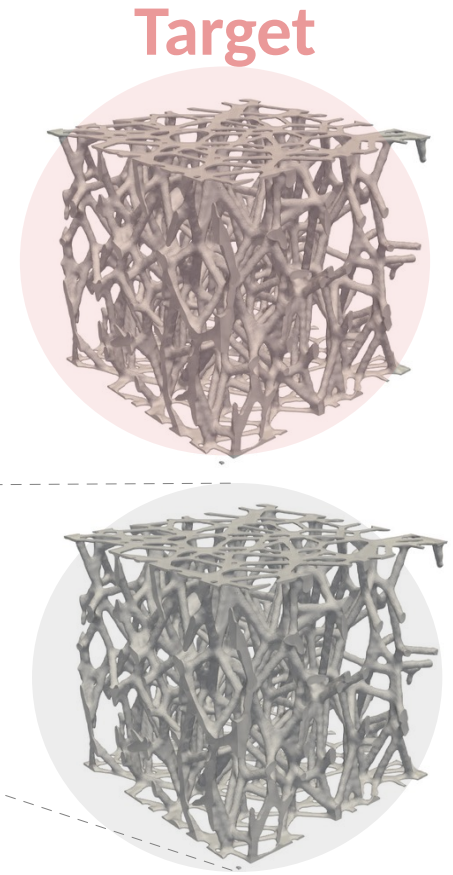
Data Science: Application of Bayesian optimization methods

- Optimizing/searching simulation parameters [1]





Rapid parameter search using BO



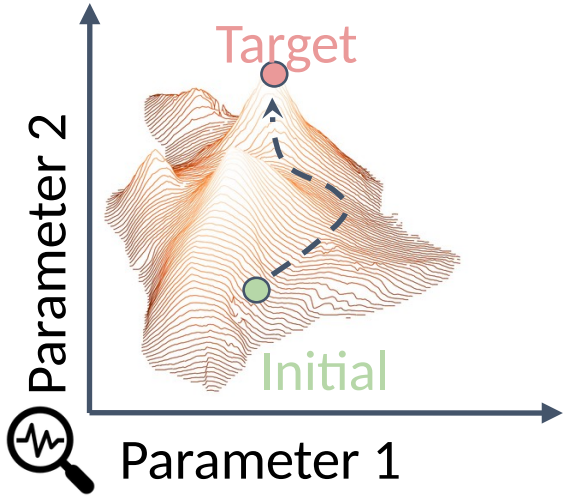
Best found

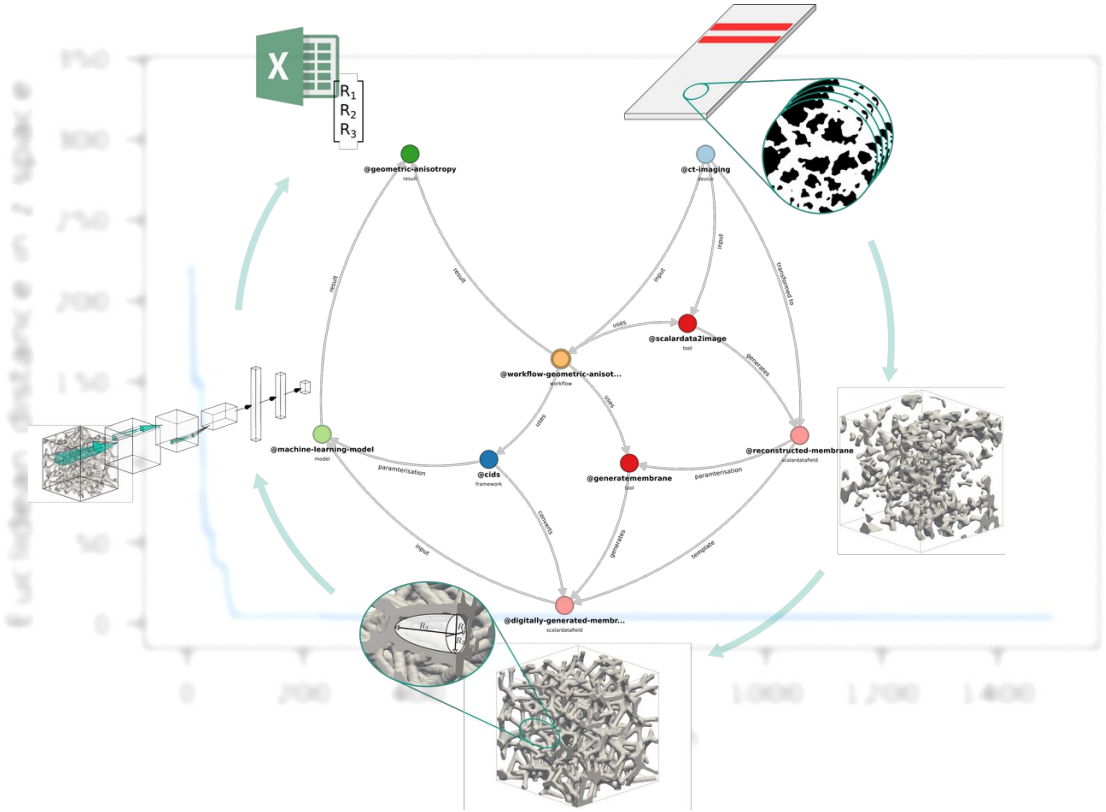
[1] work with Patrick Altschulh *et al.*, KIT. BO applied in simulated porous membranes.

Data Science: Application of Bayesian optimization methods

- Optimizing/searching simulation parameters [1]

Image:
Lars Griem (KIT)



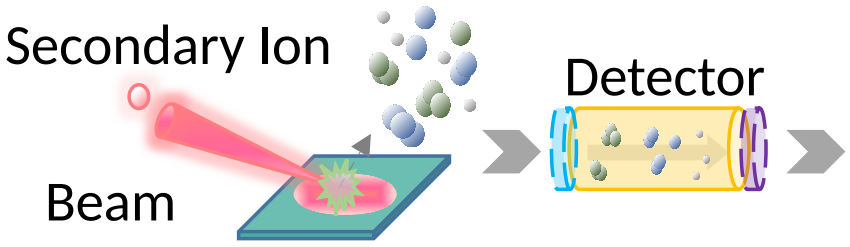


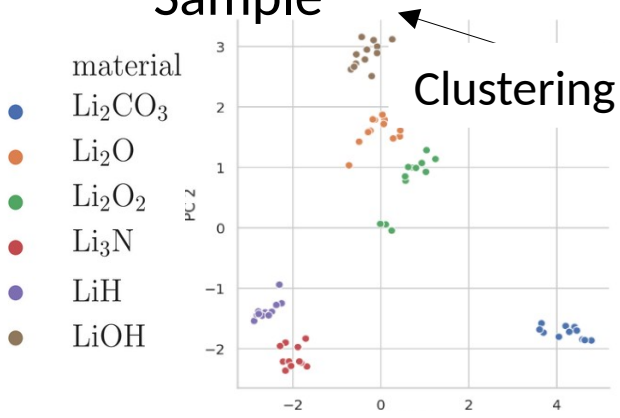
Microstructure-Property prediciton

[1] work with Patrick Altschulh *et al.*, KIT. BO applied in simulated porous membranes.

Data Science: Support in the analysis of complex characterization data

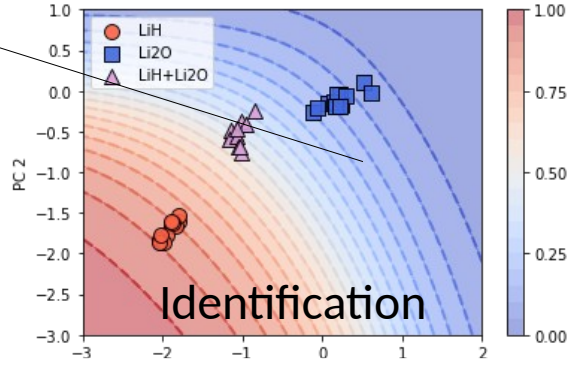
- Enhance data analysis and interpretation
- Collect data from open database and prepare models





PCA of 6 lithium-compounds [1]

Clustering

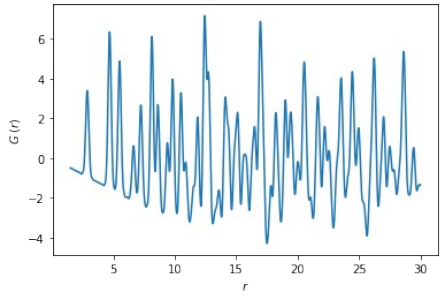


Analysis of mixtures

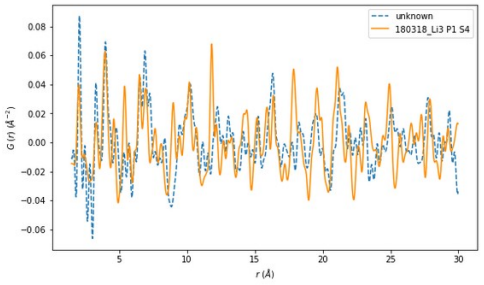
ICSD FIZ Karlsruhe

MATERIALS PROJECT

COD



Unknown spectrum



Best matching

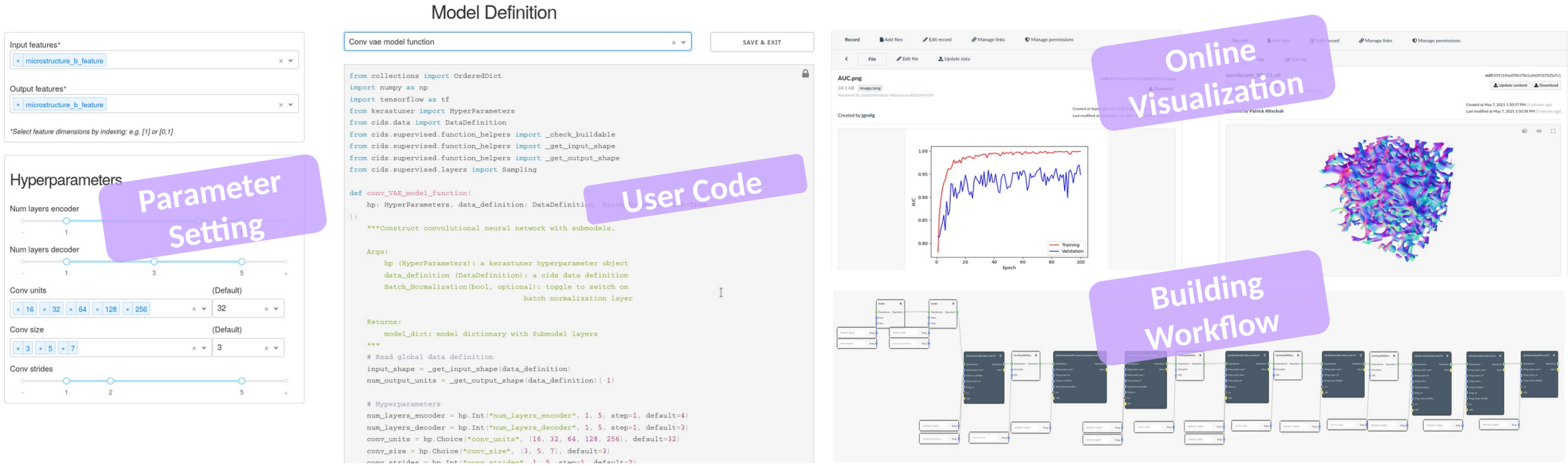
Search PDF candidates in database [2]

[1] Yinghan *et al.*, 2020, Procedia Computer Science, work with *S.-K. Otto et al.,* JLU. Extented draft in preparation
PCA: principal component analysis, used to extract most important information and reduce data dimension
[2] work with *A.-L. Hansen et al.,* KIT PDF: pair distribution function

Data Science

Exploration of more application scenarios & **integration of data science tools**

- Building appropriate models based on the amount/types of data
- Integrating data-science tools (CIDS framework [1]) and providing example use-cases



Interactive data-science tool integration, data visualization and workflow tools

[1] work with *Arnd Koeppe et al.*, KIT

Summary

- The future goal of Kadi4Mat is to support electronically recording the entire **scientific workflow** of daily research work **FAIRly** (*Findable*, *Accessible*, *Interoperable*, *Reusable*)
- Due to the heterogeneous nature of materials science, most features of Kadi4Mat are kept **generic** and are developed **bottom-up**

- All relevant information about Kadi4Mat can also be found at **https://kadi.iam.kit.edu**

Ackowledgments

- This work is funded by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK-BW), in the project MoMaF--Science Data Center (grant number: 34-7547.222).
- We would like to acknowledge the German Federal Ministry of Education and Research (BMBF) for its financial support within the project FB2 TheoDat, under the grant number 03XP0435D.
- This work is funded by the BMBF and MWK-BW as part of the Excellence Strategy of the German Federal and State Governments in the project Kadi4X.
- This work was partly carried out with the support of the Karlsruhe Nano Micro Facility (KNMFi, www.knmf.kit.edu), a Helmholtz Research Infrastructure at Karlsruhe Institute of Technology (KIT, www.kit.edu).
- Part of the model formulation was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), under project ID 390874152, POLiS Cluster of Excellence (grant number: UP 33/1). This work contributes to the research performed at CELEST (Center for Electrochemical Energy Storage Ulm-Karlsruhe).
- The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713."
- This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), under the project number SE 2842/3-1.