Combining high-performance computing, workflows and ML to get new insights in material sciences

NHR4CES Community Workshop 2025 – Michael Selzer, Giovanna Tosato

Overview

1. Introduction
2. Kadi4Mat
3. KadiAI and CIDS
   - Active Learning Optimization Framework
4. Tools and usecases:
   - Model Parametrization for Solid Oxide Fuel Cells
   - Outlook: Human-in-the-loop, automatized experiments
5. Conclusion

Introduction

1

Motivation

- Research data play an increasingly important role in almost all research disciplines, but their structured management is often neglected
  - **Research community:** Lack of reproducibility of research results, no data exchange, data loss (→ duplication of work)
  - **Data science:** Not only require large amounts of data, but the data must be of high quality and machine operable

→ Funding institutions also require a **FAIR research data management**

DFG Deutsche Forschungsgemeinschaft
https://www.dfg.de

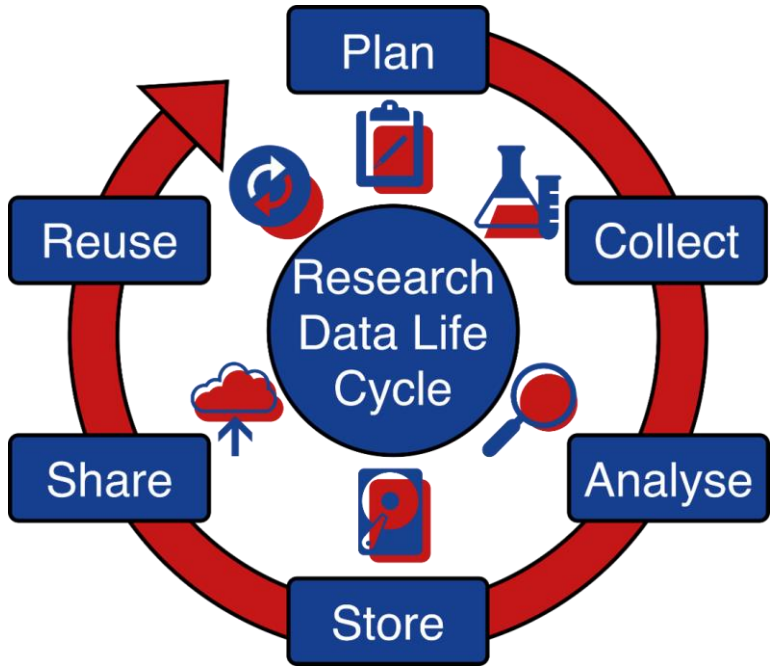
Bundesministerium für Bildung und Forschung
https://www.bmbf.de

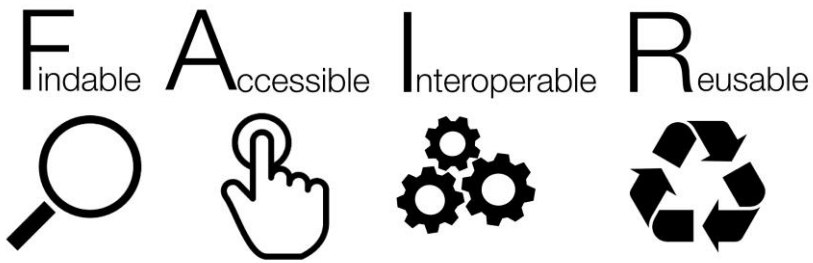
European Commission
https://commission.europa.eu

Research Data Management (RDM)

- A structured RDM requires flexible solutions in order to accomodate user- and discipline-specific applications

- **Goals**
  - Consideration of all phases of the research data life cycle, where possible
  - Support of "step-by-step" FAIRness of data at different levels
  - Generic functionalities and integration of existing standards and systems → complementing instead of replacing





https://de.wikipedia.org/wiki/FAIR-Prinzipien#/media/Datei:FAIR_data_principles.jpg

Kadi4Mat

2

Kadi4Mat

- Kadi4Mat is the open source **research data infrastructure** that is being actively developed in the context of various research projects
- Originally developed in the context of materials science, it can be used for the management of any type of research data or process within any research discipline
- Multiple instances of Kadi4Mat are hosted at KIT [1] and other institutes, including a publicly available instance at **https://demo-kadi4mat.iam.kit.edu/***

Karlsruhe Data Infrastructure
for Materials Science

[1] https://kadi.iam.kit.edu/#instances

Conceptual Overview of Kadi4Mat

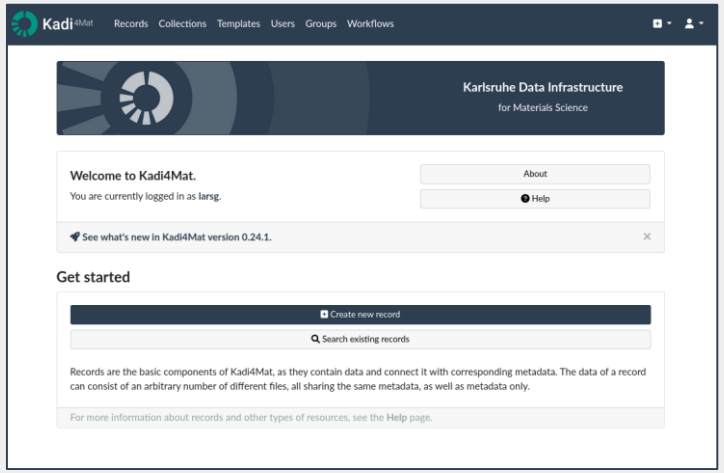
- The concept of **Kadi4Mat** is a modular and generic architecture that combines and interfaces with functionalities of **ELNs** and **repositories**

- The ELN part can be described as **ELN 2.0**, with focus on automation via a web-based API and data provenance (data origin), besides typical ELN functionality

- The repository part can be described as **community repository**, with focus on managing and sharing *warm data* (data yet to be analysed/processed further) within a community

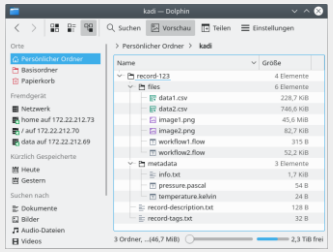
The Kadi Ecosystem

KadiWeb Repository / Web



KadiFS

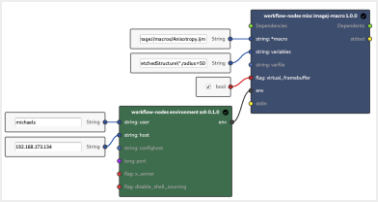
Filesystem



Streamline your research

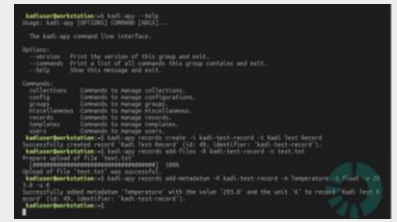
KadiStudio

Workflows



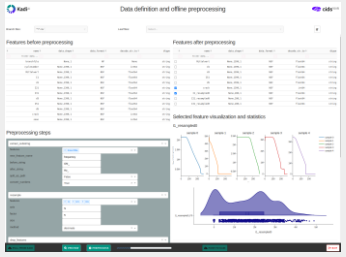
KadiAPY

(Python) API



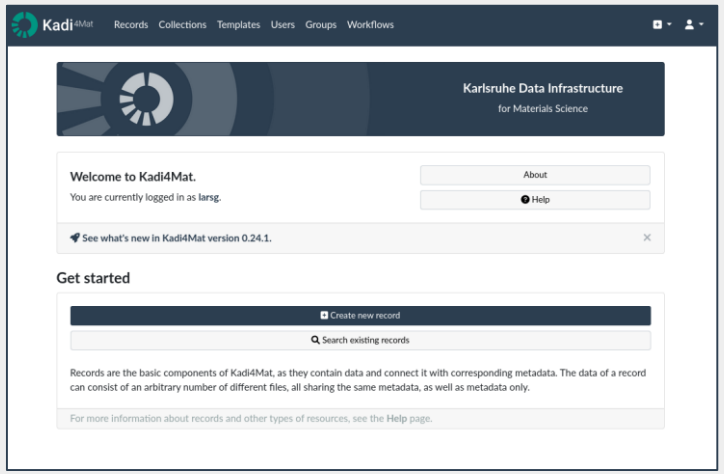
KadiAI

ML Interface



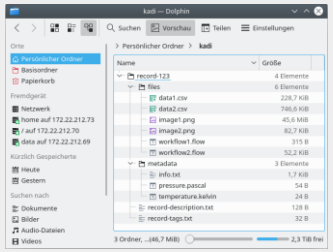
The Kadi Ecosystem

KadiWeb Repository / Web



KadiFS

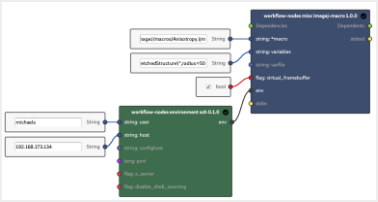
Filesystem



Streamline your research

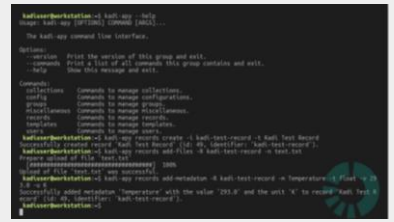
KadiStudio

Workflows



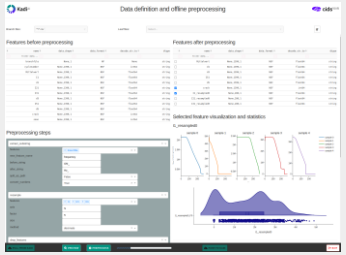
KadiAPY

(Python) API



KadiAI

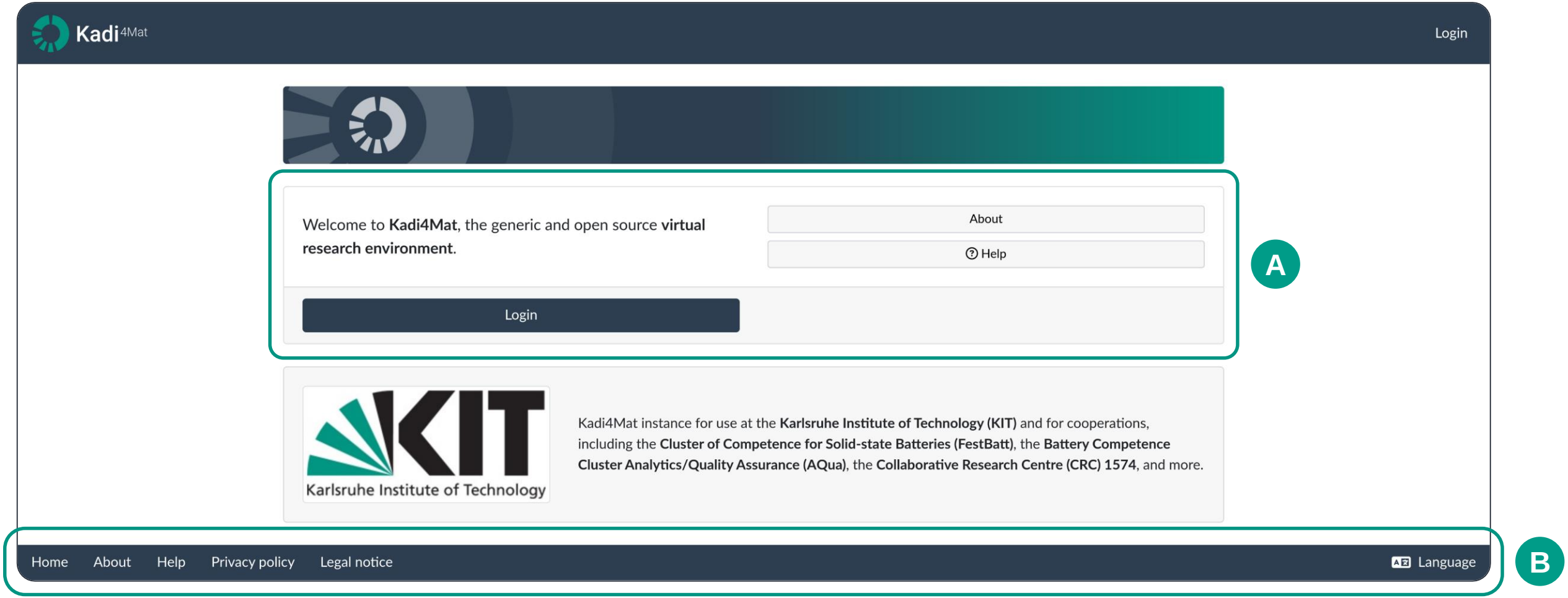
ML Interface



The Graphical User Interface of Kadi4Mat

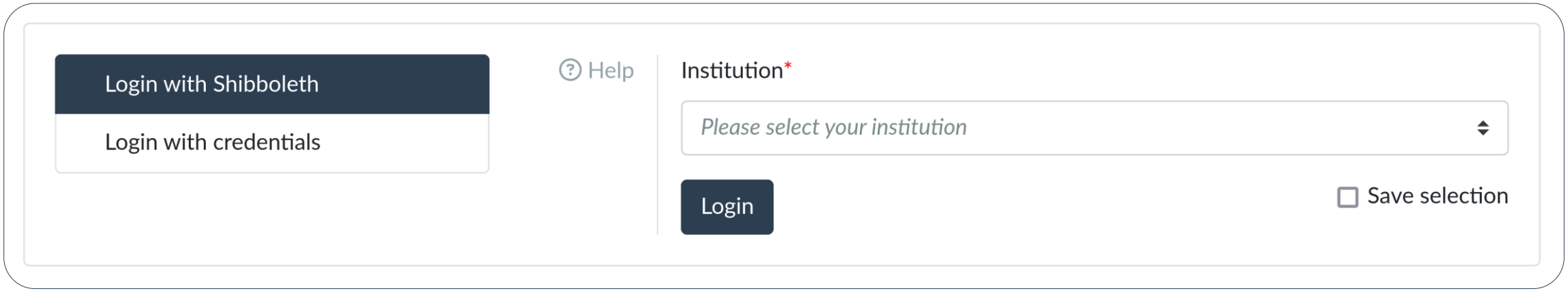
- The graphical user interface (GUI) is one of the main ways to use Kadi4Mat, and one of the easiest to get started with
- It can be used with all up-to-date web browsers and requires no additional software
- The REST-like HTTP API (Application Programming Interface) of Kadi4Mat exists next to the graphical user interface and enables programmatic usage of most functionality to facilitate automation and integration with existing programs and software

Kadi4Mat – Index Page



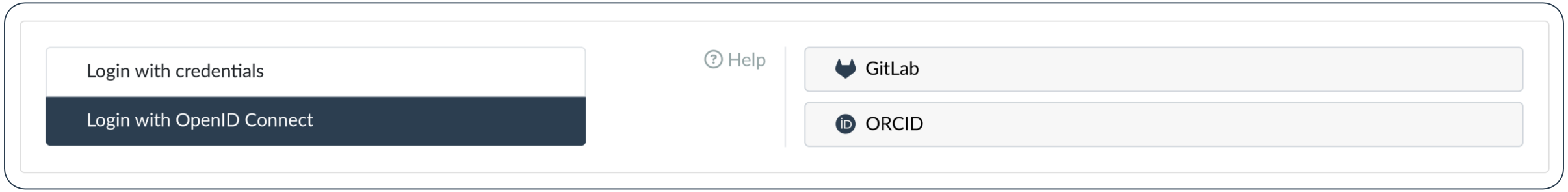
(A) Informational links and login **(B)** Navigation footer and language selection

Kadi4Mat – Login



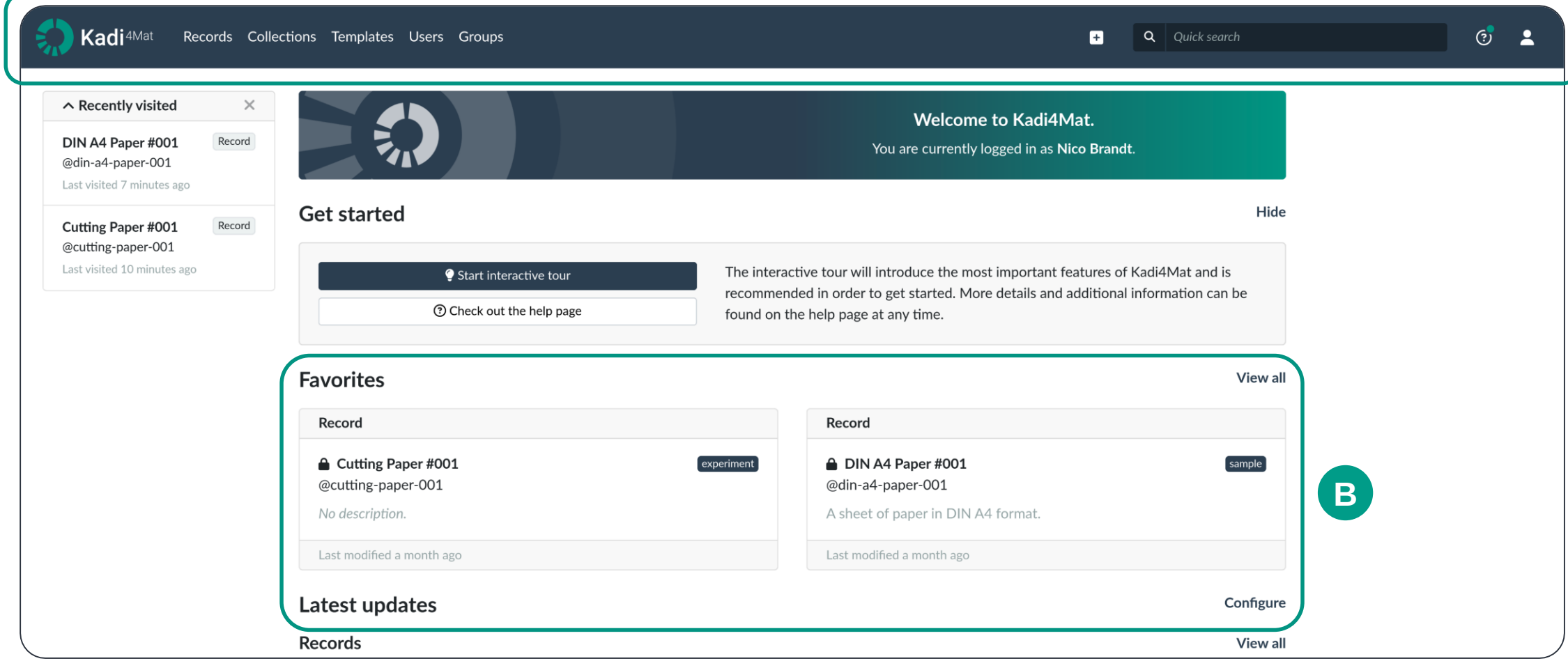
- In our main Kadi4Mat instance, login is usually done using institutional accounts via **Shibboleth**
- If necessary (e.g. due to technical limitations), login via local accounts is also possible (accounts can be created manually by us on demand)
- Login scheme can be defined per instance

Kadi4Mat – OpenID Connect Integration



- Other instances may have different authentication providers, such as our public demo instance at https://demo-kadi4mat.iam.kit.edu, which also supports OpenID Connect and free registration of local accounts

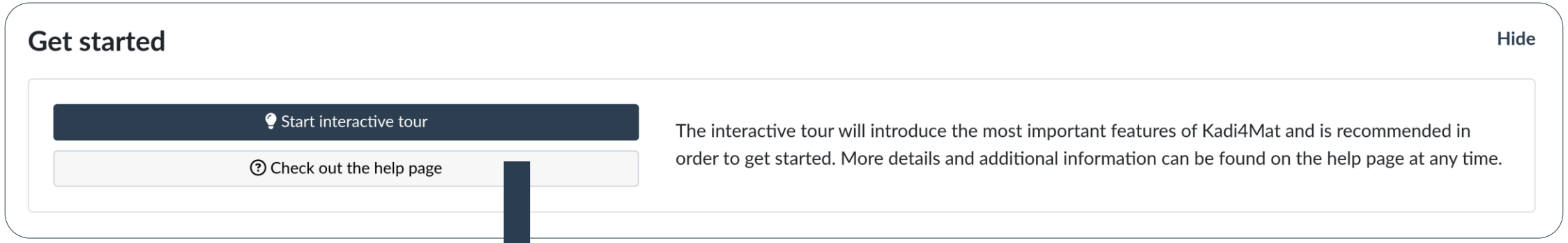
Kadi4Mat – Home Page



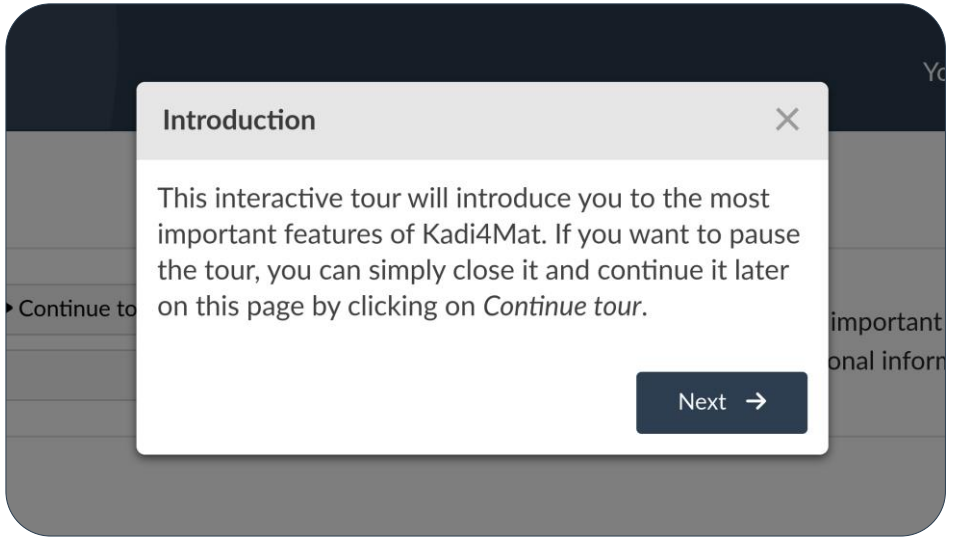
(A) Navigation bar with links to the different resources and various menus/quick actions

(B) Favorited resources and latest updates (customizable)

Kadi4Mat – Interactive Tour



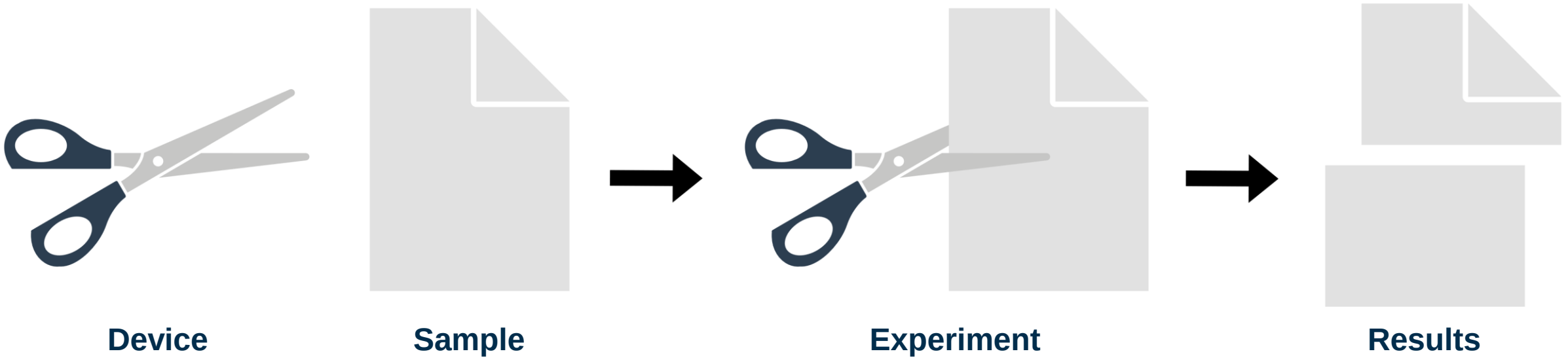
- In addition to the help page, the interactive tour introduces some of the basic functionality of Kadi4Mat, with a focus on creating and managing records



Kadi4Mat – Resources Overview

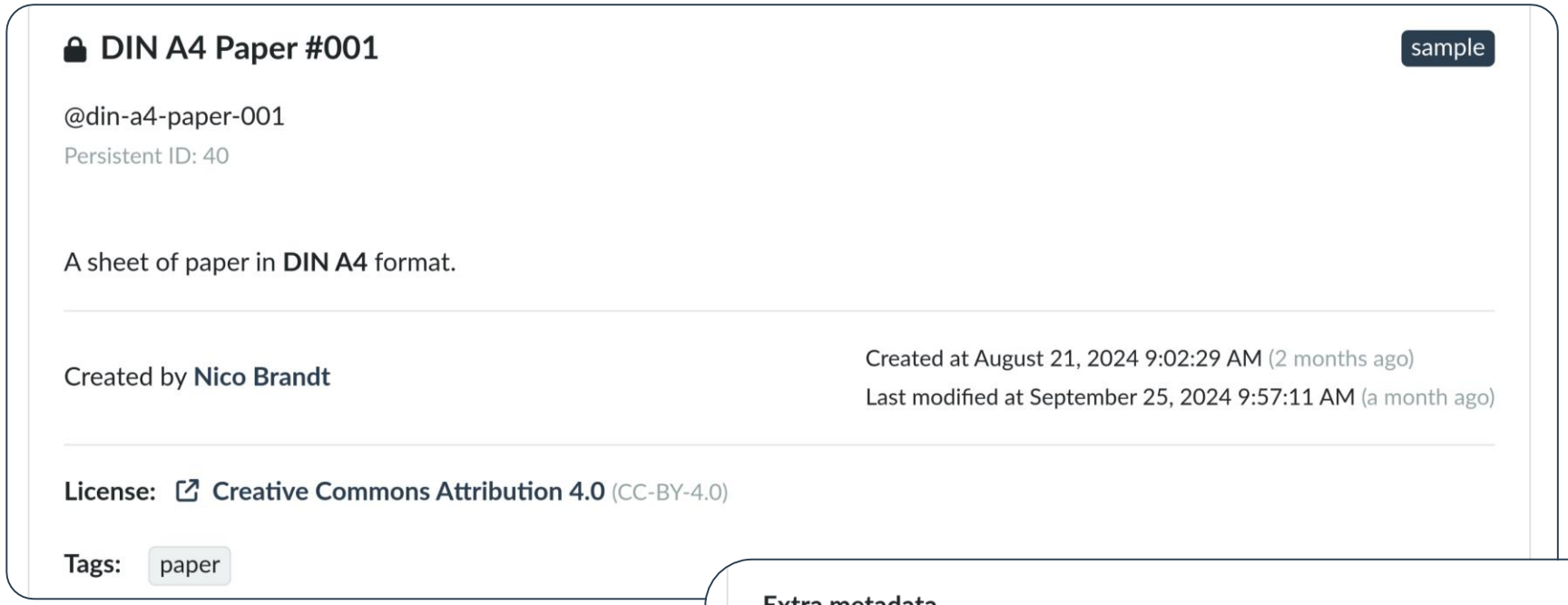
- **Records:** Basic components of Kadi4Mat, can represent any kind of digital or digitalized object via suitable metadata and/or corresponding data and can be linked together
- **Collections:** Represent logical groupings of multiple records (e.g. a single experiment) or other collections (e.g. a project with multiple experiment child-collections)
- **Templates:** Provide blueprints for either whole records, including record links and permissions, or just their generic metadata
- **Groups:** Group multiple users to facilitate permission management for different resources

Example Experiment Overview



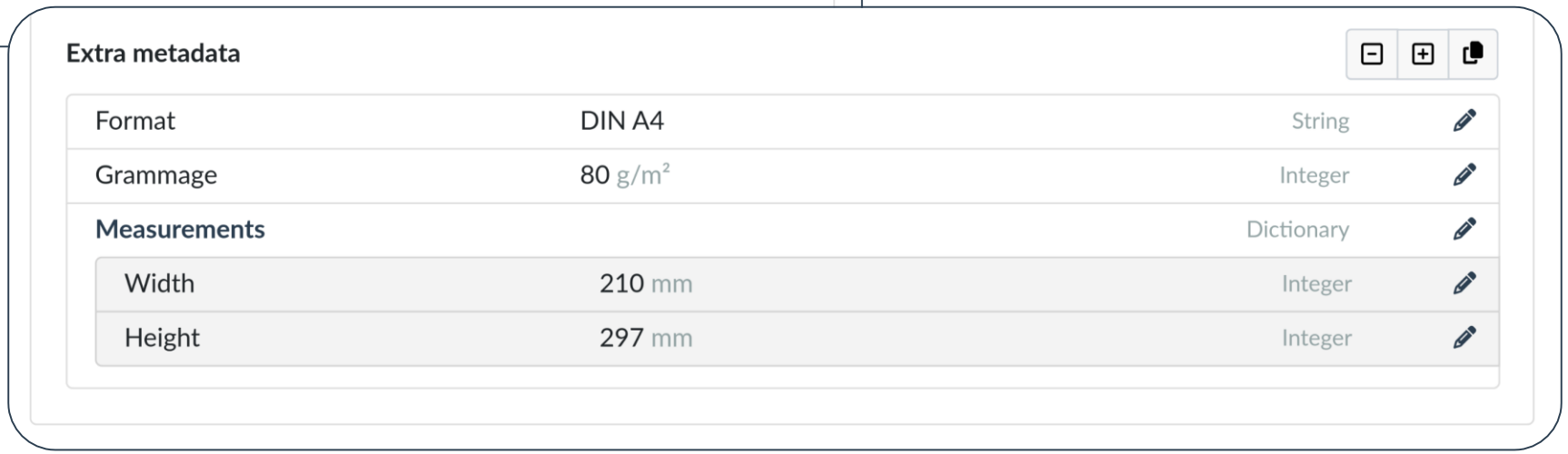
→ All **inputs**, **processes** and **outputs** involved in the execution of the experiment can be represented as (metadata-only) records within Kadi4Mat (and optionally be grouped into a collection)

Example Experiment - Record of a Sample

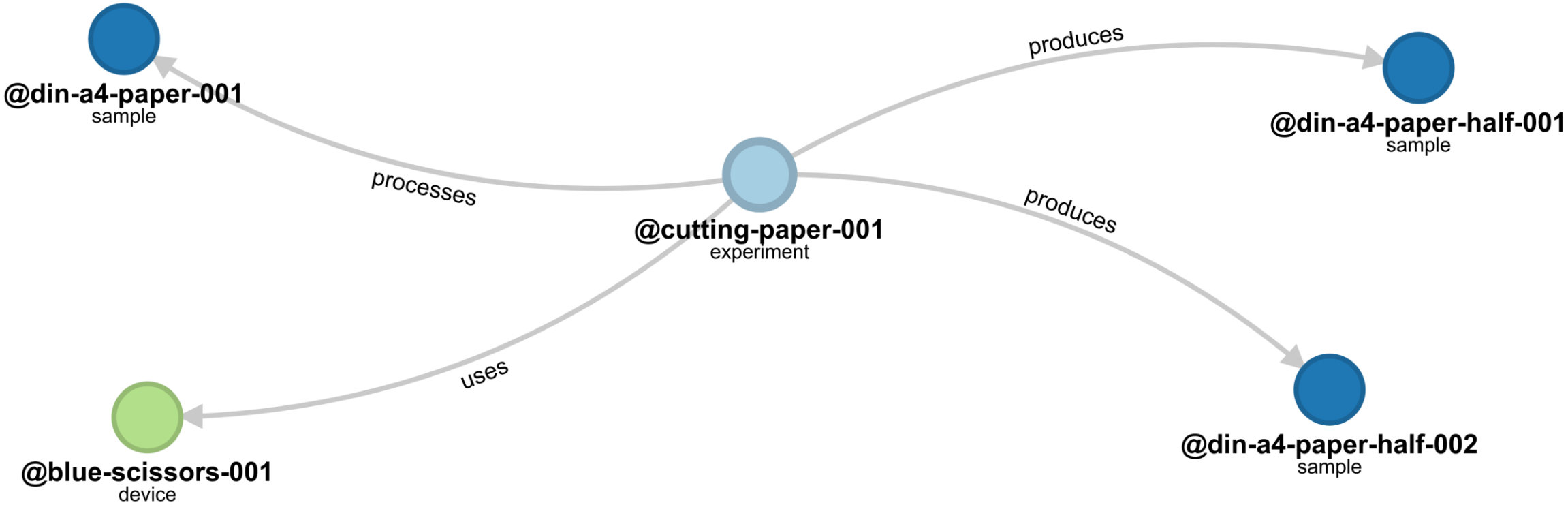


Basic Metadata
Fixed schema

Generic Metadata
No (fixed) schema



Example Experiment - Record Links and Data Provenance

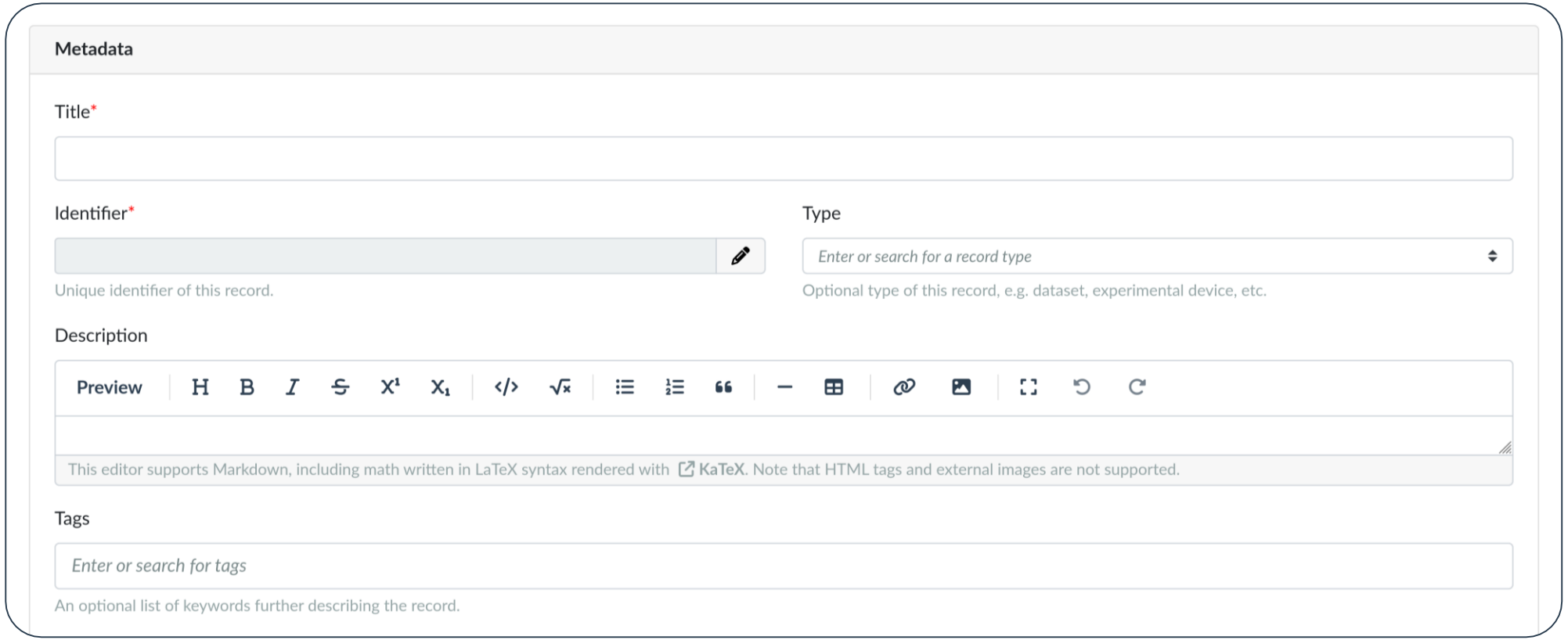


- Record links can be seen as separate resources, with their own metadata (e.g. their name)
- They specify the **data provenance**, which creates a simplified version of a **knowledge graph**

Kadi4Mat – Creating New Record

Basic metadata

- Title
- (Unique) Identifier
- Type
- (Markdown) Description
- Tags
- License
- Visibility

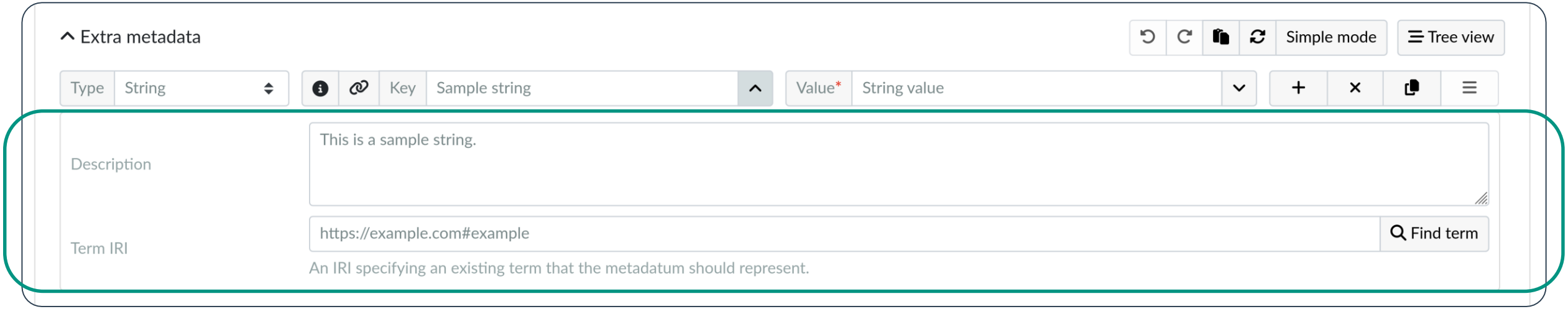


Kadi4Mat – Creating New Record

Generic metadata

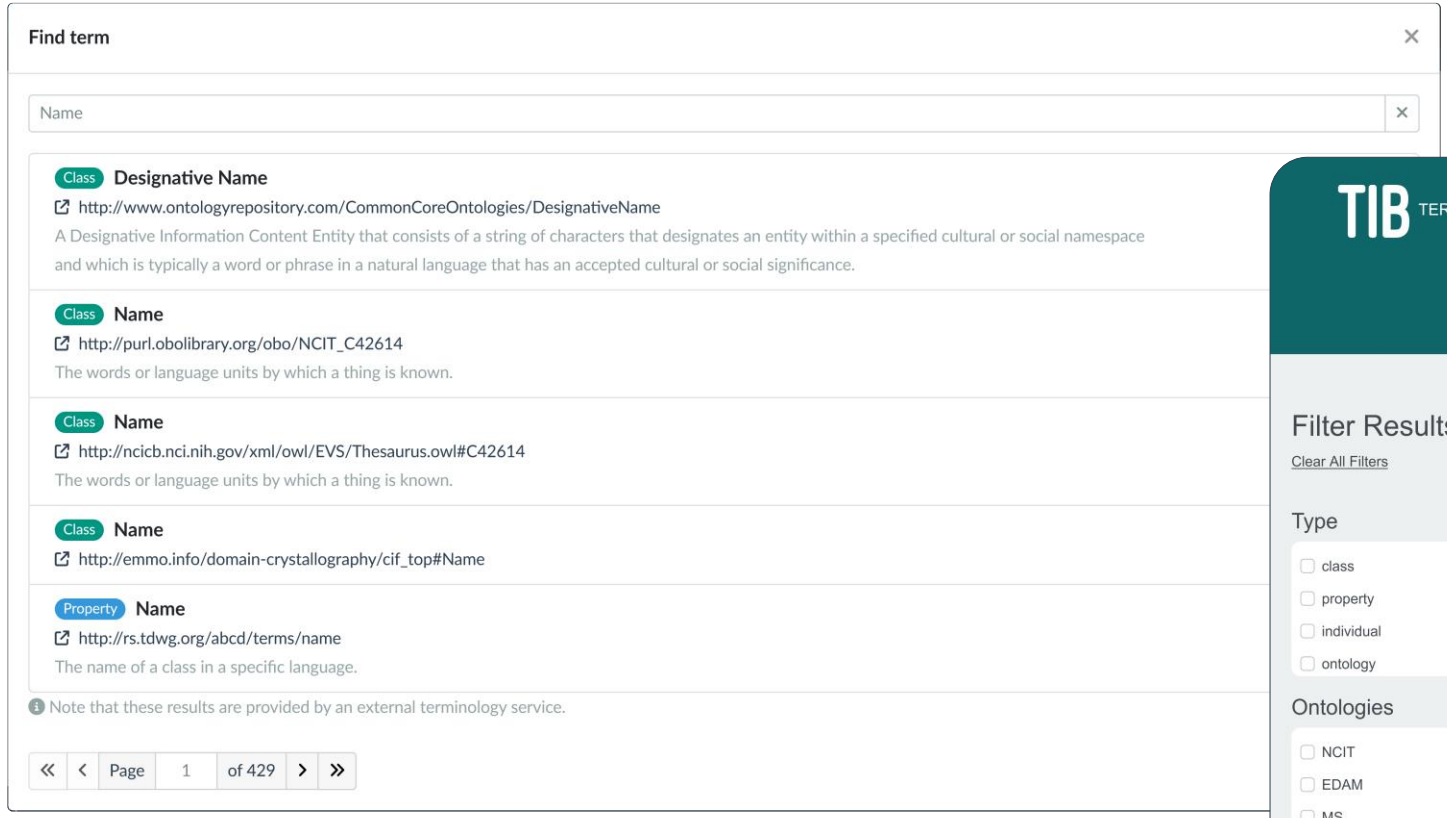
- Enhanced and (potentially) nested key/value-pairs of different types
- Supports units for numeric values, basic validation instructions (depending on the type), custom descriptions and the specification of standardized metadata terms
- The corresponding editor has different modes and views to facilitate easier metadata entry and supports loading metadata from previously defined templates

Kadi4Mat – Generic Metadata Editor

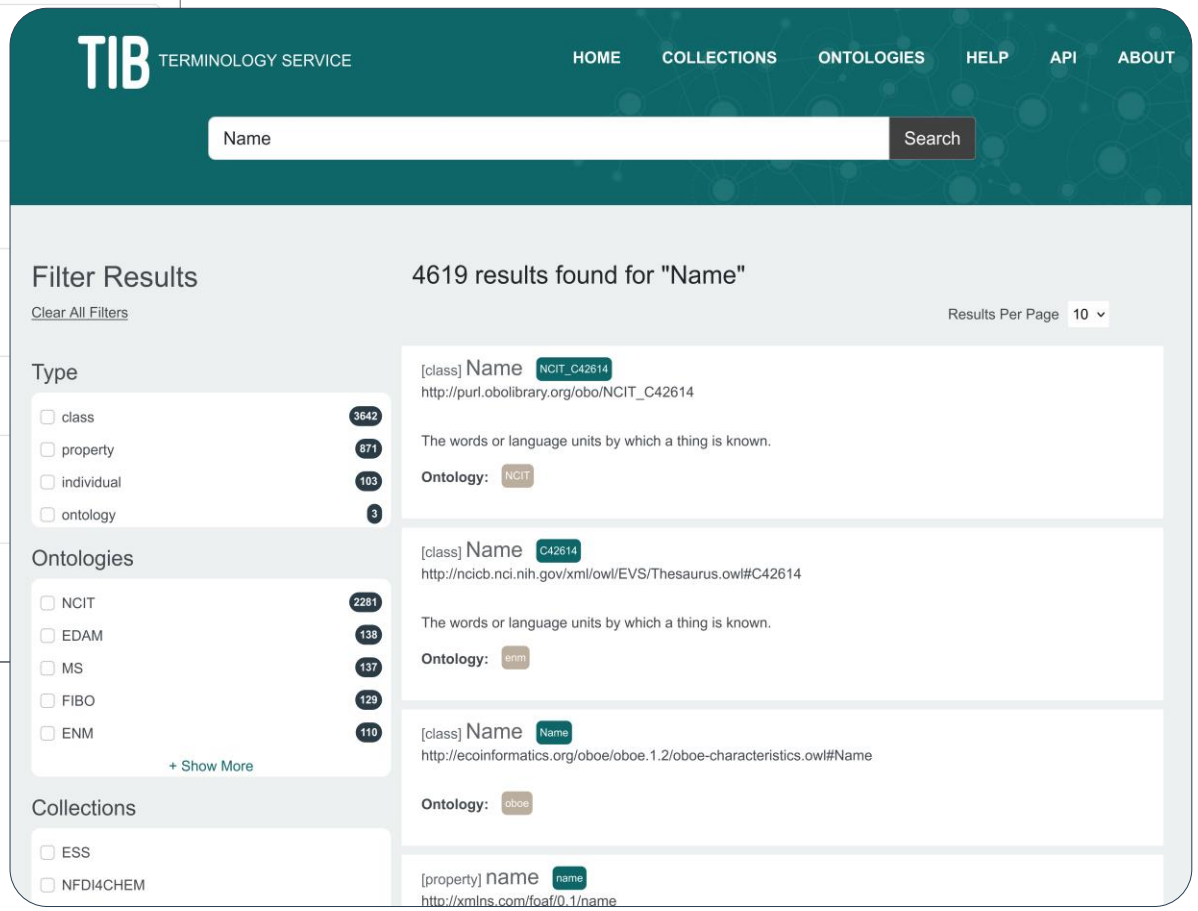


- While descriptions can provide a user-readable description of a metadatum, metadata terms are focused on machine readability and interoperability
- They enable the specification of identifiers (via Internationalized Resource Identifiers = IRIs), representing terms of existing metadata schemas/standards, vocabularies or ontologies

Kadi4Mat – Generic Metadata Term Search

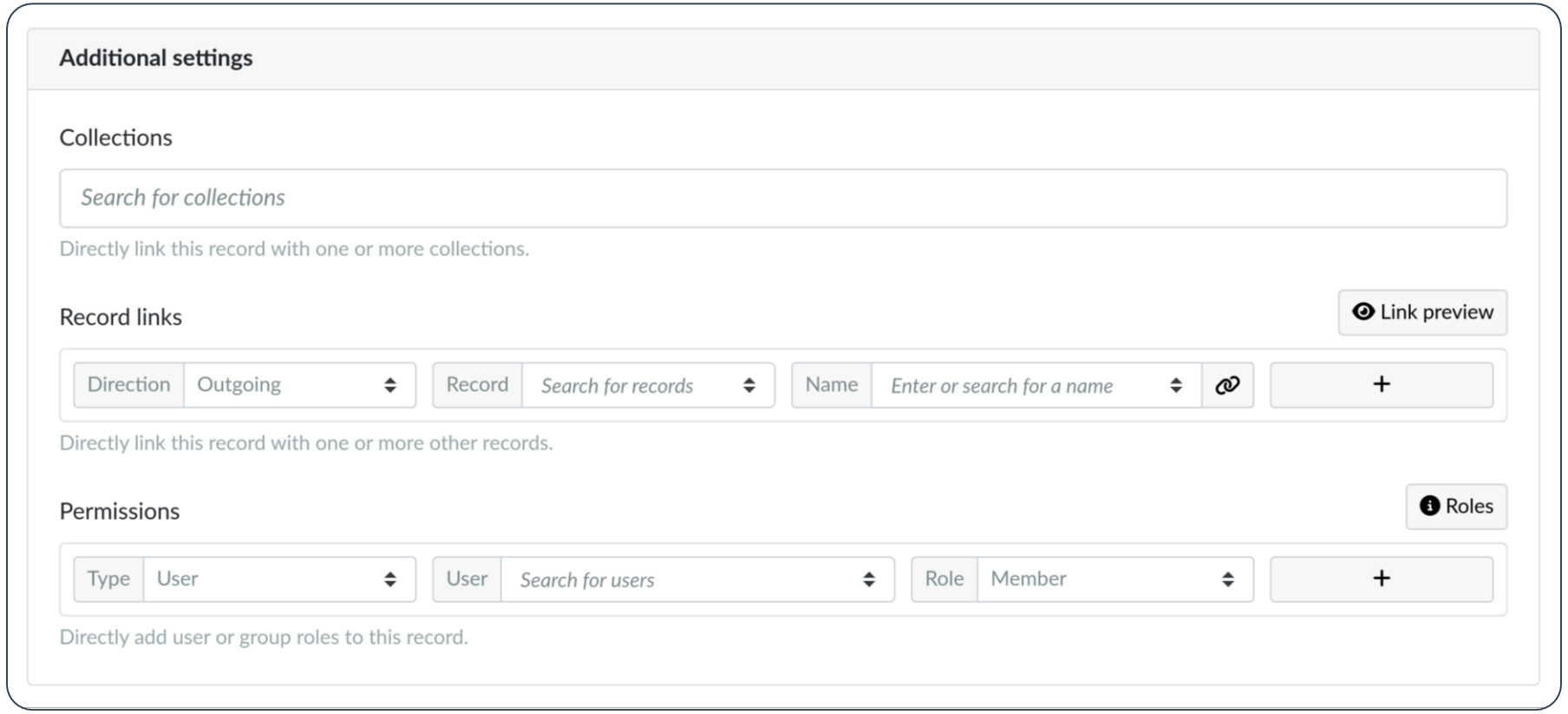


https://terminology.tib.eu/ts



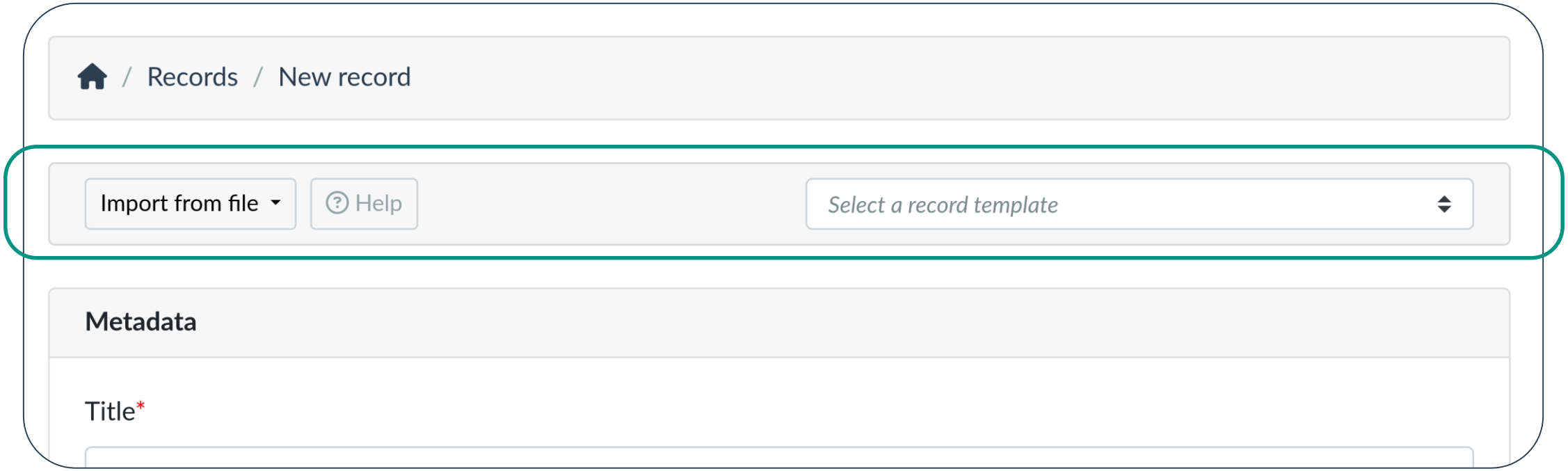
Terms can also be searched for using existing terminology services

Kadi4Mat – Creating New Record



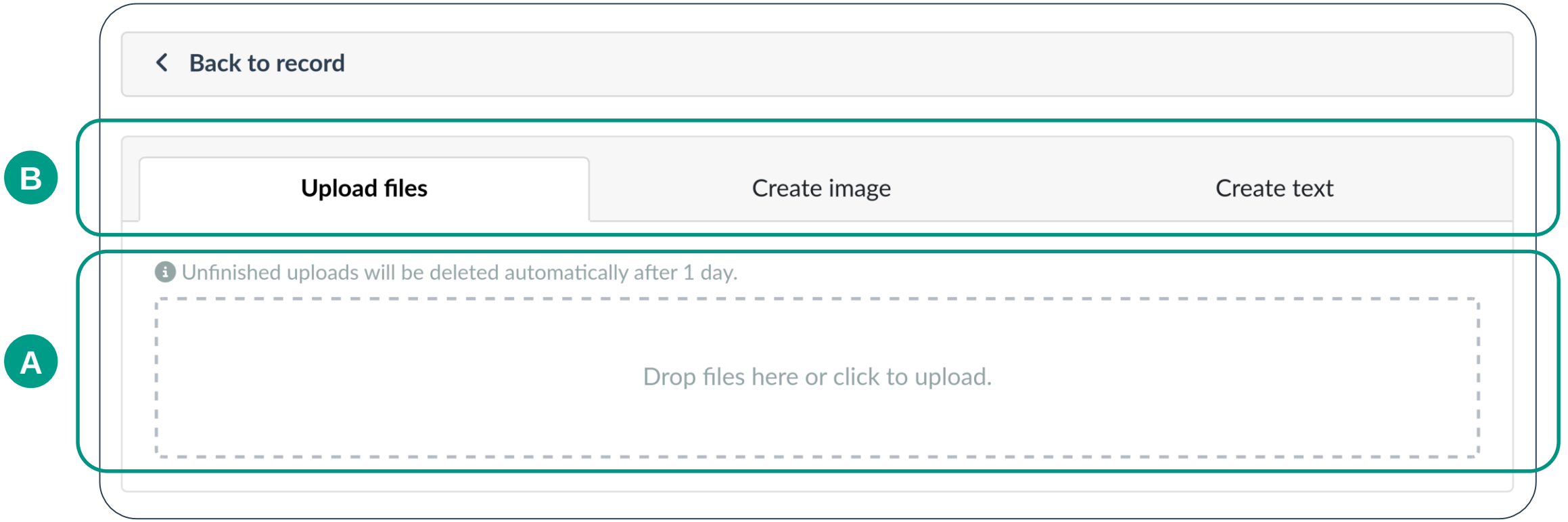
- In addition to the metadata of the record itself (describing its content), additional settings can be specified while creating a record
- These include collections, record links and access permissions for users or groups

Kadi4Mat – File Import and Templates



- When creating records, metadata and additional settings can also be imported from files (currently, only JSON files are supported within the GUI)
- In addition, previously defined (record) templates can be selected

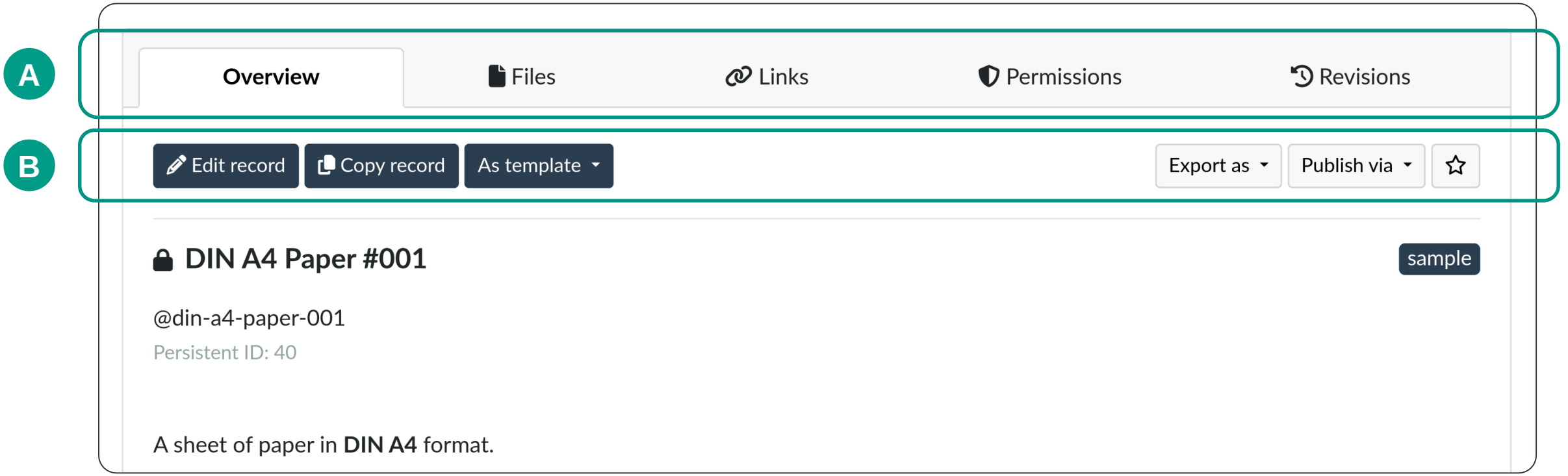
Kadi4Mat – Adding Files to a Record



(A) Regular upload of local files

(B) Some file types can also be created (and edited) directly in the web interface

Kadi4Mat – Record Overview



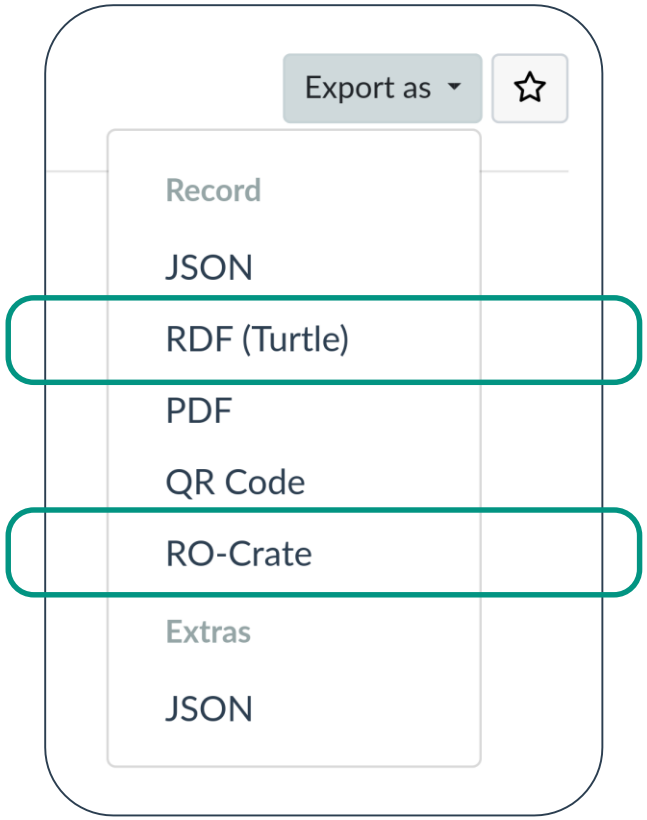
(A) Overviews of different contents of the record

(B) Additional (general and record-specific) actions, including exporting and publishing of records

Kadi4Mat – Export & Interoperability

- All relevant resource types in Kadi4Mat support exporting their (meta)data in different formats
- Exported data can be used for different purposes, including interoperability with other systems

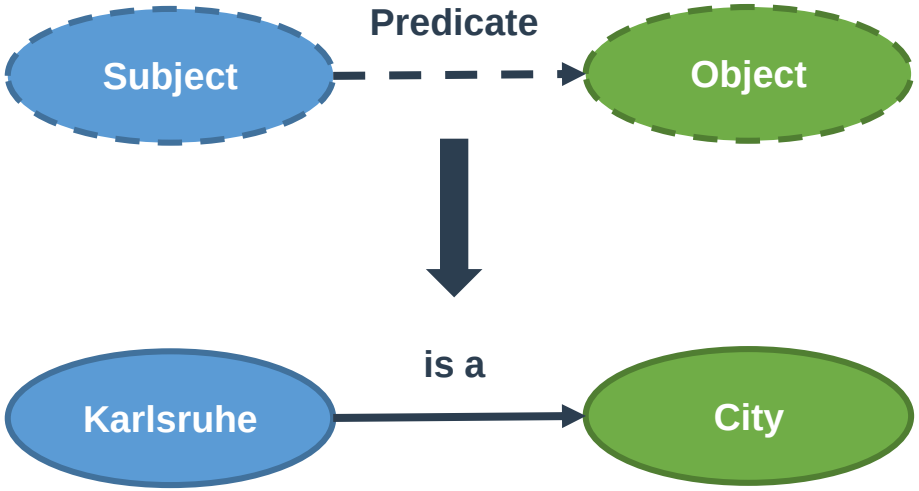
Record export



Template export

Kadi4Mat – RDF Export

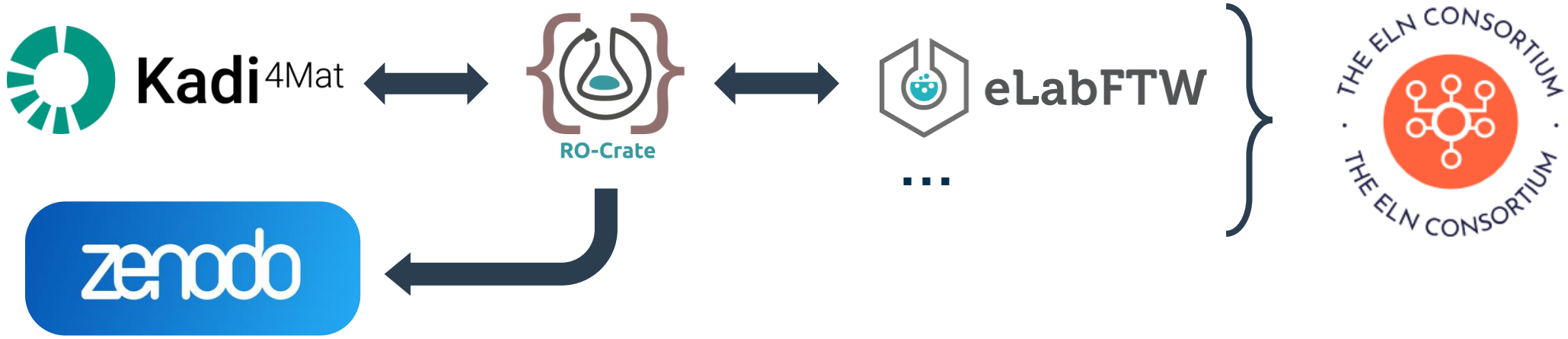
- Reminder: RDF (Resource Description Framework) is a standard for describing metadata in a graph and is the base of many ontology modelling languages
- Each graph is composed of triples specifying a **subject**, **predicate** and **object**
- Subjects, predicates and (optionally) objects make use of IRIs to describe the metadata in an interoperable manner (similar to the metadata term functionality from earlier)




```
<http://localhost:5000/records/40> a schema:Dataset ;
    rdfs:isDefinedBy [ k4m40:Format "DIN A4" ;
            k4m40:Grammage [ a schema:QuantitativeValue ;
                    schema:unitText "g/m²" ;
                    schema:value 80 ] ;
            k4m40:Measurements [ k4m40:Height [ a schema:QuantitativeValue ;
                                schema:unitText "mm" ;
```

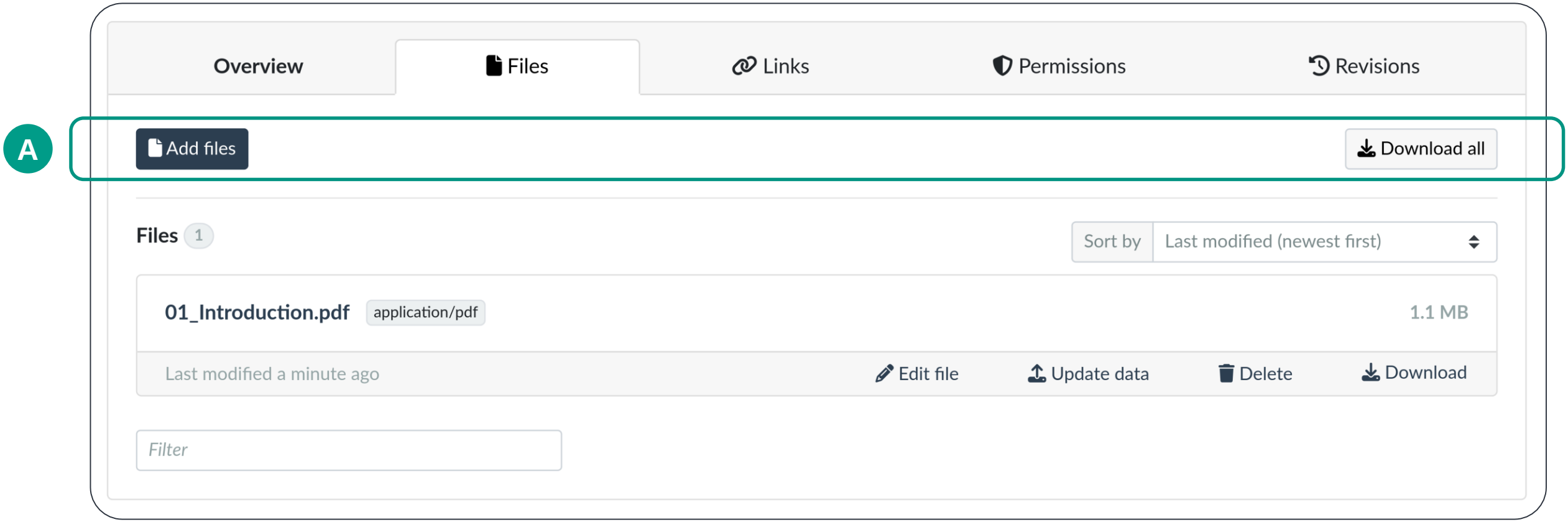
Kadi4Mat – RO-Create Export and Zenodo Publication

- RO-Crates are a container format for packaging research data and metadata of individual or multiple records (when exporting collections) in a ZIP file
- They are also used as container when publishing resources to **Zenodo** via Kadi4Mat
- Furthermore, various different ELN-like systems are collaborating to establish RO-Crate as a means of exporting/importing data in an interoperable manner as part of the ELN Consortium [1]



[1] https://github.com/theelnconsortium

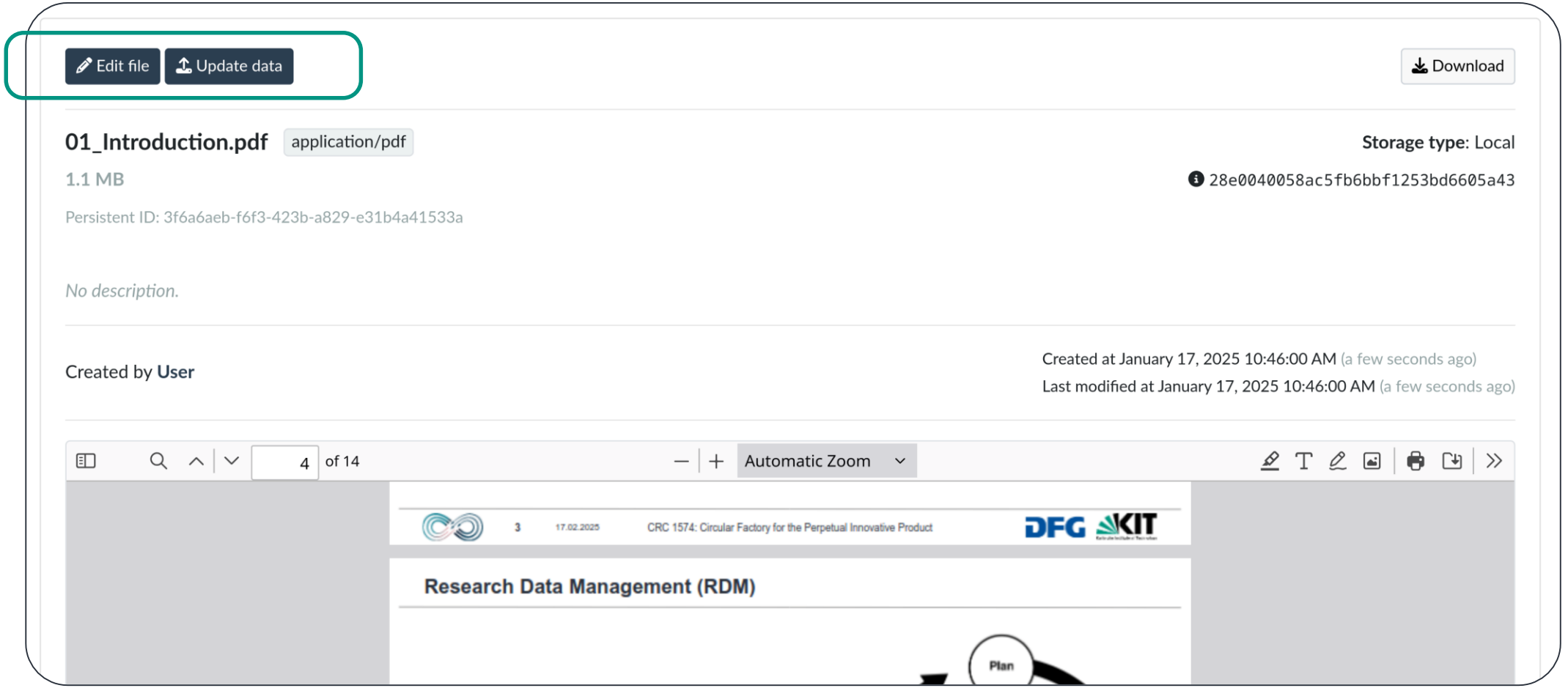
Kadi4Mat – File Overview



(A) Adding new or downloading existing files

Kadi4Mat – File Metadata and Preview

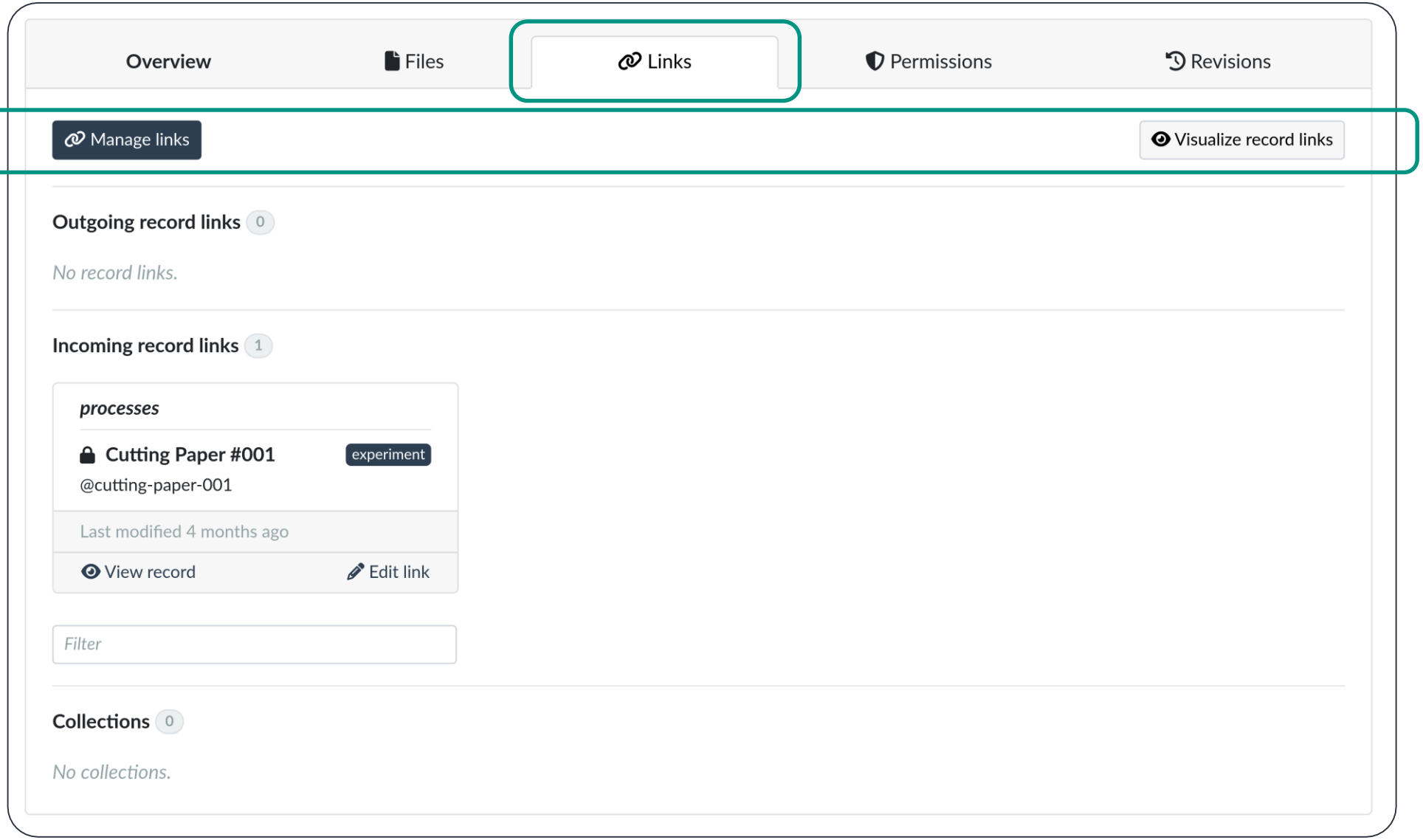
A



(A) Editing the file metadata or updating its data (potentially directly in the web interface, if supported)

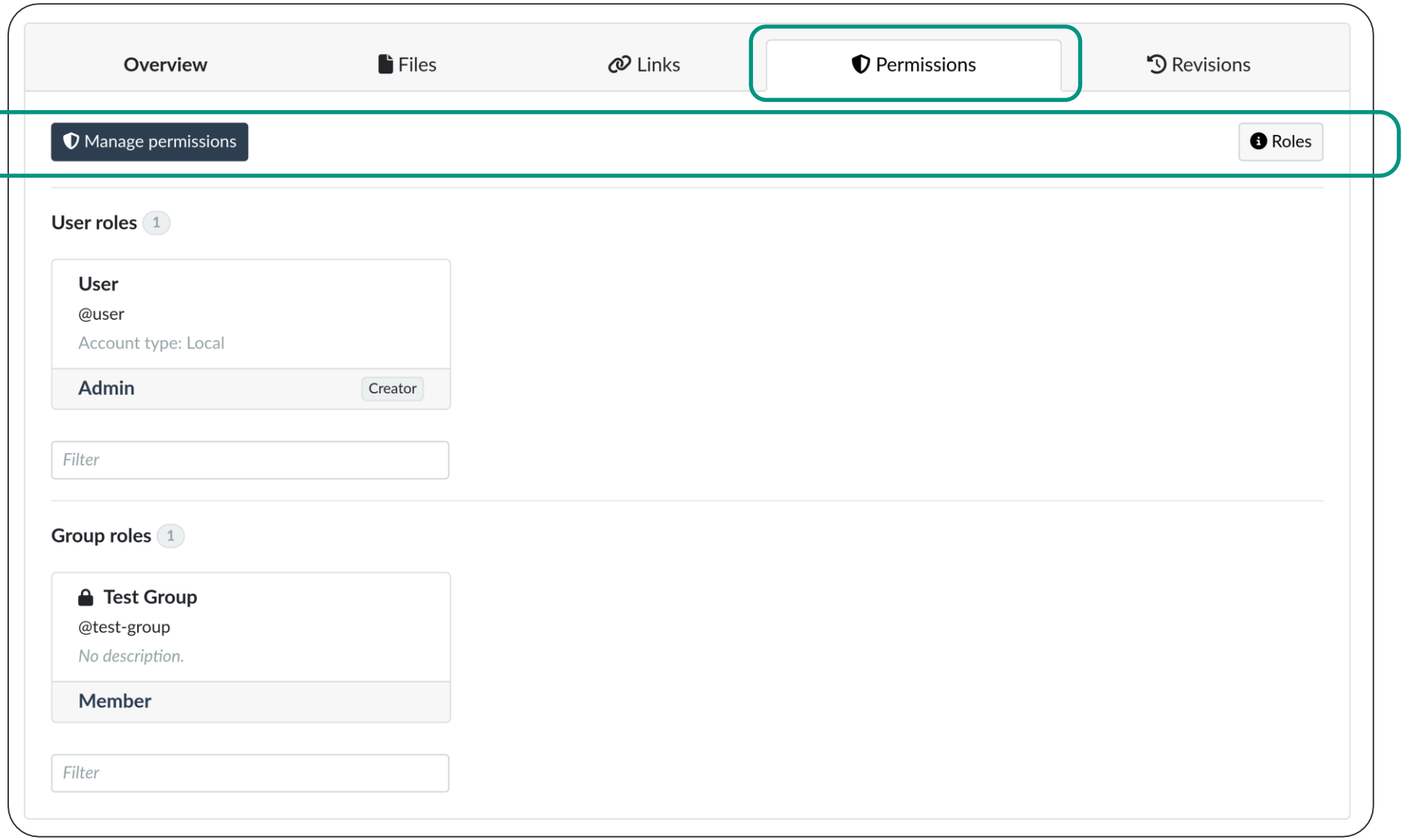
Kadi4Mat – Linked Resources Overview

(A) Managing linked resources (record links and collections) or visualizing record links (as seen previously)

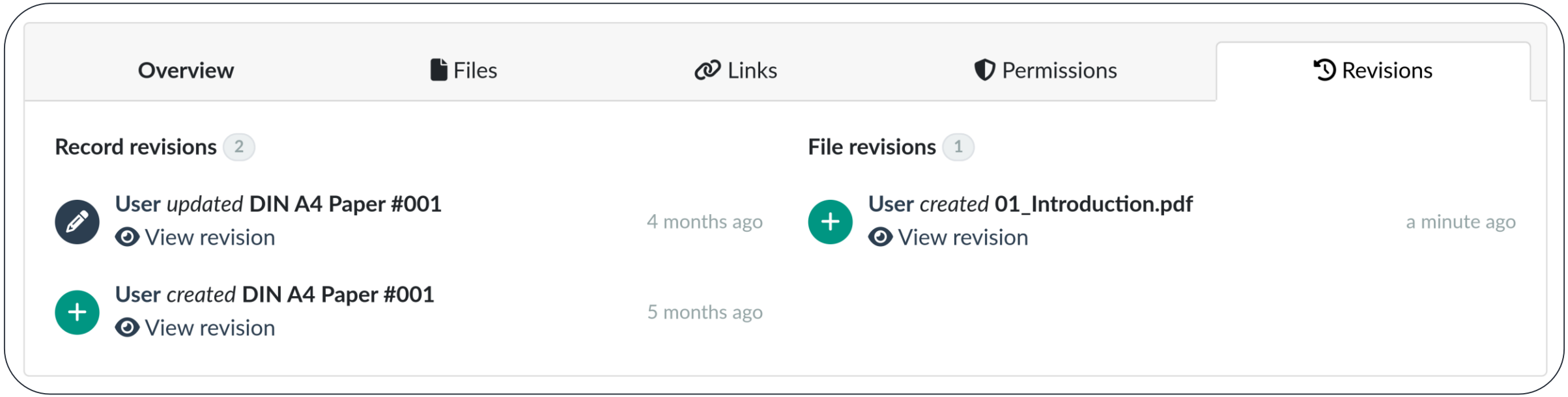


Kadi4Mat – Permission Overview

(A) Managing user/group roles or displaying available roles



Kadi4Mat – Revision Overview

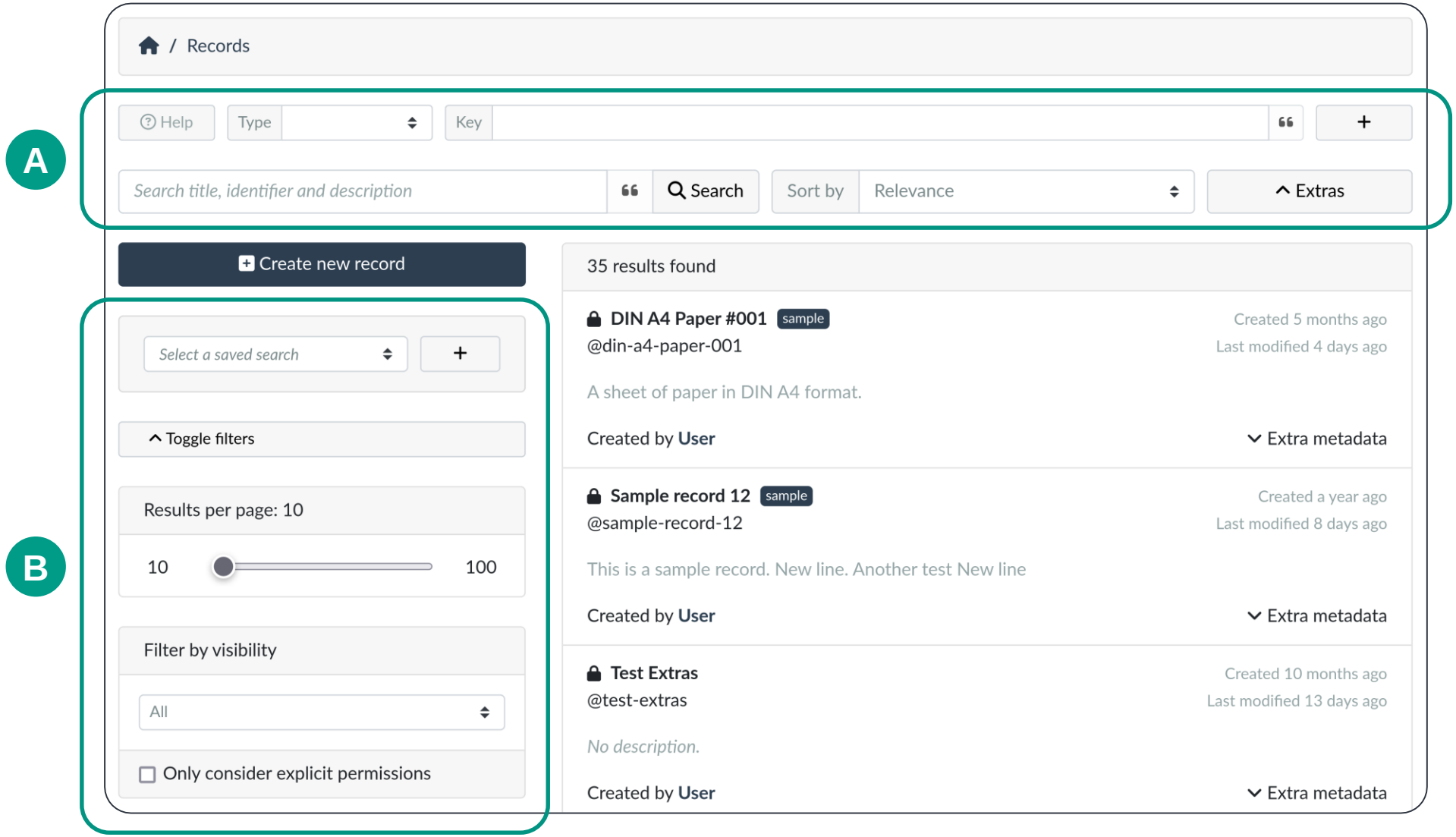


- Records include both record revisions, for record metadata and record link data, revisions of corresponding files
- Note that file revisions currently only include the file metadata

Kadi4Mat – Record Search

(A) Searching of basic (title, identifier, description) or generic record metadata

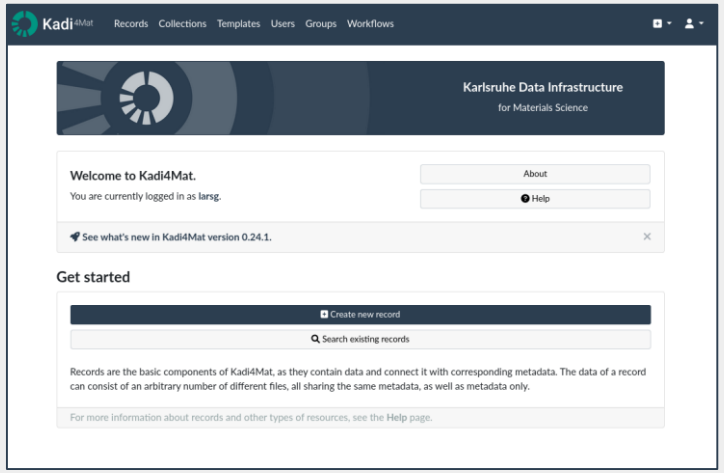
(B) Further filtering of results via different attributes and management of saved search filters



KadiAI and CIDS

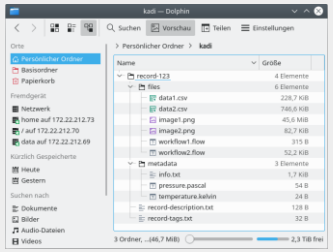
The Kadi Ecosystem

KadiWeb Repository / Web



KadiFS

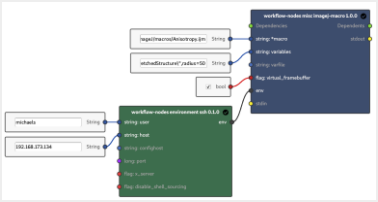
Filesystem



Streamline your research

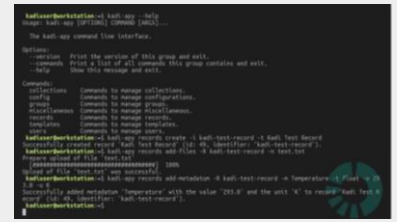
KadiStudio

Workflows



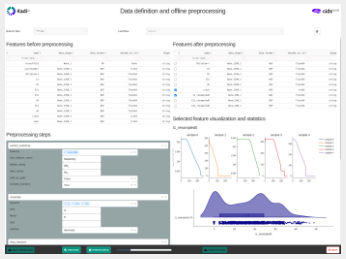
KadiAPY

(Python) API



KadiAI

ML Interface



KadiAI and CIDS

Kadi Artificial Intelligence

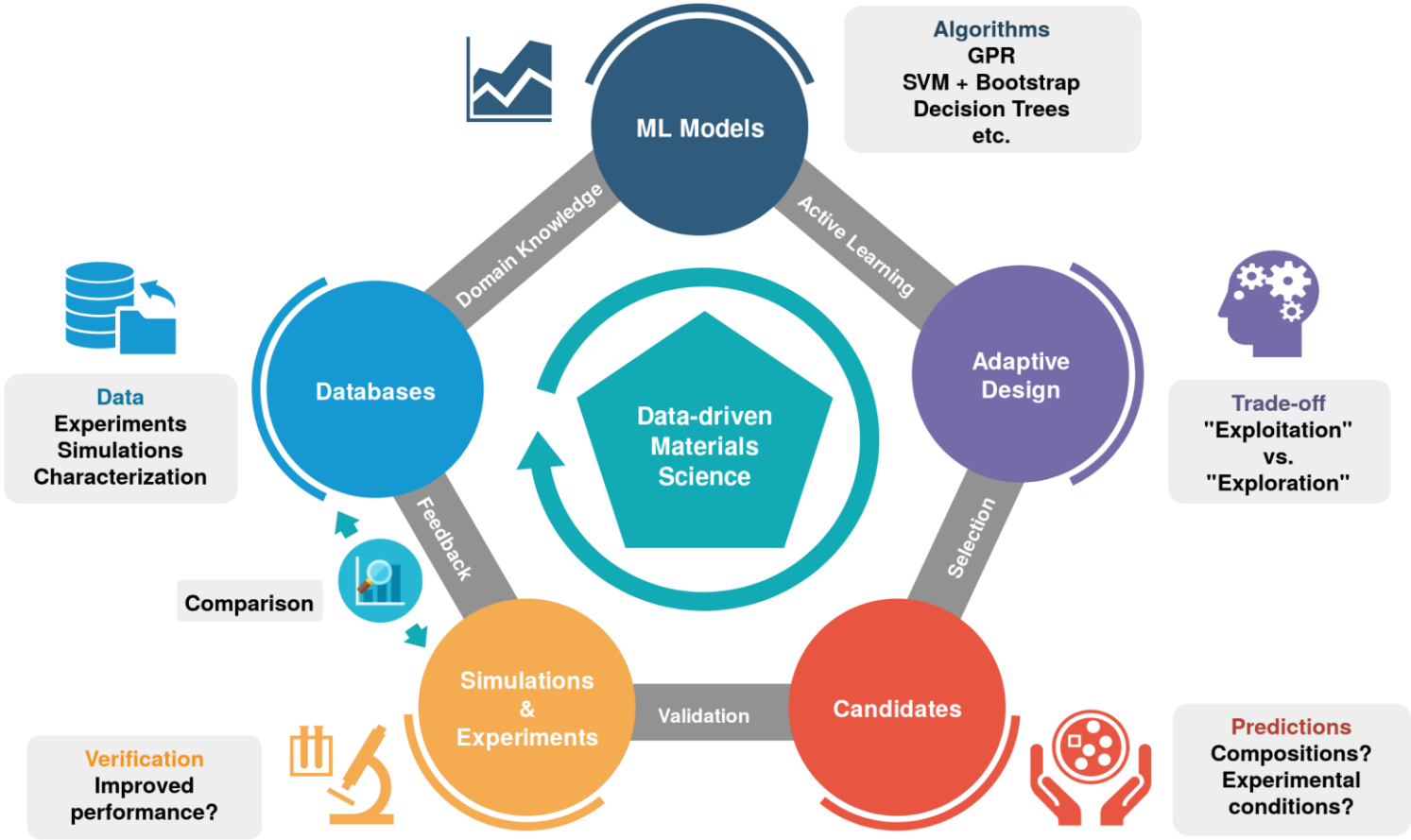
- MLOps: Standardized AI projects and workflows integrated into Kadi
- Reproducible and comparable results
- Data provenance tailored to AI applications
- Interactive definition of the data, model, and training using dashboards

Computational Intelligence and Data Science

- Python framework for AI methods in the STEM sciences
  - Based on Tensorflow/Keras, SciKit-Learn, and SciPy (More to come…)
  - Python script and Jupyter notebooks
- Workflow nodes for Kadi's workflow engine KadiStudio
- Development and standardization of machine learning code

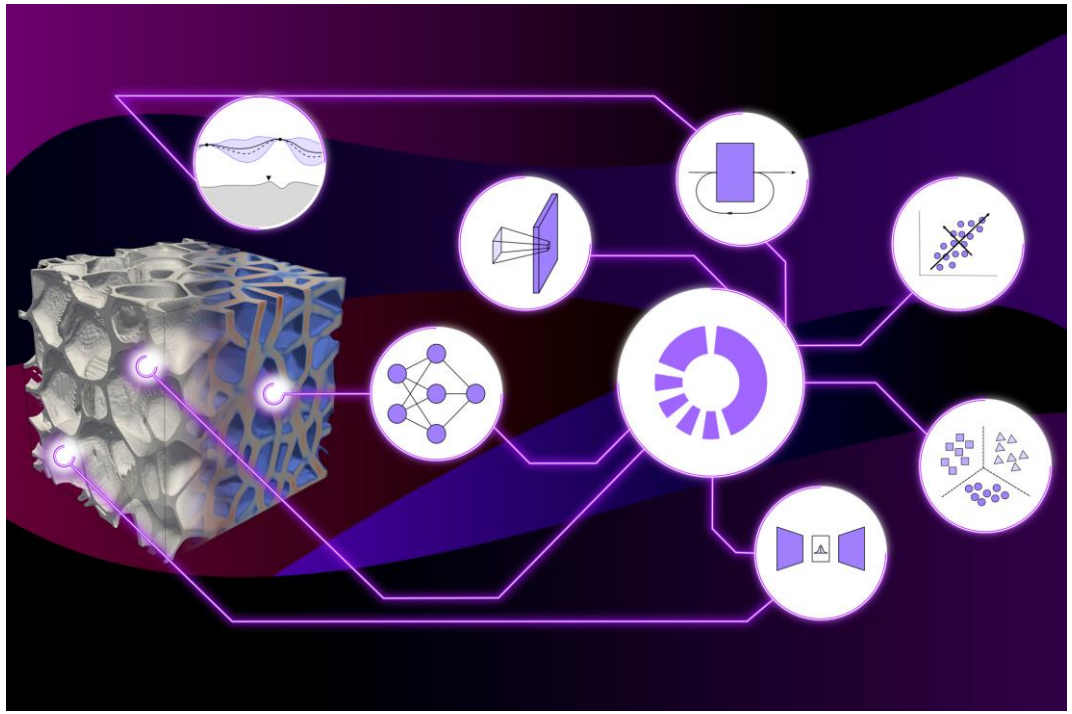
Data-integrated Artificial Intelligence

- Repositories to store and describe data
- Workflows for reproducible simulations, experiments and analyses
- Standardized yet customizable AI models
- AI models and machine learning needs to interact with repository and workflows
  - AI-driven design
  - Active Learning



Data-integrated AI for Systematic Parameter Exploration

- Many studies that involve data or statistics (e.g., in engineering or natural sciences) **investigate parameter spaces** based on performance metrics or objective functions
- Most of these studies have explorative or exploitative characteristics
  - Mapping of parameter sensitivity (e.g., computational benchmarks, clinical trials)
  - Discovery of new materials (e.g., new electrolytes)
  - Optimizing processes or experimental settings (e.g., optimal amount of additives, cost minimization)

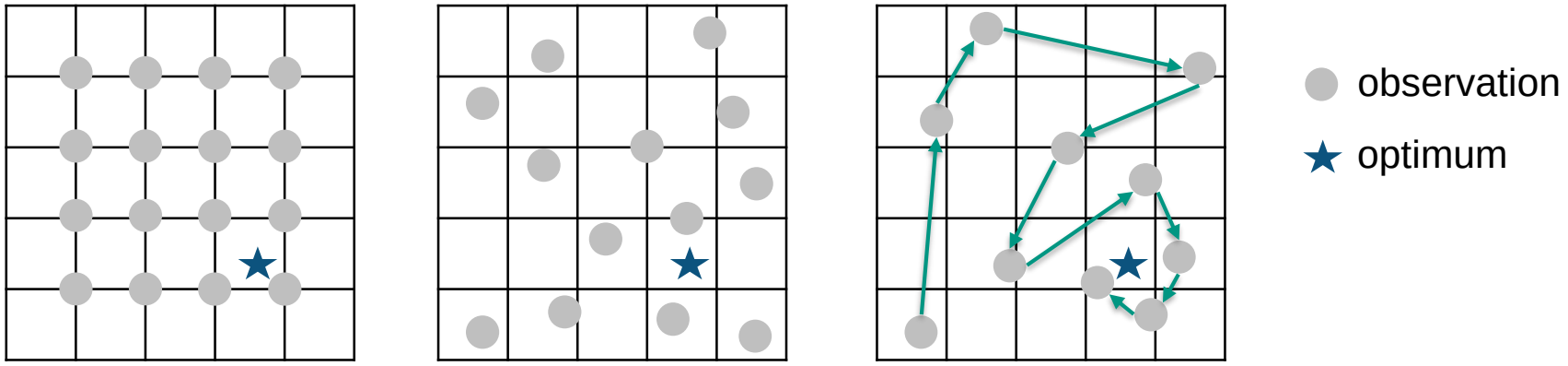


Data-driven modeling / statistics / Machine Learning (ML) / Artificial Intelligence (AI) **aims to accelerate these kind of investigations**

- KadiAIgent plugins and FINALES tenants

Method: Active Learning

- Investigation strategies for global optimization of black-box functions (Design-of-experiments)
  - Sequential: Bayesian Optimization (BO)
  - More efficient than grid search or random sampling (terms and conditions apply)

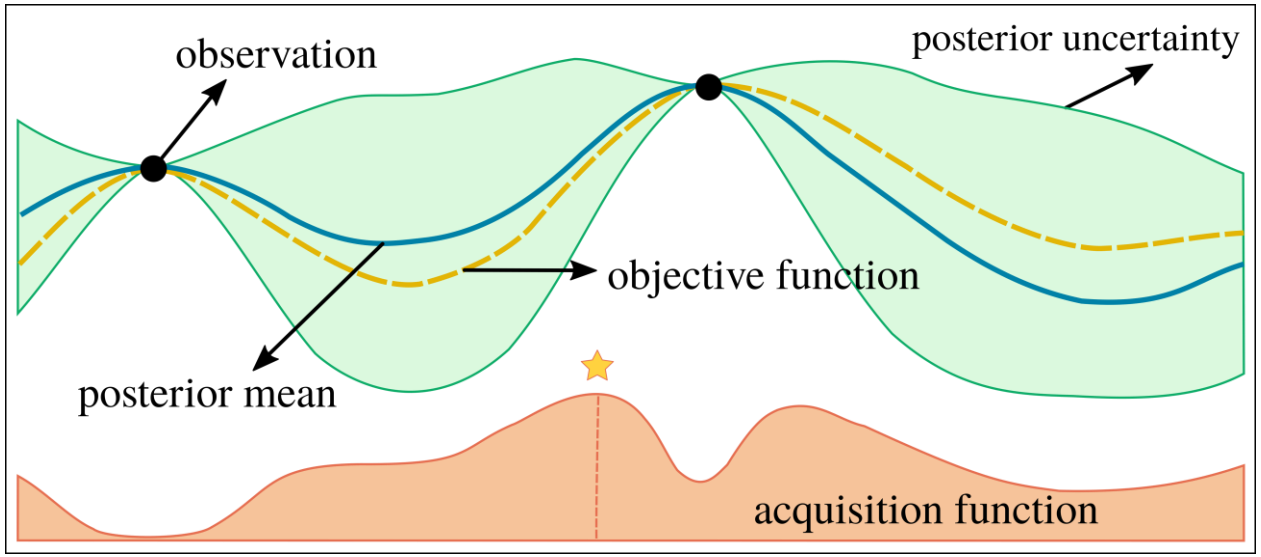


Grid search | Random search | BO search

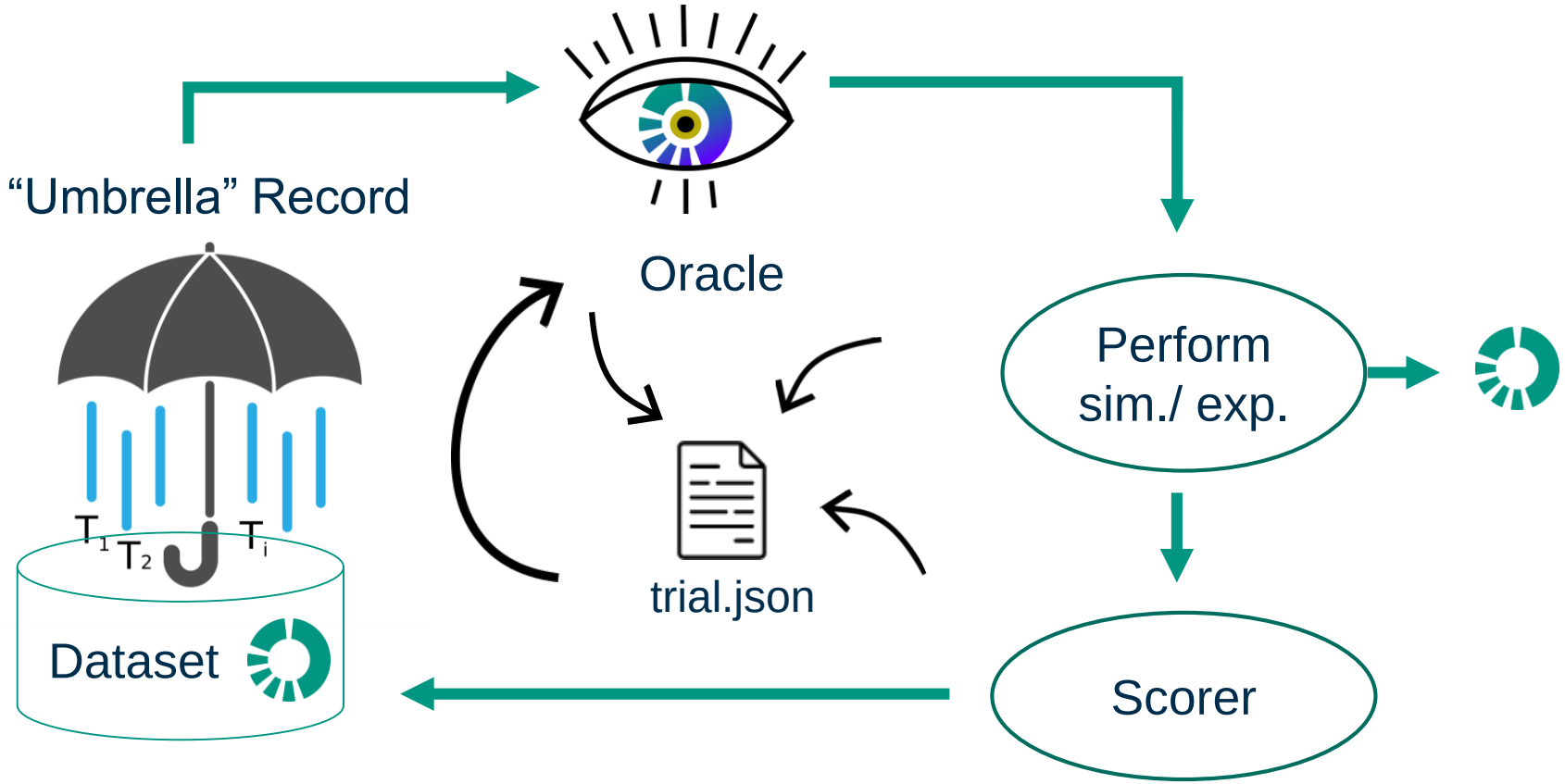
- Advantages
  - Predict promising configuration / candidates
  - Reduce required time for experimental / numerical studies

Bayesian Optimization

- Find the most informative new data point guided by previous evaluations
- Prediction with uncertainty
- Surrogate model
  - Gaussian Process
  - Incorporate the belief on the objective
  - Quantifies the uncertainty in form of a posterior distribution



Active Learning Oracle Schema



Active Learning in KadiStudio

Oracle workflow node

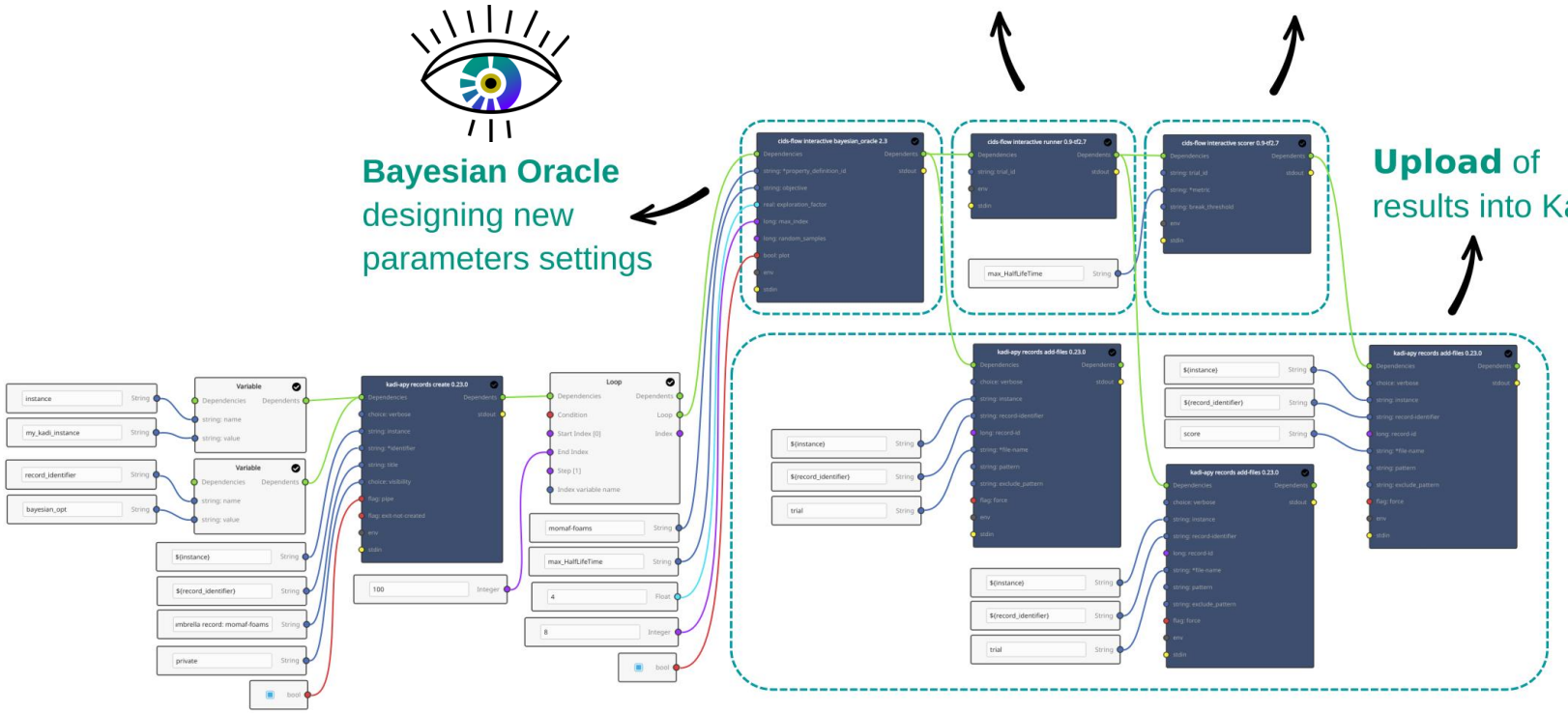
- Perform Bayesian optimization in KadiStudio and link Records in Kadi4Mat repository

Execution of simulation/ experiment

Validation and score of the result

Upload of results into Kadi

Bayesian Oracle designing new parameters settings

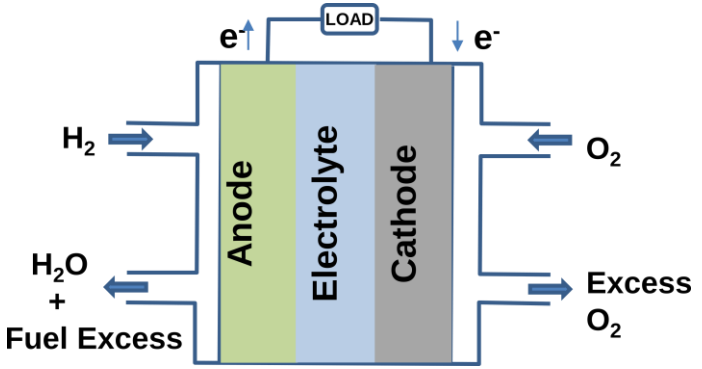


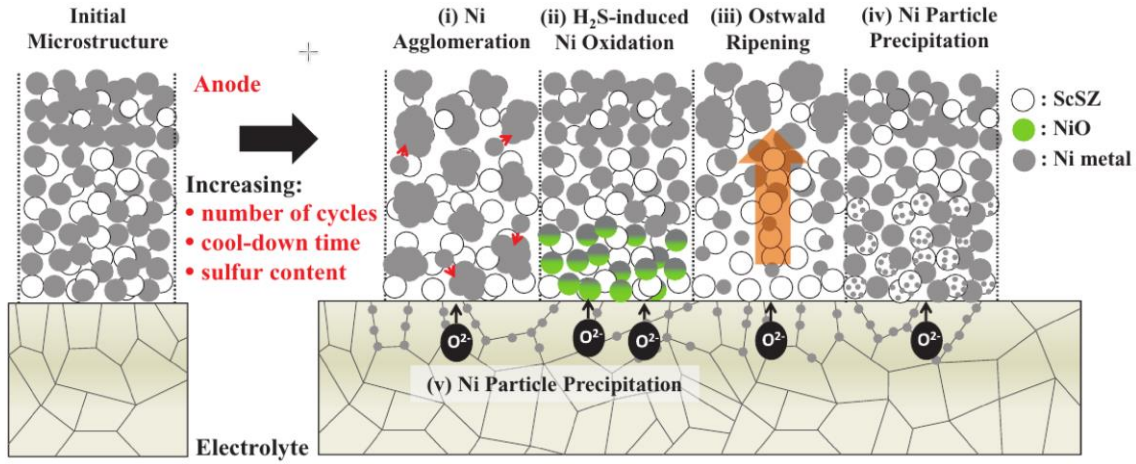
Tools and Use Cases

4

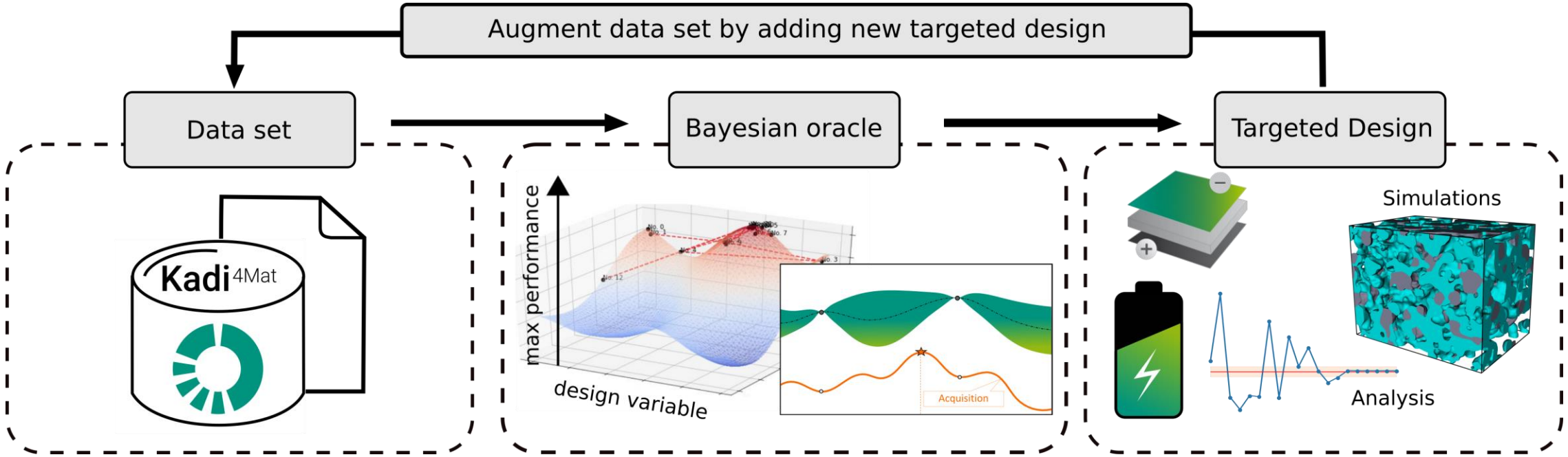
Model Parametrization for Solid Oxide Fuel Cells

- **Solid Oxide Fuel Cell (SOFC)**
  - High efficiency & fuel flexibility
  - Operating temperature 750°C − 1000°C
  - Key micro. properties: TPB density, effective conductivity, specific surface area…
- Degradation
  - Microstructural changes within SOFC anode functional layer
  - Coarsening
- Ni/GDC vs. Ni/YSZ

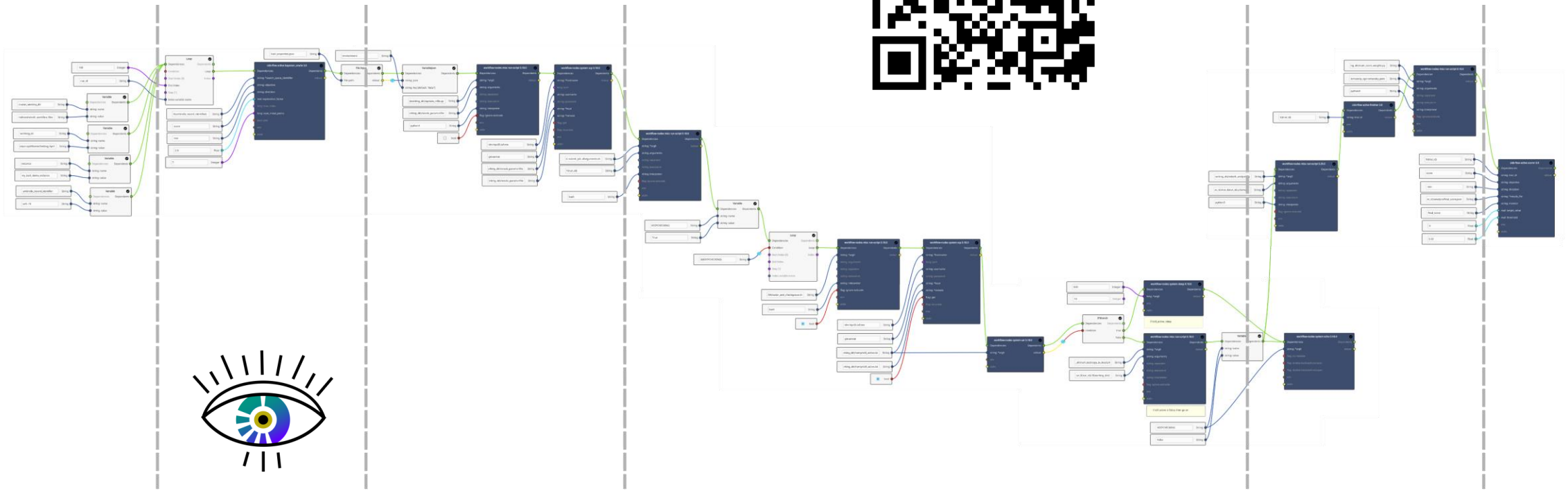




Active Learning Bayesian Optimization – Framework Integration



KadiStudio Workflow



Workspace set up

Oracle

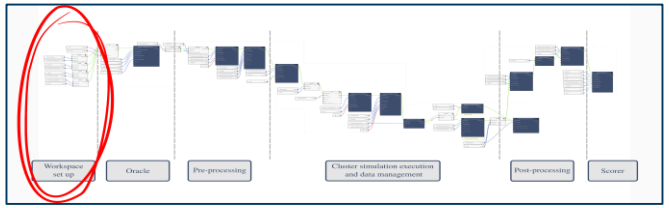
Pre-processing

Cluster simulation execution and data management

Post-processing

Scorer

Workspace – Umbrella Record in KadiWeb



Kadi4Mat Records Collections Templates Users Groups Quick search

Overview | Files | Links | Permissions | Revisions

Edit record | Copy record | As template | Export as | Publish via

Enhancing Solid Oxide Fuel Cells Development through Bayesian Active Learning umbrella

@enhancing-sofcs-through-al

Persistent ID: 42014

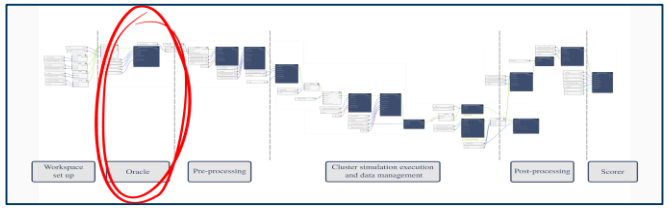
Umbrella record for 7-dim parameter study of Ni/GDC anode in Soli
(scalarization approach).

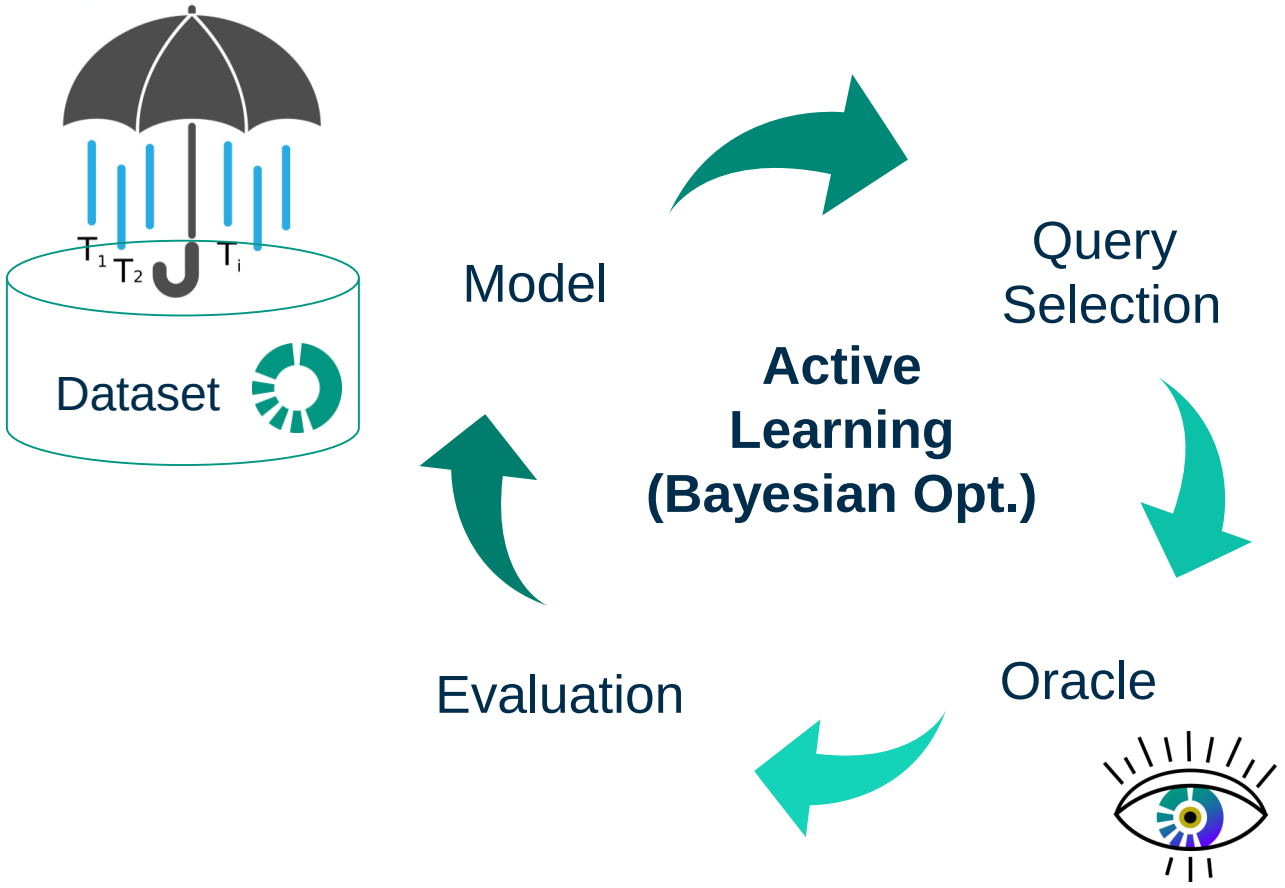
Extra metadata

g_ni_gdc		
min	0.4	
max	1.2	
step	0.01	
g_pore_gdc		
min	0.28	
max	0.8	Float
step	0.01	Float

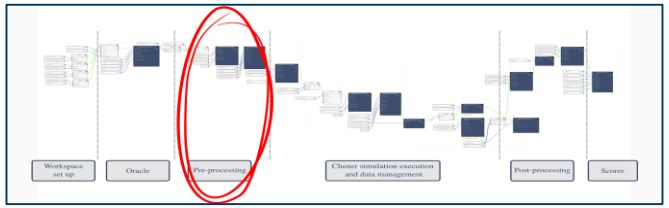
Parameter (symbol)	Value (model units)	Value (physical units)
Interfacial energy Ni-GDC (γ_{NiGDC})	$1.2 \dots 0.4\ u_E/u_l^2$	$3.0 \dots 1.0\ \mathrm{J\ m^{-2}}$
Interfacial energy Pore-GDC ($\gamma_{\mathrm{GDCPore}}$)	$0.8 \dots 0.28\ u_E/u_l^2$	$2.0 \dots 0.7\ \mathrm{J\ m^{-2}}$
Interfacial energy Ni-Pore (γ_{NiPore})	$1.0 \dots 0.6\ u_E/u_l^2$	$2.5 \dots 1.5\ \mathrm{J\ m^{-2}}$
Diff. coeff. GDC-surf. ($M_{\mathrm{GDC}}^{\mathrm{GDCPore}}$)	$(2.11 \times 10^{-6} \dots 2.11 \times 10^{-9})\ u_l^6/(u_E u_t)$	$(3.65 \times 10^{-34} \dots 3.65 \times 10^{-37})\ \mathrm{m^6/J/s}$
Diff. coeff. GDC-nickel/GDC interf. ($M_{\mathrm{GDC}}^{\mathrm{NiGDC}}$)	$(2.11 \times 10^{-8} \dots 2.11 \times 10^{-14})\ u_l^6/(u_E u_t)$	$(3.65 \times 10^{-36} \dots 3.65 \times 10^{-42})\ \mathrm{m^6/J/s}$
Diff. coeff. Ni-nickel/GDC interf. ($M_{\mathrm{Ni}}^{\mathrm{NiGDC}}$)	$(10^{-3} \dots 10^{-8})\ u_l^6/(u_E u_t)$	$(1.73 \times 10^{-31} \dots 1.73 \times 10^{-36})\ \mathrm{m^6/J/s}$
Diff. coeff. Ni-surf. ($M_{\mathrm{Ni}}^{\mathrm{NiPore}}$)	$(10^{-1} \dots 10^{-3})\ u_l^6/(u_E u_t)$	$(1.73 \times 10^{-29} \dots 1.73 \times 10^{-31})\ \mathrm{m^6/J/s}$

Active Learning Oracle



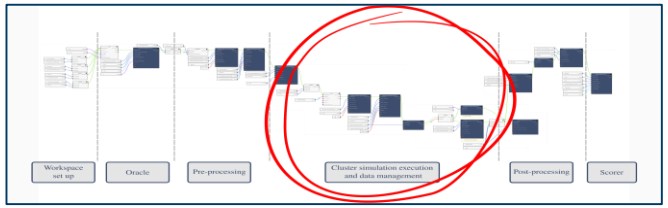


Pre-processing and Data Management

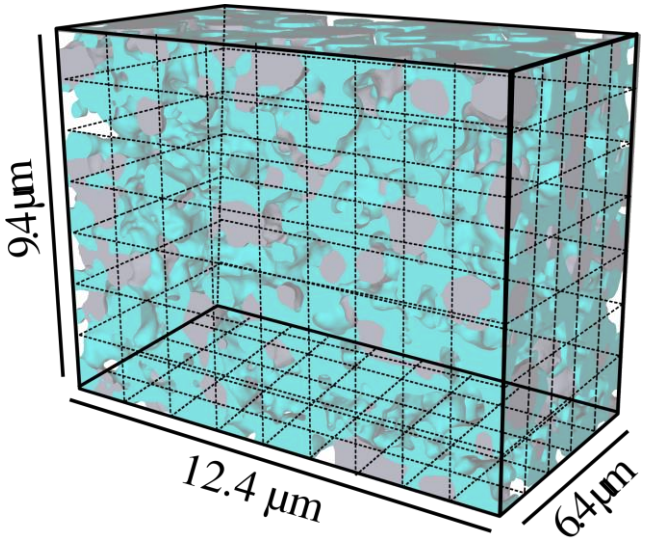


- KadiStudio allows to integrate:
  - Experimental devices
  - Numerical methods
  - Pre-built tool nodes
  - Customizable functions
- For the AL framework:
  - Conversion of Oracle output information into simulation inputs

HPC-Cluster (idm-hpc03)



- 8 computing nodes
- Each node equipped with two AMD EPYC™ 7452 processors
- 64 cores per node (32 cores per processor)
- Interconnect: via 100 Gbit/s InfiniBand network
- Job Scheduler: SLURM

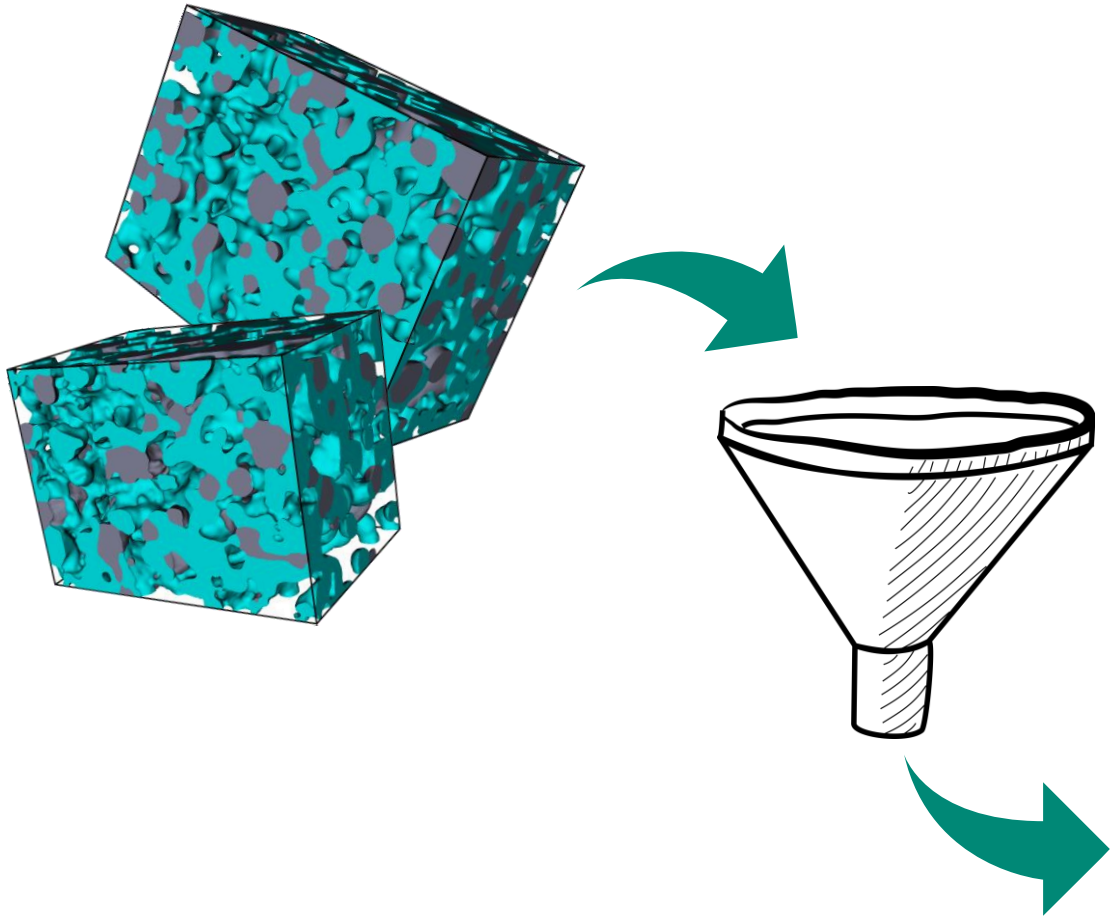


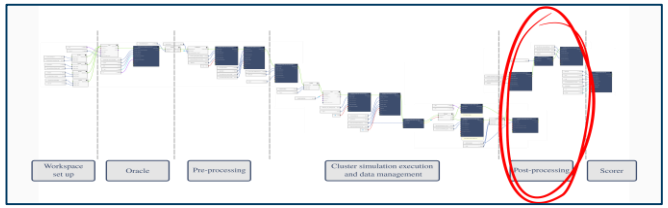
- For our study:

  Domain decomposition partition: (7, 9, 5)

  `#SBATCH -n 315`

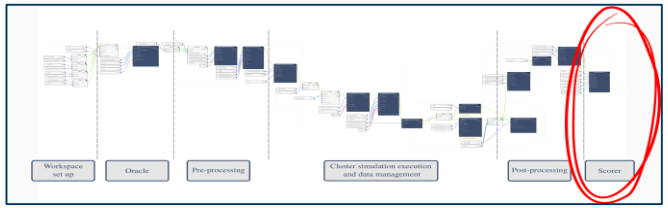
Post-processing



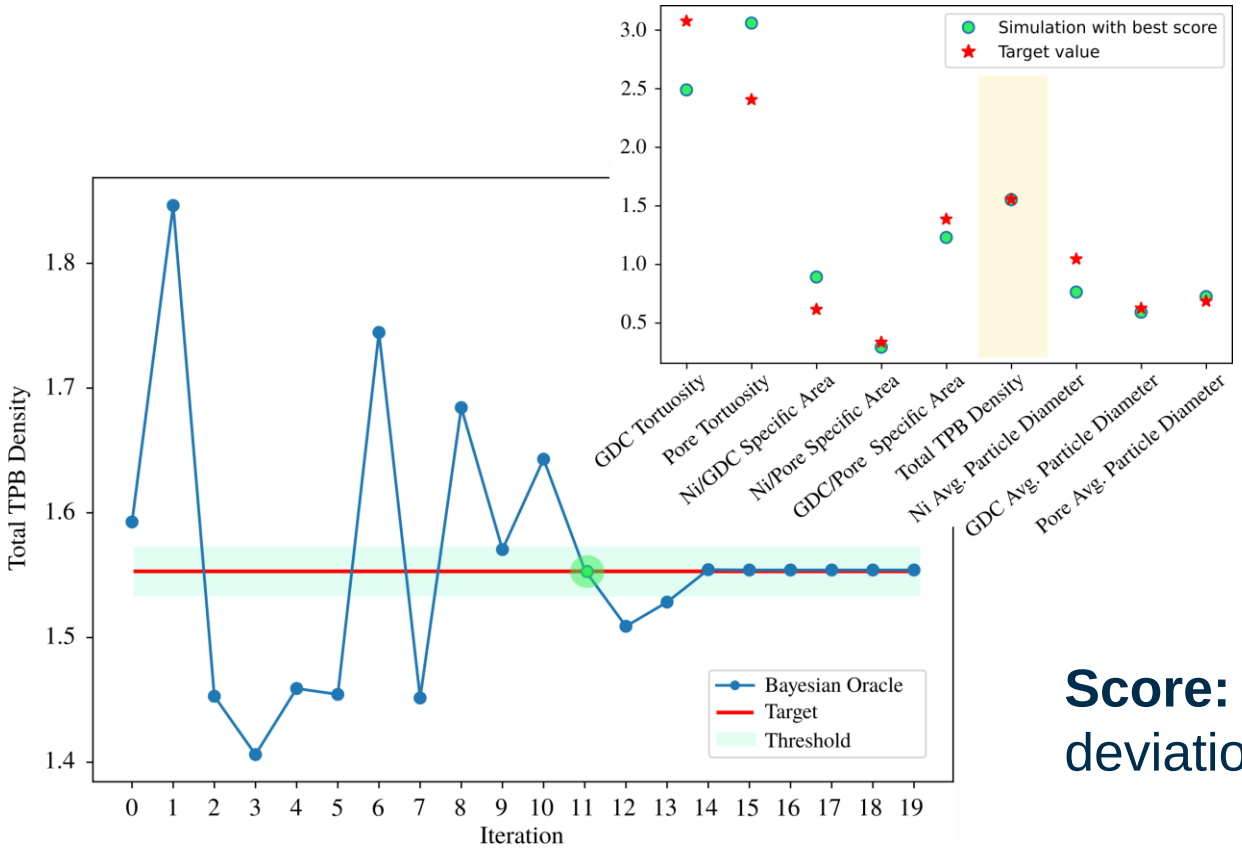


GDC Tortuosity	3.2e-3
Pore Tortuosity	9.8e-4
Ni/GDC Specific Area	7.9e-5
Ni/Pore Specific Area	7.5e-4
GDC/Pore Specific Area	3.3e-5
Total TPB Density	5.2e-4
Ni Avg. Particle Diameter	6.0e-4
GDC Avg. Particle Diameter.	1.5e-7
Pore Avg. Particle Diameter.	2.5e-5

Scoring System (naive)

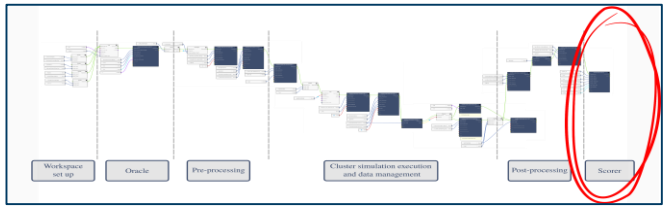


- Single objective optimization: one microstructural property as objective



Score: Triple Phase Boundary Density deviation

Scoring System – Scalarization Approach



- (Degenerative) multi-objective optimization

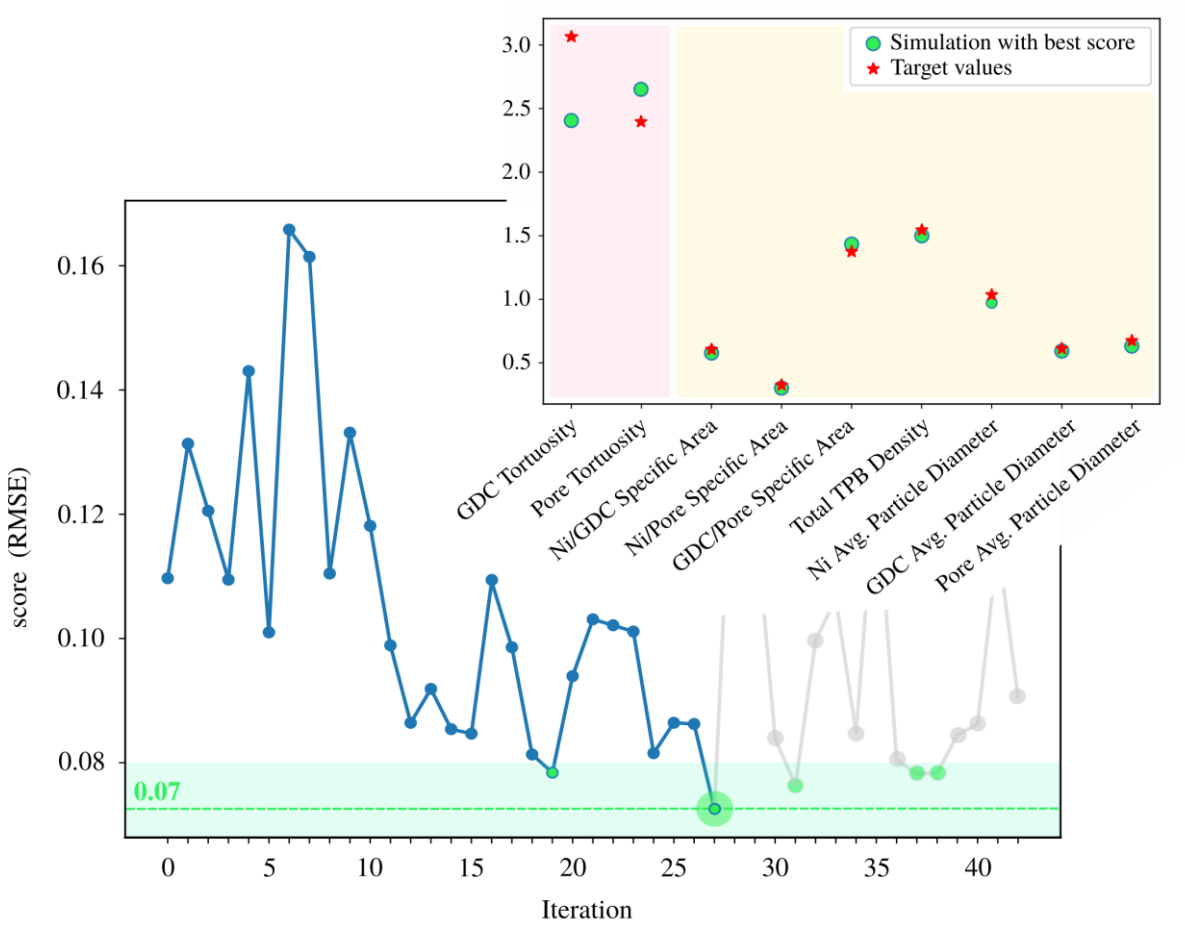
$$F(x) = w_1 * f_1(x) + w_2 * f_2(x) + \cdots + w_n * f_n(x)$$

where,

$f_1(x), \cdots, f_n(x) \rightarrow objective\ functions$
$w_1, \cdots, w_n \rightarrow weights$

Score: RMSE of 7 meaningful microstructure parameters

GDC Tortuosity
Pore Tortuosity
Ni/GDC Specific Area
Ni/Pore Specific Area
GDC/Pore Specific Area
Total TPB Density
Ni Avg. Particle Diameter
GDC Avg. Particle Diameter.
Pore Avg. Particle Diameter.

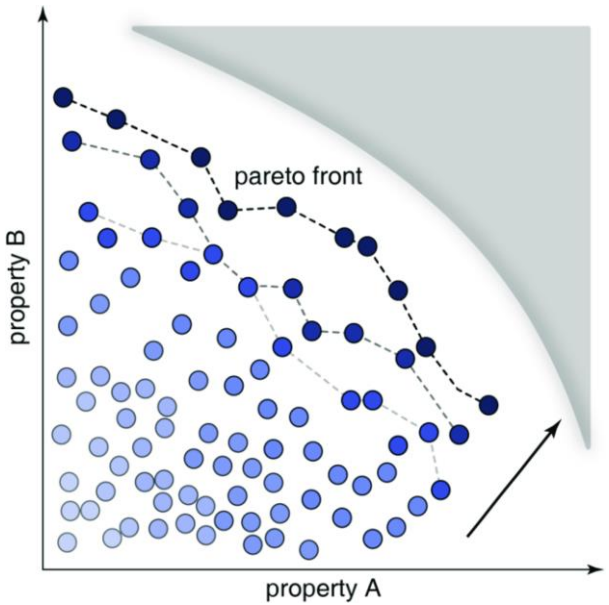


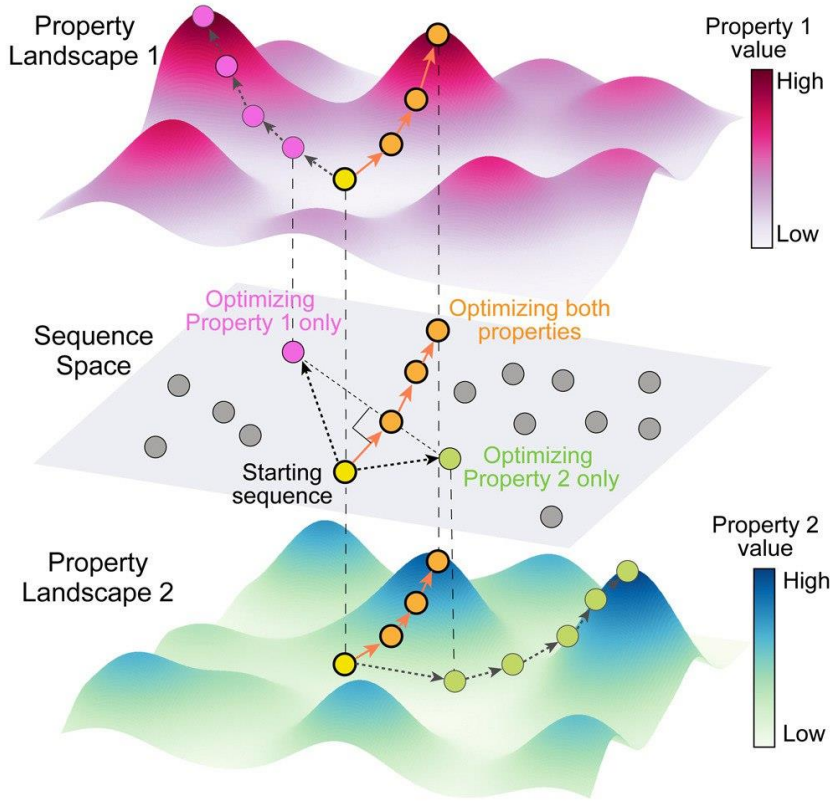
Multi-objective Optimization

- Design problems often involve more than one objective
  - Possibly conflicting objective functions

$$\begin{aligned} \min_{\mathbf{x}} \quad & \{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})\} \\ \text{subject to} \quad & \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ & \mathbf{h}(\mathbf{x}) = \mathbf{0} \end{aligned}$$

- Set-valued solutions: Pareto front (objective function space), Pareto set (design variable space)
  - Search for non-dominated points
  - Best possible trade-offs between conflicting objectives

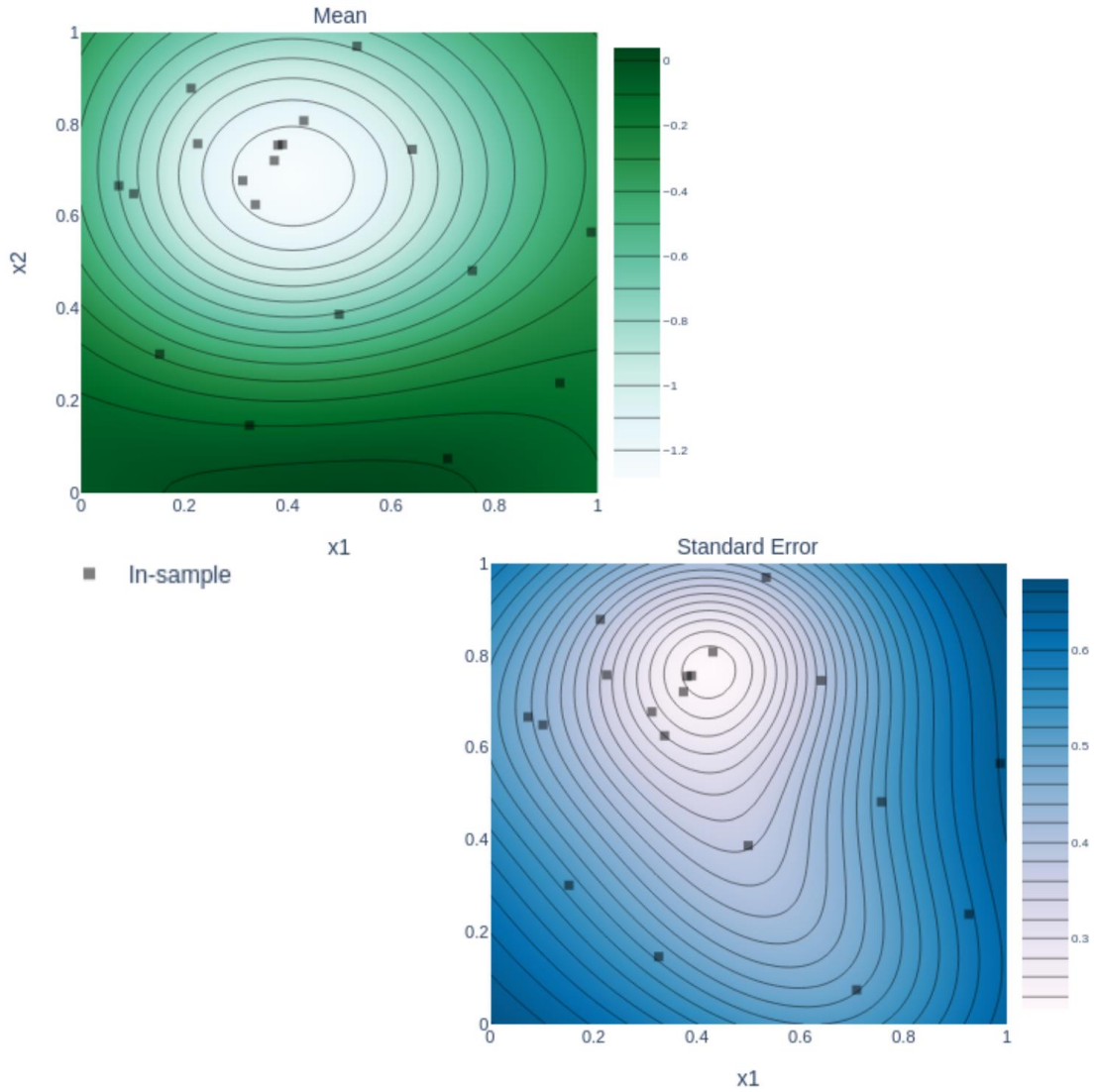




Luo, Jiaqi et al., iScience, Volume 28, Issue 3, 112119

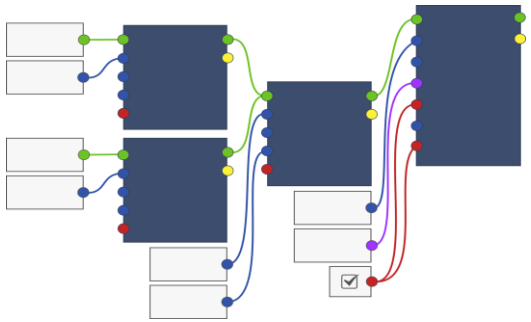
KadiAlgent
- new system

- Ax - Adaptive Experimentation Platform
  - User-friendly wrapper for BoTorch
  - Provides higher-level APIs and infrastructure for experiment management
  - Multi-objective optimization
  - Combines built-in optimization algorithms (Sobol initialization followed by Bayesian)
  - Built-in acquisition functions (e.g., qNEHVI)
  - Visualization options (Plotly)
    - Contour plots
    - Slice plots
    - Trade-offs (plot_objective_vs_constraints)
    - Pareto frontier

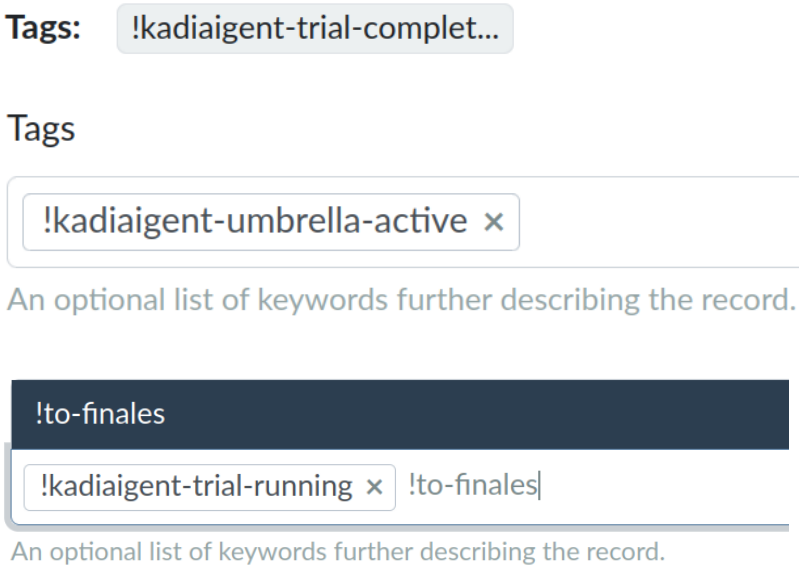


KadiAIgent – Trigger Options

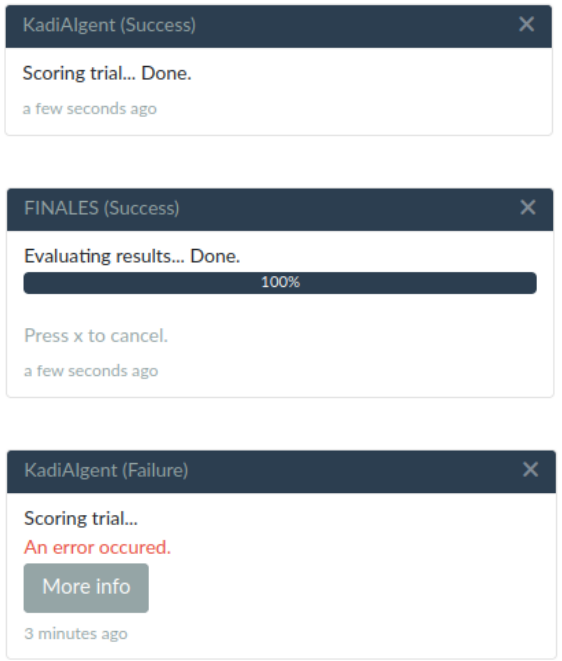
- KadiStudio



- Kadi4Mat
  - control tags



…also allow task notifications



FINALES: Human-in-the-loop automatized experiments

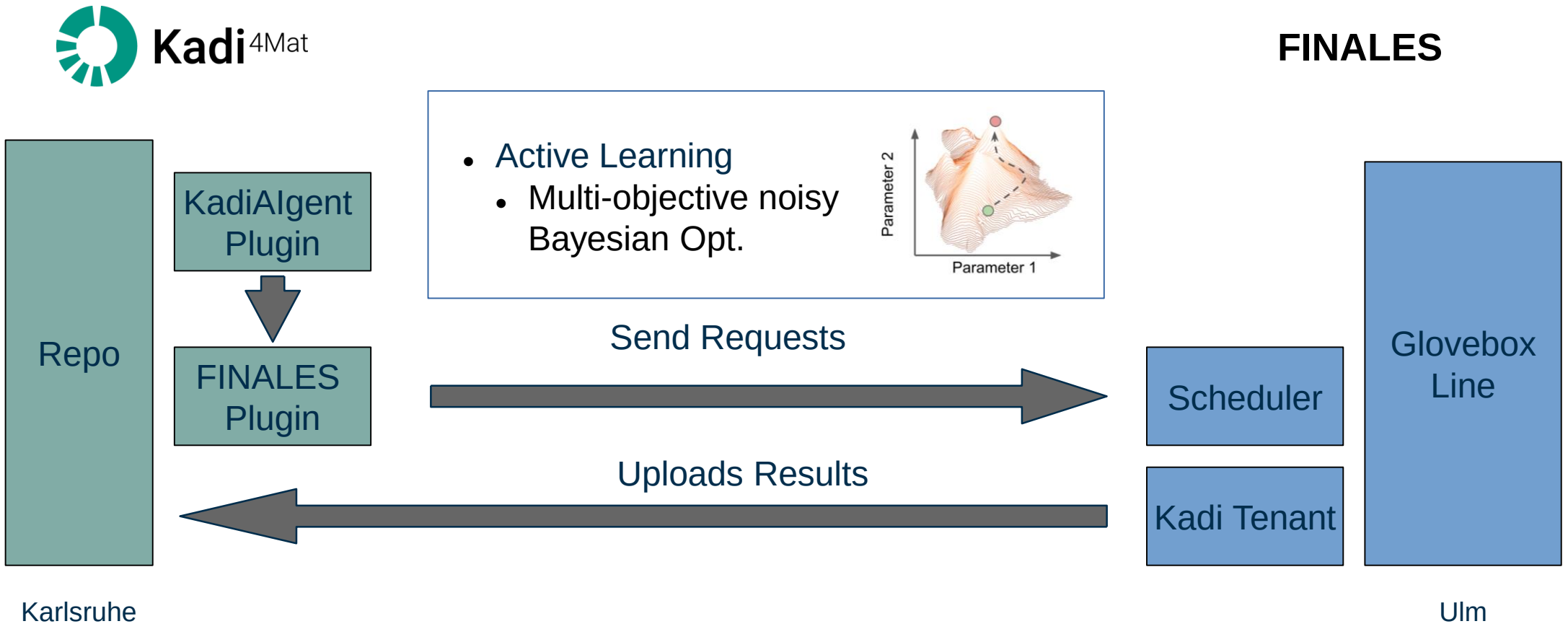
Use-case: optimizing the formation of Sodium battery systems

- Improve the efficiency of battery production vs. the lifetime of batteries
- Multi-objective optimization problem with noisy parameters and objective functions
- POLiS Cluster of Excellence: automated battery assembly glovebox in Ulm

AutoBASS

- Project team:
  - FINALES: Monika Vogler, Leon Merker
  - Kadi4Mat: Arnd Koeppe, Giovanna Tosato, Johannes Steinhülb, Nico Brandt

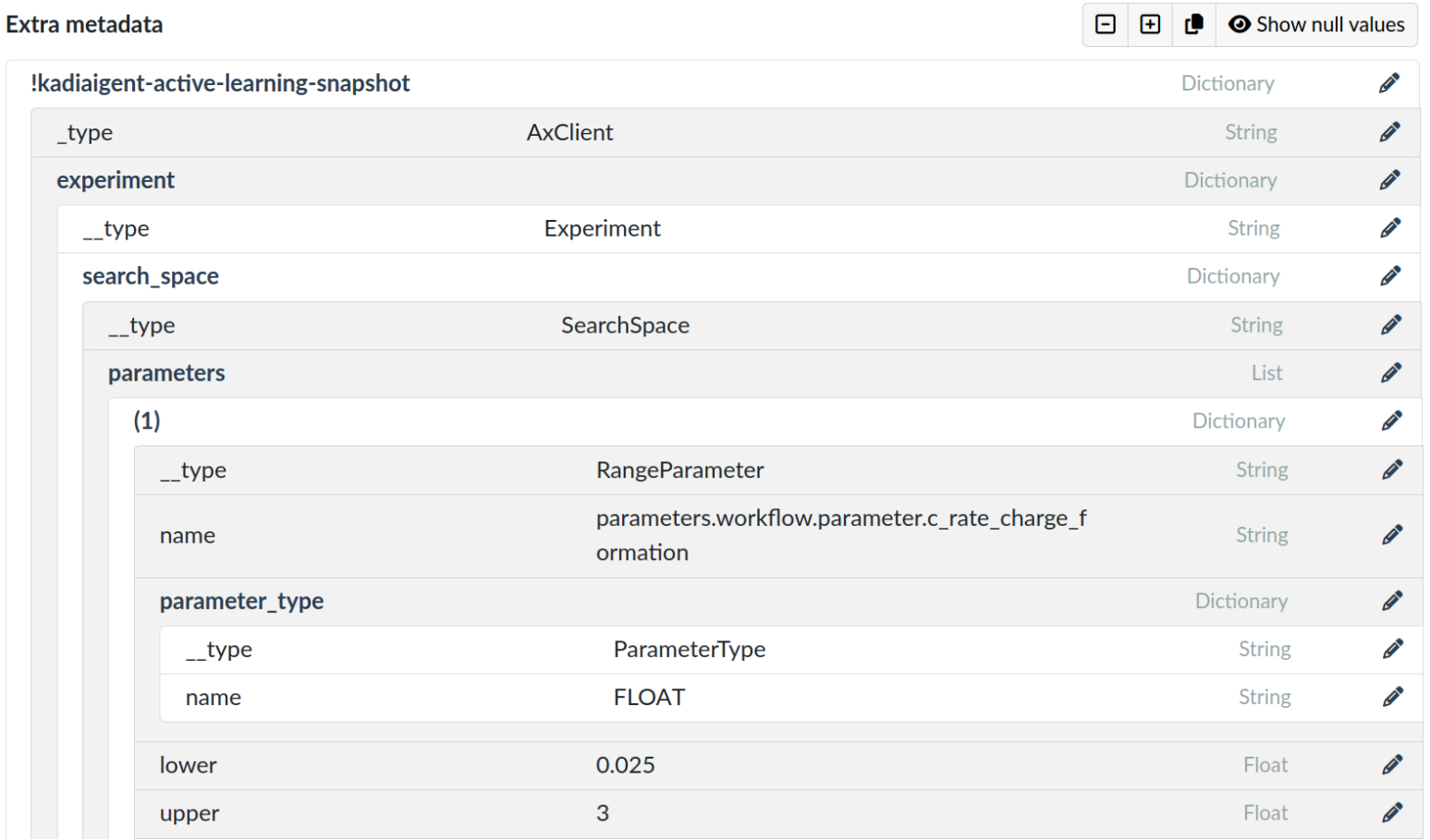
Data Exchange Concept



KadiAIgent: Umbrella – Trial – Interface Records

Umbrella

- Design space and optimization settings



KadiAIgent: Umbrella – Trial – Interface Records

!use-finales-template

Umbrella

- Design space and optimization settings

FINALES Baseinterface

- Blueprint for requests to FINALES (workflow)

!kadiaigent-trial-for

Extra metadata

!kadiaigent-active-learning-trial-parameters

parameters.workflow.parameter.c_rate_charge_formation	2.0072262854627922
parameters.workflow.parameter.c_rate_discharge_formation	1.8230733814819184
parameters.workflow.parameter.repetions_formation_cycle	2

Trial

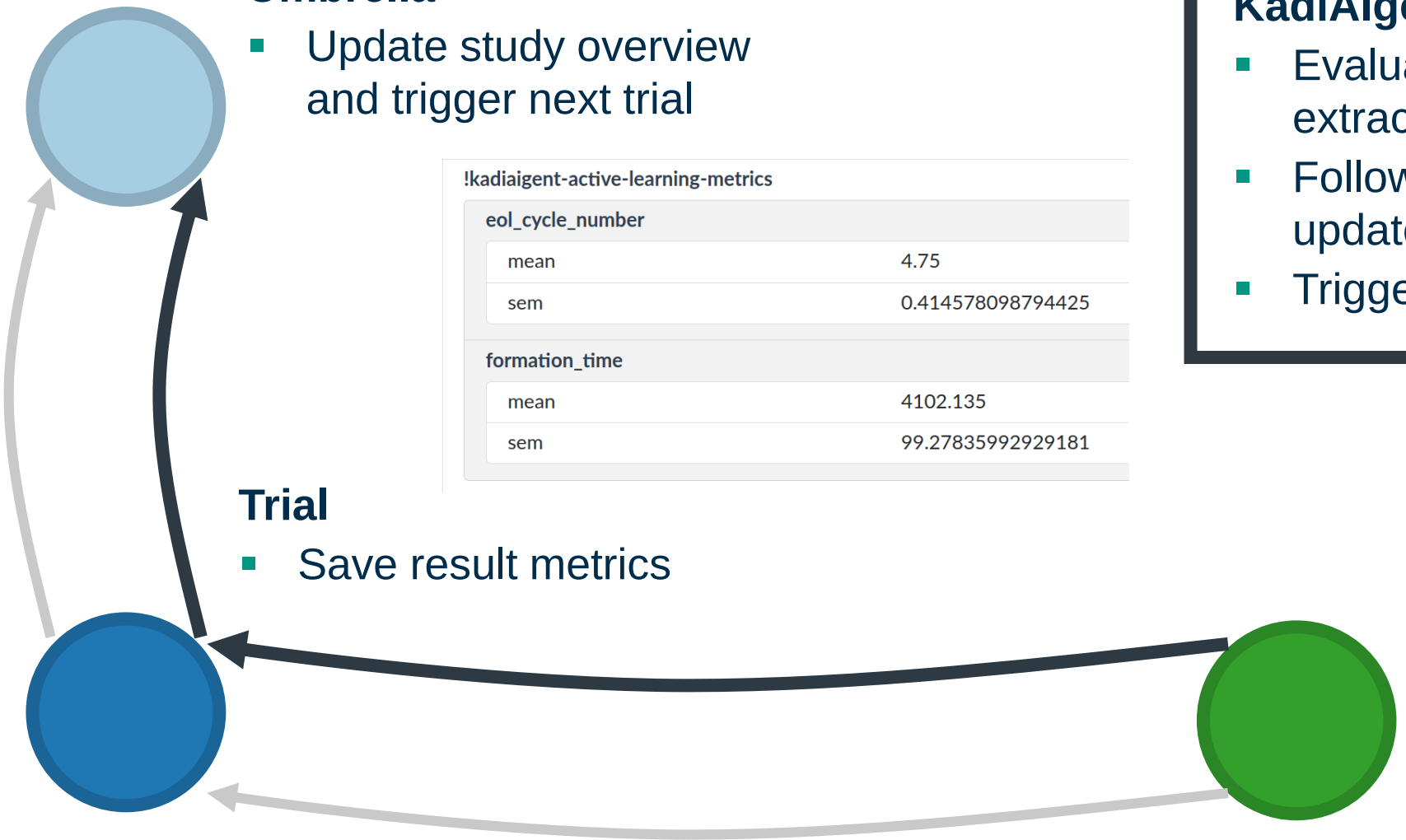
- Selected design parameter configuration

FINALES Interface

- Replace values in base request and send to FINALES (unique identifier)

!finales-request-for

KadiAlgent: Evaluation



Umbrella

- Update study overview and trigger next trial

!kadiaigent-active-learning-metrics

eol_cycle_number	
mean	4.75
sem	0.414578098794425
formation_time	
mean	4102.135
sem	99.27835992929181

Trial

- Save result metrics

KadiAlgent and FINALES plugins

- Evaluate FINALES response and extract result metrics
- Follow links back to umbrella and update study
- Trigger next iteration

FINALES Interface

- Receives/finds response from FINALES

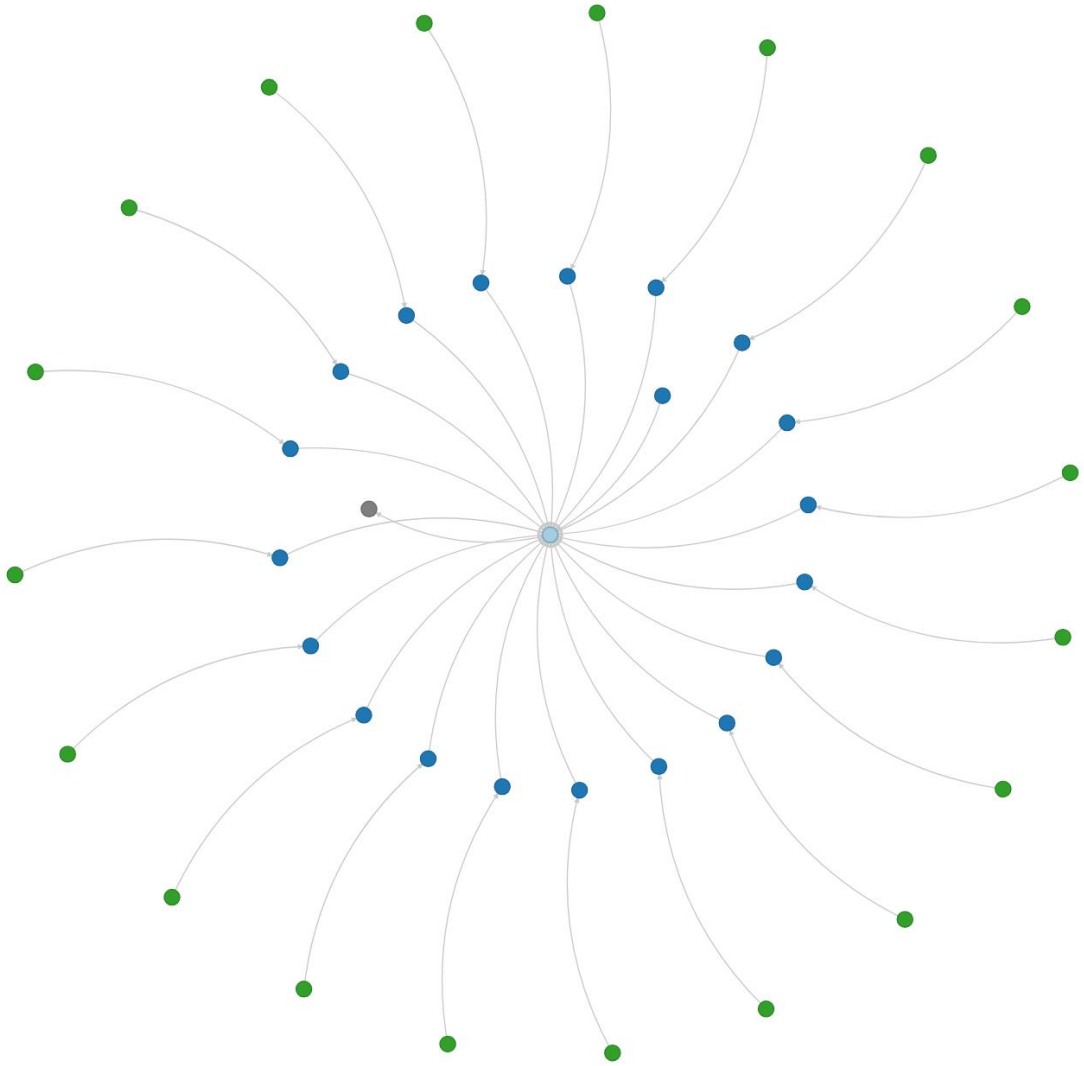
KadiAIgent: Umbrella – Trial – Interface Records

Umbrella

Base FINALES interface

Trial

FINALES interface



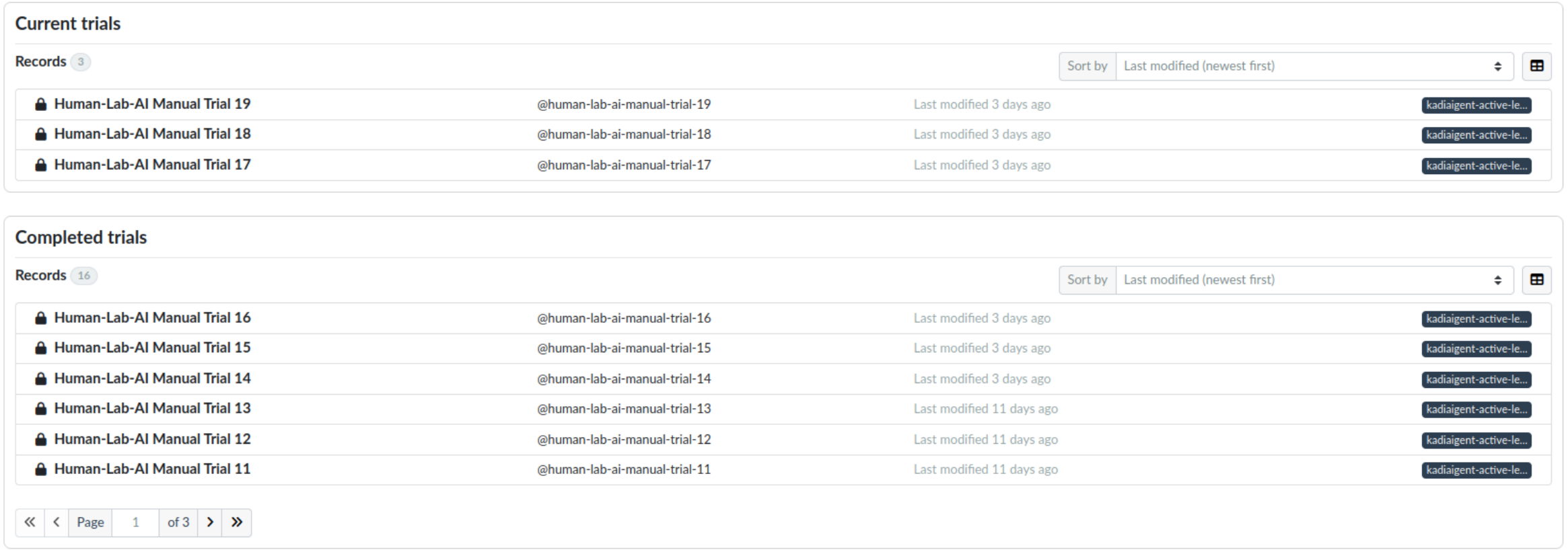
Human interface: dashboard

Plotly figure (scatter)

- Use it as interactive figure for your project

- Here: trade-off between multiple objectives as status of study

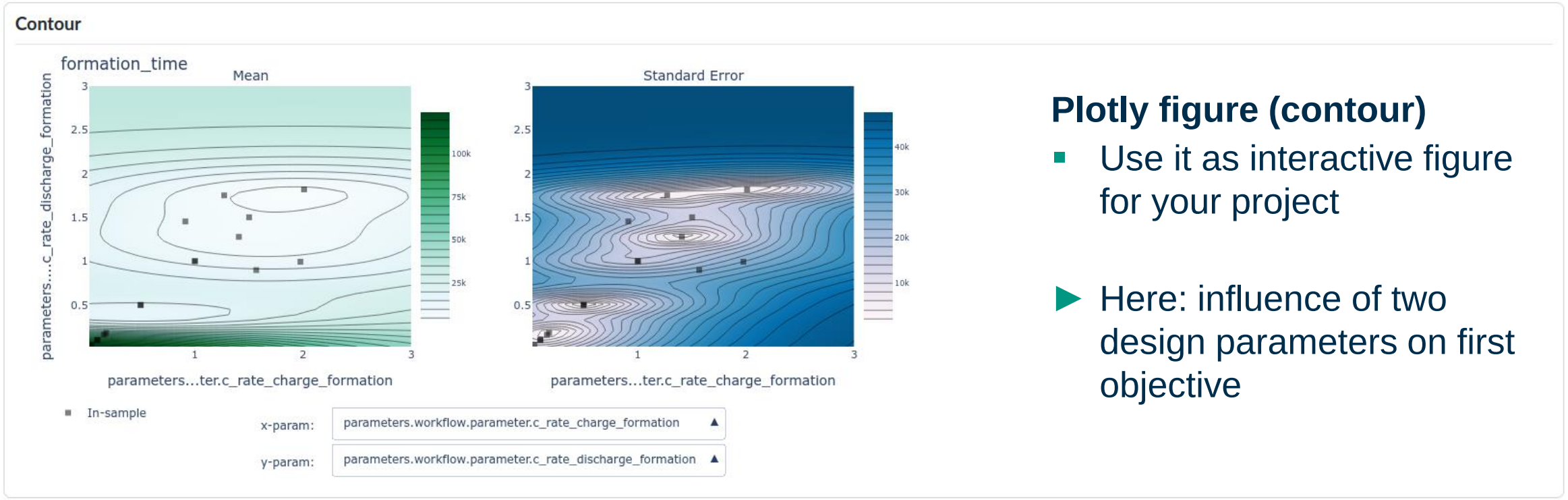
Human interface: dashboard



Record view

- Use it to quickly skim, access, or create specific records from your project
- Here: shows the currently running and completed trials

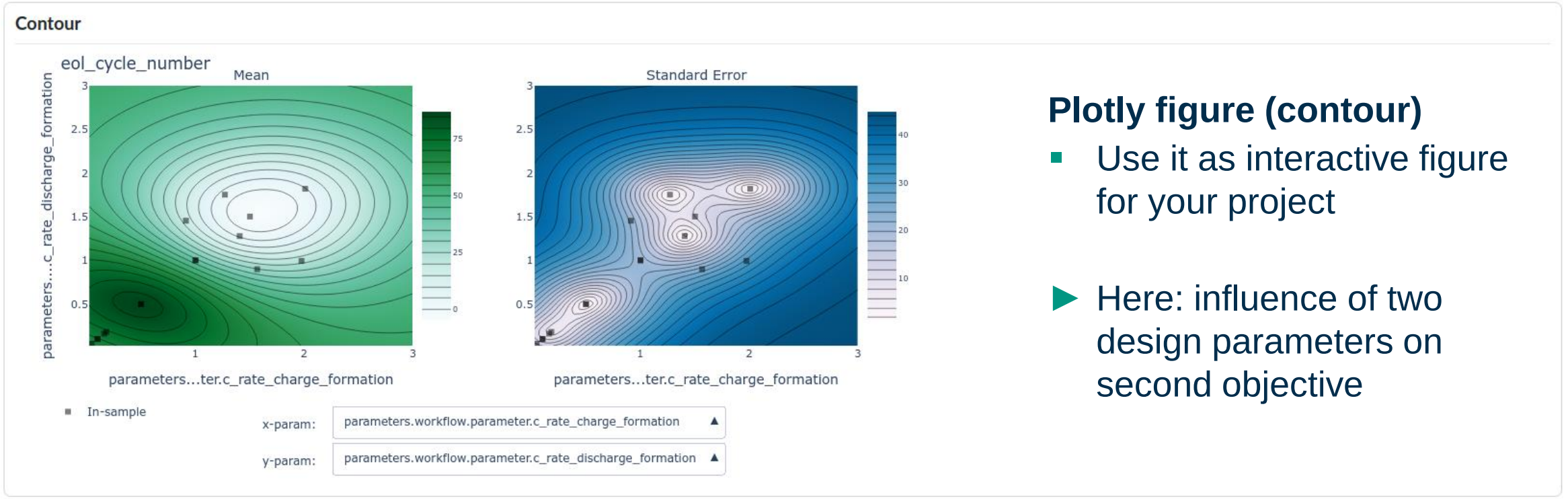
Human interface: dashboard



Plotly figure (contour)

- Use it as interactive figure for your project
- Here: influence of two design parameters on first objective

Human interface: dashboard

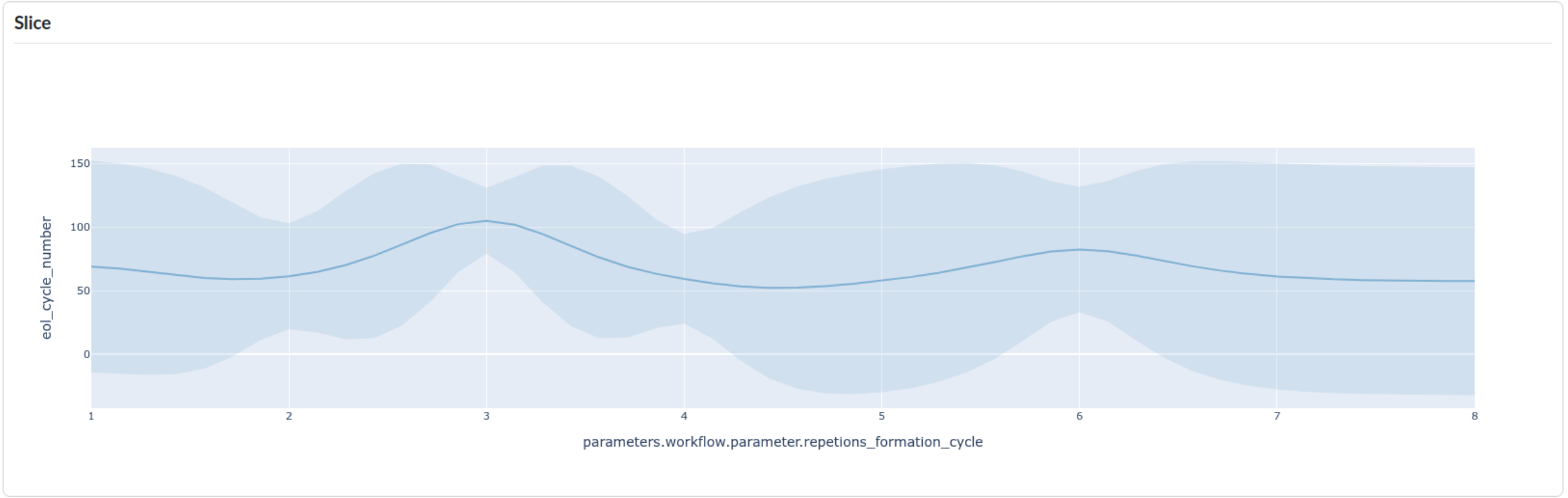


Plotly figure (contour)

- Use it as interactive figure for your project

- Here: influence of two design parameters on second objective

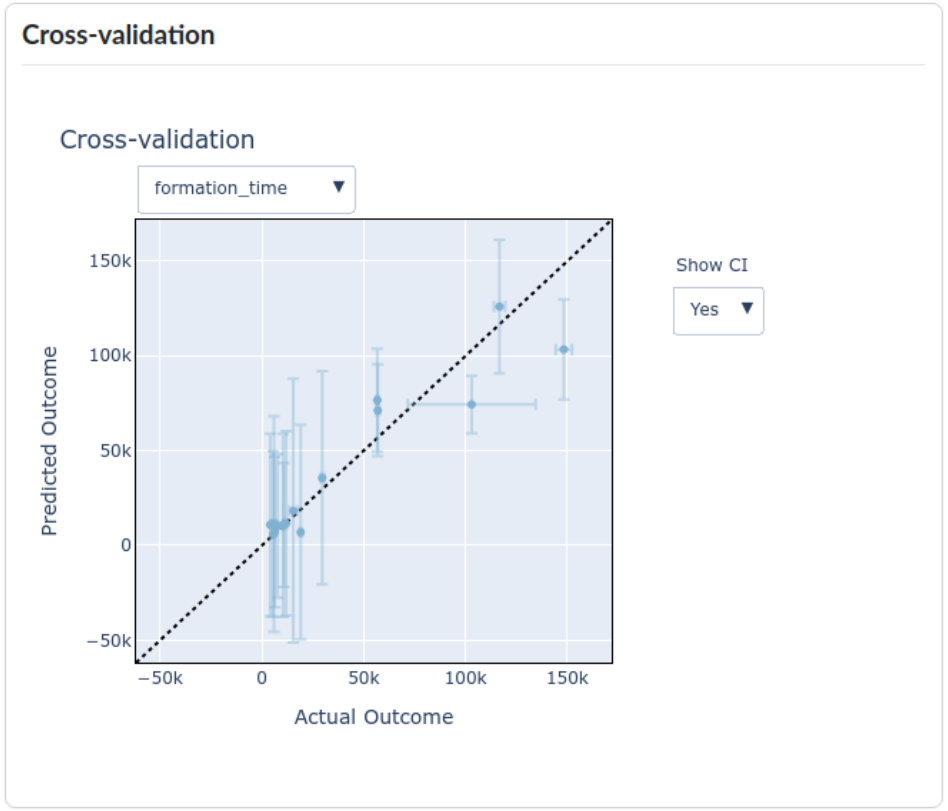
Human interface: dashboard

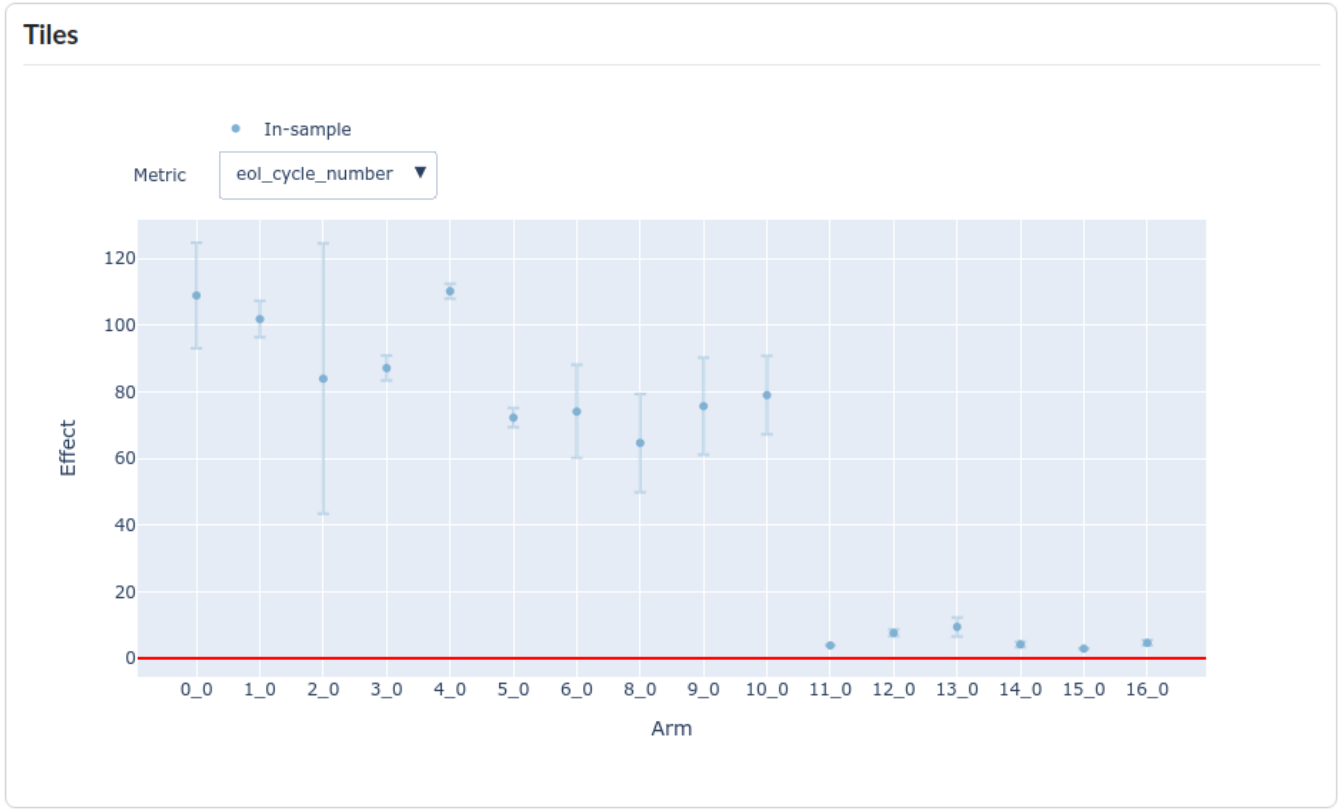


Plotly figure (lines)

- Use it as interactive figure for your project
- Here: model-based estimate of influence of design parameter on objective

Human interface: dashboard





Plotly figure (lines)

- Use it as interactive figure for your project
- ▶ Here: model performance and study progress

Conclusion

5

Conclusion

RDM

- Belongs to the everyday life of all researchers
- Is essential to make research data FAIR
- Promotes structured documentation and seamless provenance tracking
- Supports interoperability and reuse of data components in future workflows

Data-integrated AI

- Creates data-driven solutions for materials science
- Within KadiAI and CIDS, structures and accelerate ML research
- **KadiAIgent** plugin and workflow nodes for manual and automatized experiments (multiobjective) optimization

A systematic integration of data science tools through research data infrastructure **accelerates research**

Further Information

All information about Kadi4Mat can be found under

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

Thank you for your attention!