nffa.eu
PILOT 2021 2026

DELIVERABLE REPORT

WP16 - Implementing FAIR data approach within NEP

D16.6
Final report on data provenance tools for NEP community

Due Date
M48

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

PROJECT DETAILS

PROJECT ACRONYM	PROJECT TITLE
NEP	Nanoscience Foundries and Fine Analysis - Europe|PILOT

GRANT AGREEMENT NO:	FUNDING SCHEME
101007417	RIA - Research and Innovation action

START DATE

01/03/2021

WORK PACKAGE DETAILS

WORK PACKAGE ID	WORK PACKAGE TITLE
WP16	JA6 - Implementing FAIR data approach within NEP

WORK PACKAGE LEADER

Dr. Giuseppe Piero Brandino (eXact lab)

DELIVERABLE DETAILS

DELIVERABLE ID	DELIVERABLE TITLE
D – D16.6	Final report on Data provenance tools for NEP community

DELIVERABLE DESCRIPTION

This deliverable summarizes the experience and guidelines for the NFFA community in developing, deploying, and using data provenance tools. It is articulated in 3 sections, dedicated to 1) establishing the high-level provenance of the physical and digital entities of research workflows, 2) automatically integrating this into digital provenance tools, including electronic lab notebooks and workflow management systems, and 3) disseminating this in FAIR modes to the community at large.

DUE DATE		ACTUAL SUBMISSION DATE
M48 (Month)	28/02/2025	28/02/2025

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

AUTHORS

Nicola Marzari (EPFL and PSI) Rossella Aversa (KIT) Andrea Recchia (eXact lab) Giuseppe Piero Brandino (eXact lab), Valeria Granata (EPFL), Edan Bainglass (PSI)

PERSON RESPONSIBLE FOR THE DELIVERABLE

Prof. Nicola Marzari (EPFL)

NATURE

☒ R - Report

☐ P - Prototype

☐ DEC - Websites, Patent filing, Press & media actions, Videos, etc.

☐ O - Other

DISSEMINATION LEVEL

☒ P - Public

☐ PP - Restricted to other programme participants & EC: (Specify)

☐ RE - Restricted to a group (Specify)

☐ CO - Confidential, only for members of the consortium

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

REPORT DETAILS

ACTUAL SUBMISSION DATE
28/02/2025

NUMBER OF PAGES
21

FOR MORE INFO PLEASE CONTACT

Giuseppe Piero Brandino
eXact lab srl
via Crispi 56 Trieste
email: brandino@exact-lab.it

VERSION	DATE	AUTHOR(S)	DESCRIPTION / REASON FOR MODIFICATION	STATUS
1	27/01/2025	R. Aversa	High-level provenance	Draft
2	01/02/2025	E. Bainglass	ELN and WFMS	Draft
3	03/02/2025	V. Granata	Materials Cloud Archive	Draft
4	14/02/2025	N. Marzari, A. Recchia, G.P. Brandino	Revision	Final

CONTENTS

Introduction **5**
High-level provenance 5
The workflows and lifecycles 5
Infrastructure workflow scenarios 8
Provenance conceptual realization 9
Structure of the metadata schemas 9
Fine-level provenance: Electronic lab notebooks, workflow management systems, and their interoperability 12
One-to-one integration: AiiDAlab-openBIS demonstrator 13
Platform-agnostic methods: Overcoming the N2 problem 16
Open and FAIR Archiving of data 17

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

INTRODUCTION

This deliverable summarizes the experience and guidelines for the NFFA community in developing, deploying, and using data provenance tools. It is articulated in 3 sections, dedicated to 1) establishing the high-level provenance of the physical and digital entities of research workflows, 2) automatically integrating this into digital provenance tools, including electronic lab notebooks and workflow management systems, and 3) disseminating this in FAIR modes to the community at large.

High-level provenance

The high-level provenance information is intended to reflect the lifecycle of physical and digital entities collected during nanoscience and materials science research studies, from the fabrication of a material to the scientific publication, framed in the NEP management infrastructure designed in the Deliverable D16.1 [https://doi.org/10.5281/zenodo.6984650] and further refined in the following. The definition and the description of the used terms was formalized in the MDMC-NEP Glossary of Terms [https://doi.org/10.5281/zenodo.10663833].

The workflows and lifecycles

Figure 16.1 shows a basic experimental workflow, in which each process is performed only once, using one Technique. Multiple loops are possible, even on more than one System. Not necessarily all the illustrated processes apply to all cases. As a generic experimental workflow, one or more Inputs undergo a Fabrication to produce one or more Precursors, which undergo a Sample Preparation to produce one or more Samples. The output Sample, consisting of one or more Sample Components, undergoes a Measurement and Raw Data is produced. In any of these processes, the use of Consumables may be needed. A Measurement can also be performed during Fabrication or Sample preparation, using a Technique, which requires an Instrument and other Equipment, located in a Laboratory hosted by an Institution. The Settings describe any constant or varying controlled conditions of the Instrument, the Equipment, or the general environment in which the processes take place.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

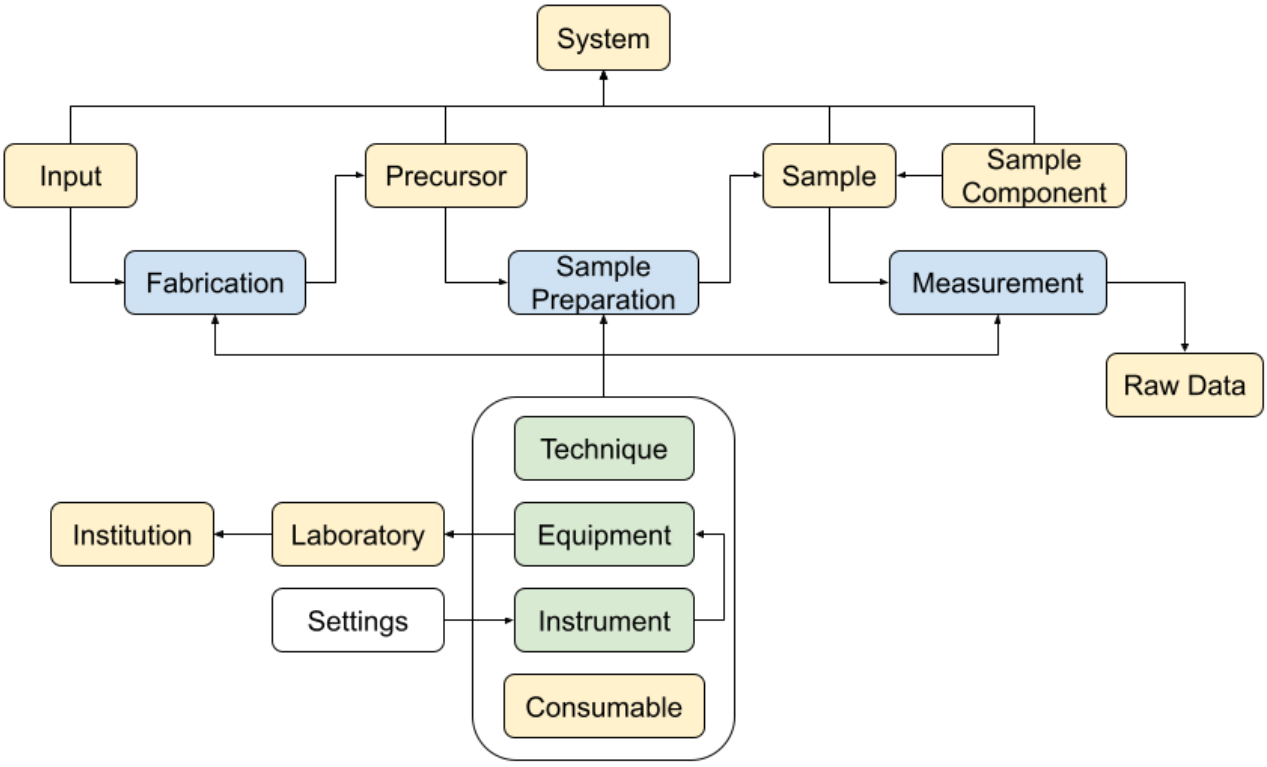


Figure 16.1: Experimental workflow. Green boxes: agents, yellow and white boxes: entities, blue boxes: processes. The rounded-corner box groups the terms connected to each of the processes, to simplify the graphical representation.

Figure 16.2 shows a basic computational workflow. Multiple loops are possible, as well as different combinations of Research Data, Systems and Techniques. As above, not necessarily all the illustrated processes apply to all cases. As a generic computational workflow, Research Data undergoes Data Processing in order to obtain Processed Data, which is then used to perform the Model Preparation. The resulting Model, optionally together with Processed Data, is the input of any type of Data Acquisition, which may be a Simulation or a Calculation, and Raw Data is produced. During any of these processes, a Technique is used, which requires a (usually virtual) Instrument (e.g., a computational resource or an HPC infrastructure) and possibly other Equipment, located in a Laboratory hosted by an Institution, and/or virtually or remotely accessed, if the Laboratory is a virtual place. Research Software is additionally included in Data Processing and Model Preparation, while it is not in Simulations or Calculations: any software used during Data Acquisition is considered part of the Instrument and should be described as such. The Settings describe any constant or varying controlled conditions of the Instrument, the Equipment, and the Research Software (whenever relevant) or the general environment in which the processes take place.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

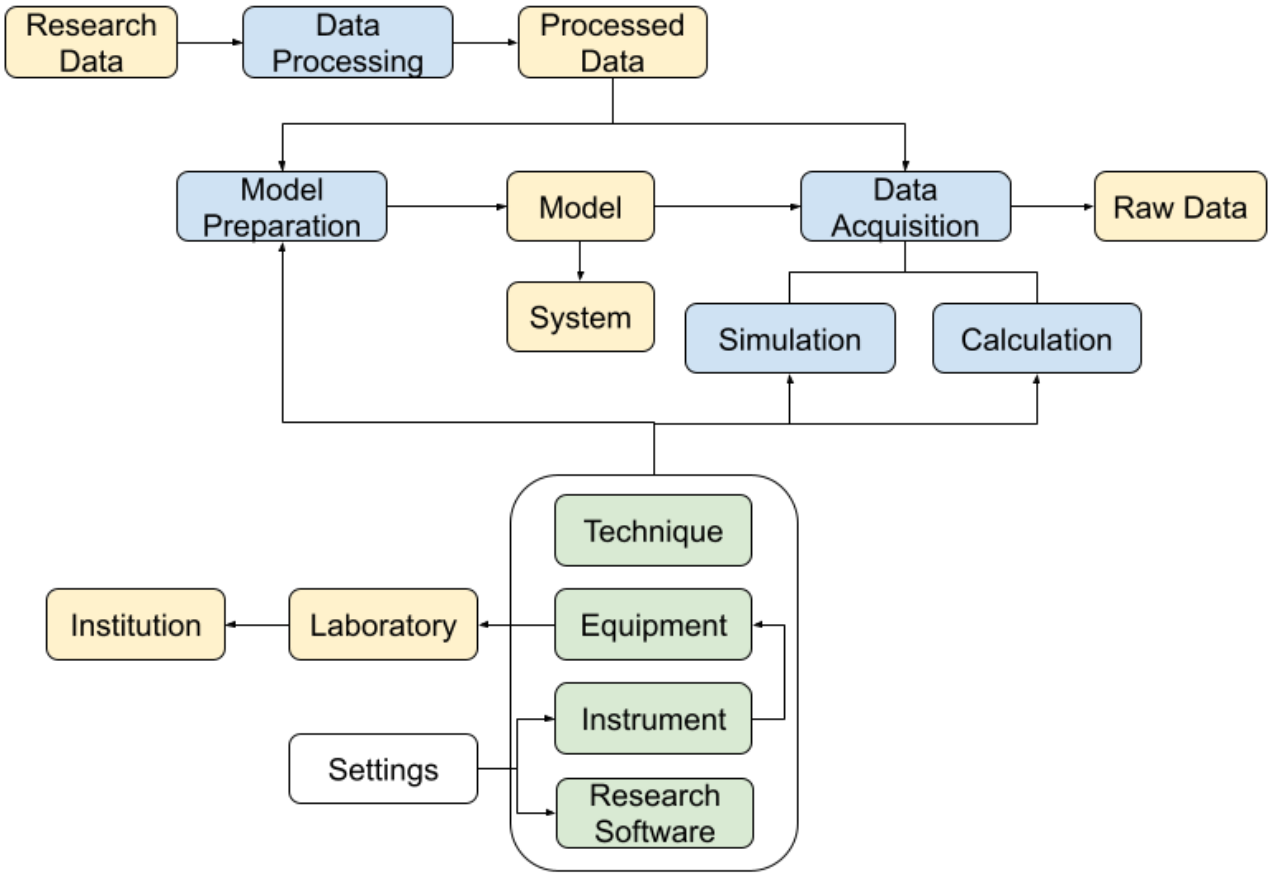


Figure 16.2: Computational workflow. Green boxes: agents, yellow and white boxes: entities, blue boxes: processes. The rounded-corner box groups the terms connected to each of the processes, to simplify the graphical representation.

Figure 16.3 illustrates the processes which may be performed in a simple Data Analysis Lifecycle. Multiple loops are possible, involving some or all the processes, even in a different order. A typical Data Analysis Lifecycle starts with the Data Processing of Raw Data to prepare it for one or more further processes. The Processed Data is usually an intermediate result, used as input of Data Analysis to obtain Analyzed Data. Data Interpretation is then performed on Analyzed Data to draw Conclusions that support the answer to some specific research question. The processes in the Data Analysis Lifecycle may require the use of Reference Data (as the main input data or for comparison) and the employment of Research Software, whose configuration parameters are described by Settings.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

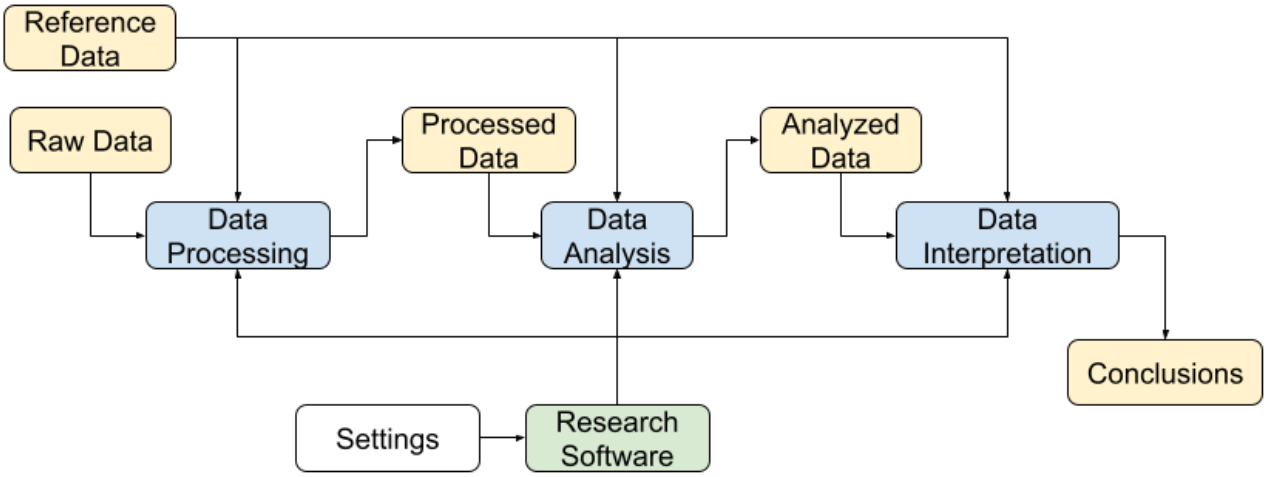


Figure 16.3: Data Analysis Lifecycle. Green boxes: agents, yellow and white boxes: entities, blue boxes: processes.

Infrastructure workflow scenarios

The scenarios described in this section and illustrated in Figure 16.4 assume at least a basic level of digitization. From a survey conducted in 2015 in NFFA-Europe and two follow-up surveys at summer schools in 2024, we know that the level of data management is very diverse among the laboratories. Some laboratories have implemented very sophisticated Electronic Lab Notebook (ELN) workflows, however many others don't use any ELN and write the procedures on paper. It is also a common practice for many users to save the produced data on personal USB drives or hard disks, rather than on an institutional data storage system.

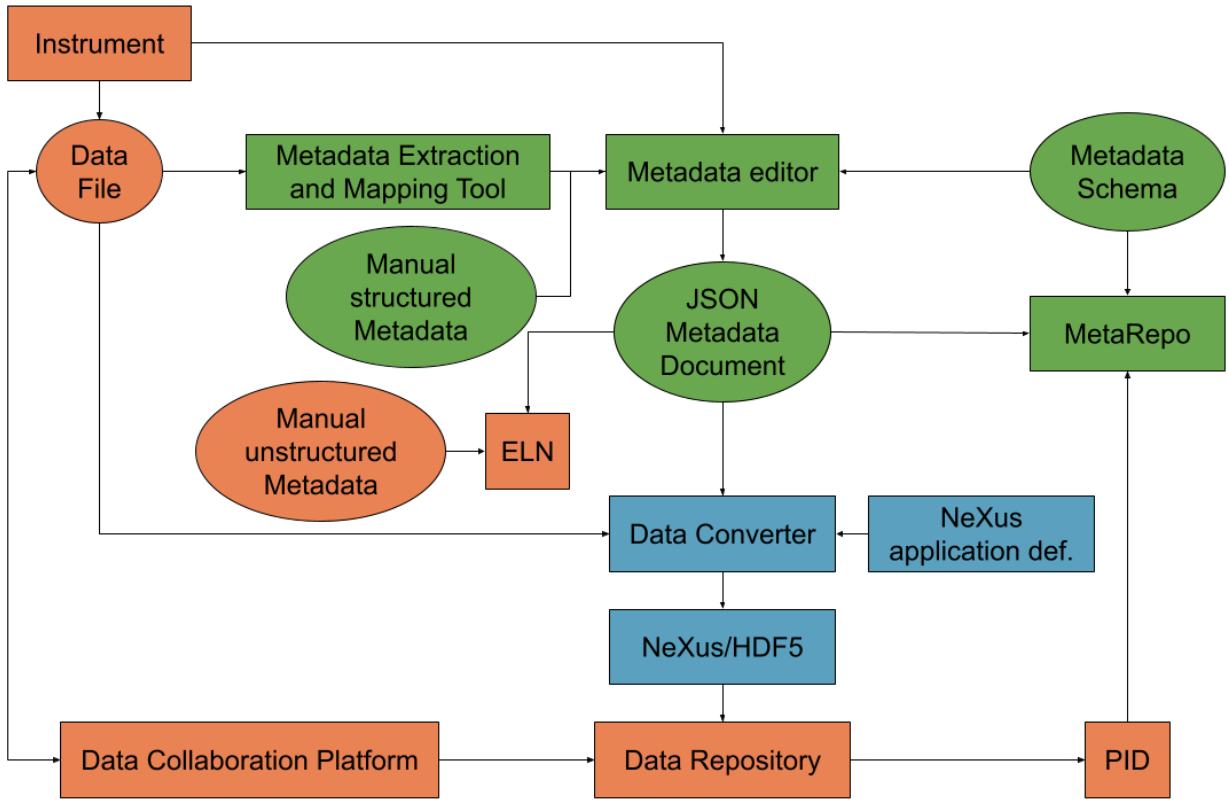


Figure 16.4: Updated infrastructure workflow, with respect to the one presented in D16.1

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

Depending on the starting level of the laboratories, different scenarios are possible.
The Beginner Scenario (orange boxes) can be recommended as the first transition from "analogical" to "digital". Even if this approach cannot be considered sufficient in a data management view, it's a good starting point in a "better than nothing" philosophy.
The Intermediate Scenario (green boxes) can be seen as a good approach to data management, and resembles what we defined as "basic scenario" in the Deliverable D16.1.
The Advanced Scenario (blue boxes) is the "ideal scenario" in the Deliverable D16.1, which includes optional features like the conversion of data and metadata to a standard format (NeXus).

The high-level provenance implementation comes into play from the Intermediate Scenario (green boxes) on, and involves the Metadata Schemas and corresponding Metadata Documents, as well as the Metadata Editor [https://metadata-editor.gitlab.io/documentation/], described in D16.4 [https://doi.org/10.5281/zenodo.10201228] and the MetaRepo [https://metarepo.nffa.eu/frontend/schema-management.html], described in D9.1 [https://doi.org/10.5281/zenodo.7702265] and D9.2 [https://doi.org/10.5281/zenodo.10156331].

Provenance conceptual realization

The suggested solution has been designed to be applied (at least) to NEP [https://nffa.eu/] and to the Joint Lab "Integrated Model and Data Driven Materials Characterization" (JL-MDMC) of the Helmholtz Association [https://jl-mdmc-helmholtz.de/] metadata, using MetaRepo [https://metarepo.nffa.eu/frontend/schema-management.html] as the metadata repository.
Taking into account:

- the experimental workflow in all its complexity
- how long a monolithic metadata document can be, if all details are inserted
- the attempt to describe the output of each process (see Figures 16.1, 16.2, 16.3) following the chronological order in which they are produced
- the possibility to adopt the same high-level solution for multiple projects rather than having many customized solutions
- the current level of development of metadata schemas for each step, as well as the variety of practices and repositories adopted in the institutes

we agreed that the most practical solution is to:

- have a metadata document record for each output of a process (see Figures 16.1, 16.2, 16.3)
- include the references to the previous step or to the needed input(s) as well as other related resources inside the metadata document as part of the schema
- in case of digital resources, like datasets, the relatedResource of the record must be the URL or the PID pointing to the resource
- in case of not-digital resources, like samples, the relatedResource of the record may be the URL or the PID pointing to the sample preparation in the ELN.

Structure of the metadata schemas

We created a modular solution, in which every output of a process (see Figures 16.1, 16.2, 16.3) is described by a metadata document record according to a metadata schema. The schema of the output contains the description of the step(s) needed to obtain it and the reference to the used

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

input(s) and to other related resources. This means that the high-level provenance is tracked "backwards". With this approach, there are no mandatory steps or inputs in the workflow. The only important thing is to relate a given output to the necessary input(s) to reproduce it.

Technically, the provenance is realized including an optional field called "parents" in every metadata schema. As shown in Figure 16.5, the "parents" field is an array of objects, allowing reference to more than one parent. The object itself contains three properties: the "parentType", to specify the type of which input was used during the process, the "parentReferenceType", to specify whether the reference (typically a URL) points internally to MetaRepo or to an external system, and the "parentReference", which is the actual URL to the parent resource.

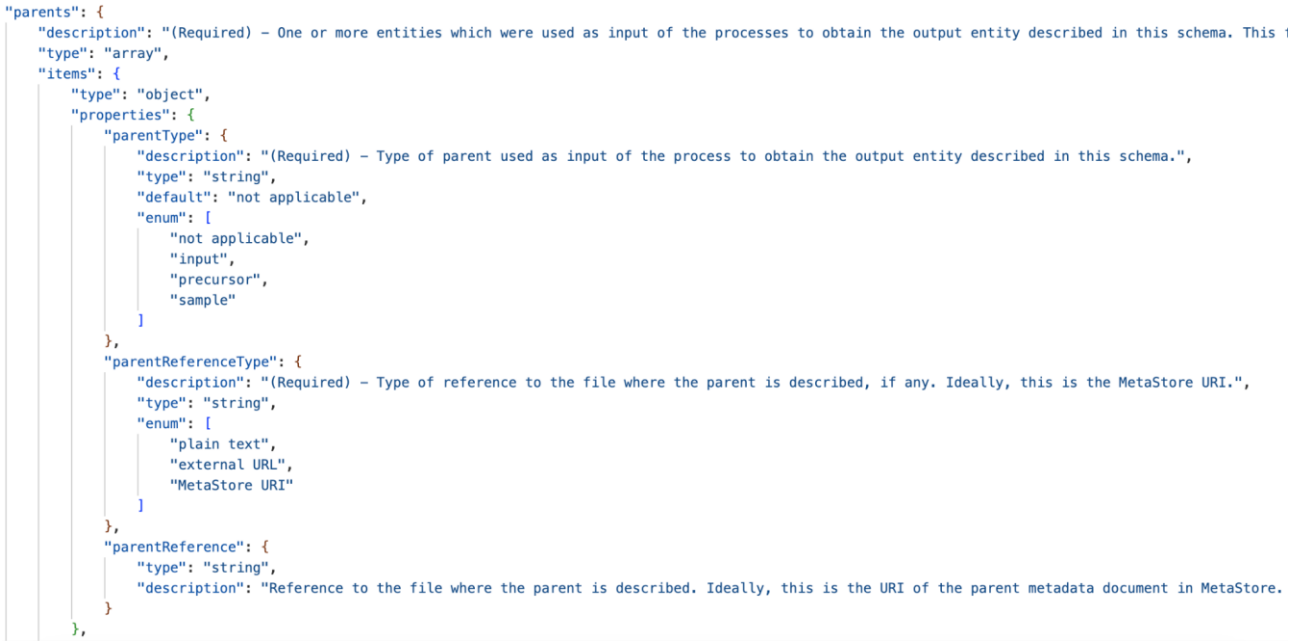


Figure 16.5: Example of the provenance realization in a metadata schema using the "parents" field.

The Metadata Editor allows authenticated users to automatically generate the full provenance of a metadata resource. To do this, the "parent" field must be either as a first level (if entry does not exist) or nested inside "entry" as a second level in the metadata schema. No further levels of nesting are possible. During the provenance generation process, the Metadata Editor looks recursively inside the metadata documents in the "parentReference" fields, provided that the "parentReferenceType" is "MetaStore URI". If this is not the case, the last available "parentReference" is tracked, but not recursively inspected, because it is not possible to guarantee that the metadata structure of the resource, if any, fulfills the requirements, i.e., it contains the "parents" field.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

```
{
    "d5cd029a-47e2-4ef6-8afb-7907e4cc4403":{
        "referenceType":"MetaStore URI",
        "reference":"https://metarepo.nffa.eu/api/v1/metadata/d5cd029a-47e2-4ef6-8
        "type":"sample",
        "parents":[
            {
                "2a13ccdd-fd33-4f49-bbc3-01d2a48580ff":{
                    "referenceType":"MetaStore URI",
                    "reference":"https://metarepo.nffa.eu/api/v1/metadata/2a13ccdd
                    "type":"precursor",
                    "parents":[
                        {
                            "6f09d66f-ef33-4ce4-9d8b-64dc6c9945b7":{
                                "referenceType":"MetaStore URI",
                                "reference":"https://metarepo.nffa.eu/api/v1/metad
                                "type":"input"
                            }
                        }
                    ]
                }
            }
        ]
    },
    "proposalIds":["120"]
}
```


Figure 16.6: Example of generated provenance document.

Figure 16.6 shows an example of a provenance document, provided as a JSON file. It consists of nested sections (as indicated by highlighted portions) that include details like “reference” , “referenceType” , Schema “type” (also referred to as label), and possibly “parents”, a nested array that in turn carries similar blocks of information describing the current block's parents. To make things clearer, the provenance document of Figure 16.6 is shown in Figure 16.7 as a graph, generated with JSON Crack [https://jsoncrack.com/].

This initiative has received funding from the EU’s H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

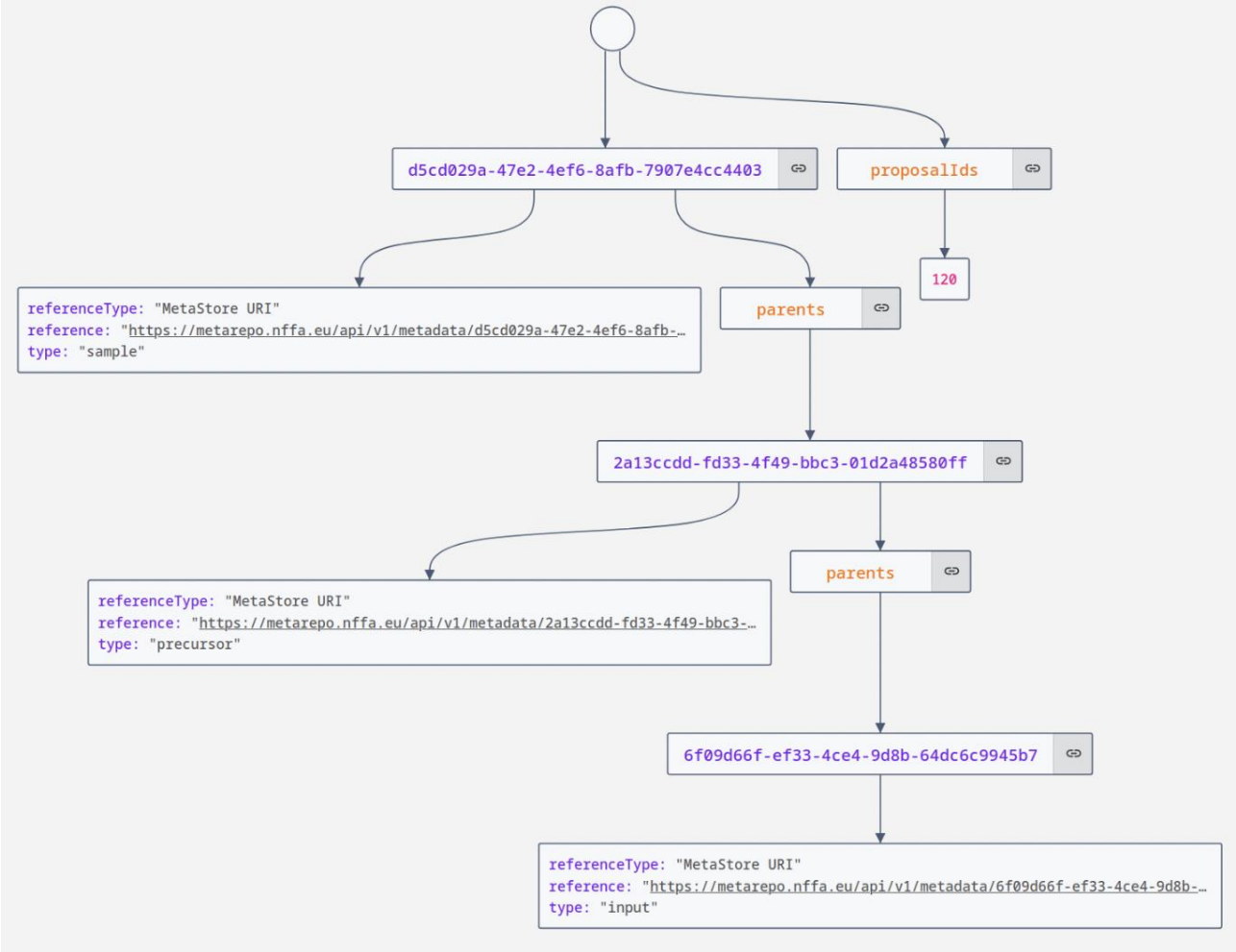


Figure 16.7: Structure of the generated provenance document, represented as a graph.

Fine-level provenance: Electronic lab notebooks, workflow management systems, and their interoperability

As discussed above, in the course of a scientific study, be it experimental and/or computational in nature, it is crucial that all experiment/simulation inputs, processes, and outputs are tracked for the sake of data provenance and reproducibility. Here, we discuss how to automate a fine-level provenance of all these entities through electronic lab notebooks and workflow management systems.

On the experimental side, ELNs often serve as the means of recording experimental samples, protocols, and outcomes. Workflow management systems, or WFMSs, though generally designed for orchestration, are often equipped to serve a similar role in computational studies.

In studies involving both experimental and computational components, as is often the case in autonomous laboratory conceptualization/implementation, the exchange of data across provenance-tracking platforms becomes an integral and essential component. To facilitate the

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

exchange, platforms may opt for dedicated one-to-one APIs, allowing each side to speak to the other. This approach benefits from full control of the API, affording it more power (can seamlessly handle features unique to the coupled platforms). However, it also requires the API to understand the specifics of each platform, in particular, their own dedicated APIs (see example below). Unfortunately, the conception and implementation of new one-to-one APIs scales as N^2 with the number of ever-growing ELNs and WFMSs. This motivates a new solution to the interoperability problem, one that is platform agnostic in nature.

In the following sections, we provide updates to the one-to-one interoperability API designed for research data exchange between the AiiDAlab GUI of the AiiDA WFMS and the openBIS ELN, highlighting the captured feature set with which a platform-agnostic API would compete. We then discuss the present state of a conceptual approach to platform-agnostic exchange of research data. Both efforts are in active development as deliverables of the ORD-PREMISE Establish project of the ETH Board ORD program, and the MADICES workshop series it sponsors.

One-to-one integration: AiiDAlab-openBIS demonstrator

Building on the integration of computational simulations into LIMS/ELNs reported in D16.3, we provide here updates to the AiiDAlab-openBIS integration developed in the ORD-PREMISE project (WP2) to facilitate the seamless integration of experimental and simulation data.

In the present iteration, to reduce the burden of adoption, significant efforts have gone into developing GUIs directly within AiiDAlab or on a dedicated platform enabling researchers to access both the ELN/LIMS and the WFMS with minimal user input. Within these GUIs, access to inventories of molecules and substrates defined in the openBIS LIMS empowers different researchers to initiate predefined simulation workflows in AiiDAlab. From the GUI, users can automatically upload simulations to the LIMS. On upload, the GUI automates the process of establishing parent/child relationships with the LIMS digital twins of the molecules and substrates.

Every researcher with access rights to a specific project can efficiently inspect the list of simulations and experiments available for that project, linked to specific inventory entries. The entire data provenance - from a selection of results, up to the experiment/simulation steps, and inventory entries - can be exported in machine-actionable formats, such as JSON-LD, JSON Schema, and LinkML, following the metadata schema format defined in the ORD-PREMISE project (WP2).

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

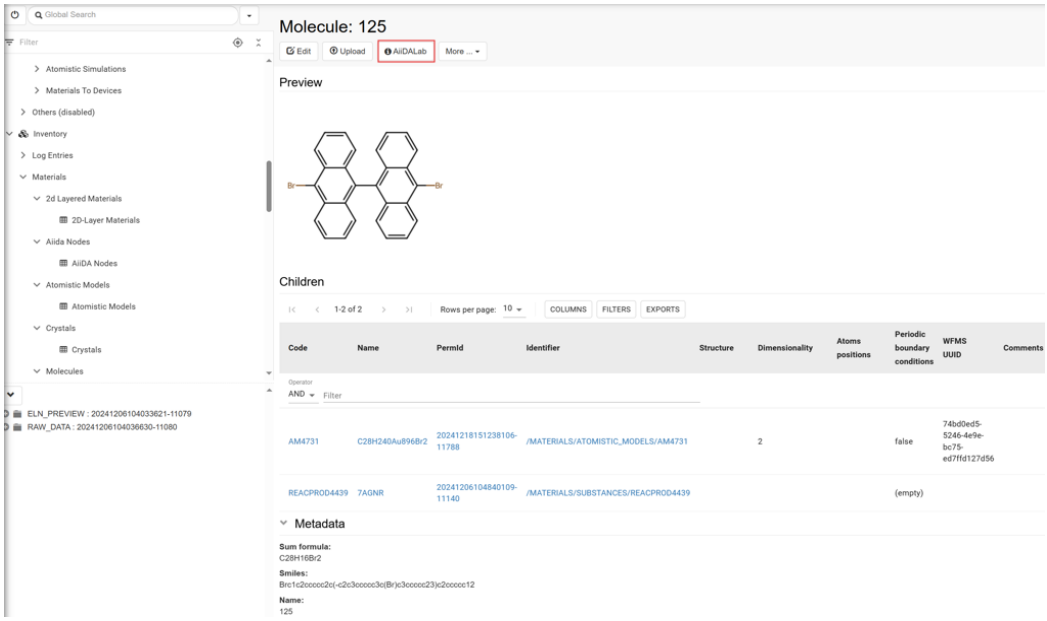


Figure 16.8: The openBIS LIMS, from which users can send objects of interest to AiiDAlab via a dedicated control (highlighted in red).

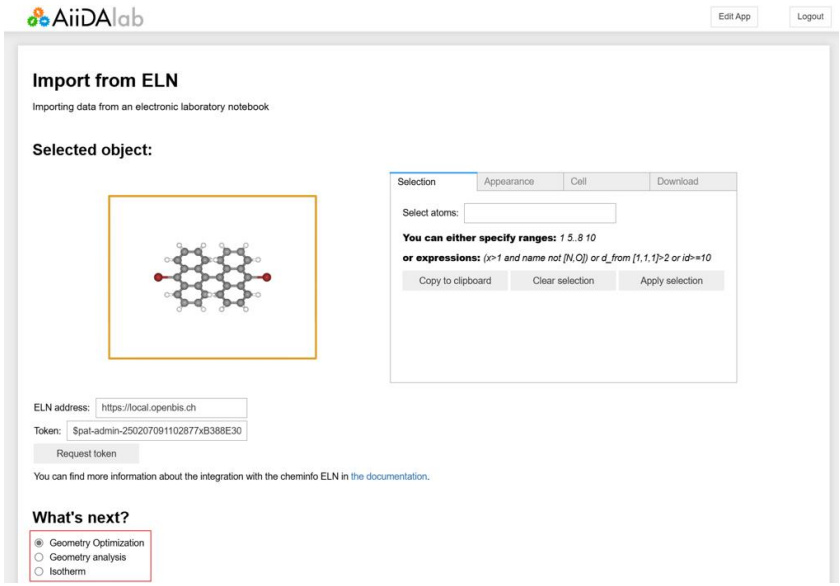


Figure 16.9: A dedicated AiiDAlab app receiving objects from openBIS, in which the user can select to perform one of the available simulations (highlighted in red).

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

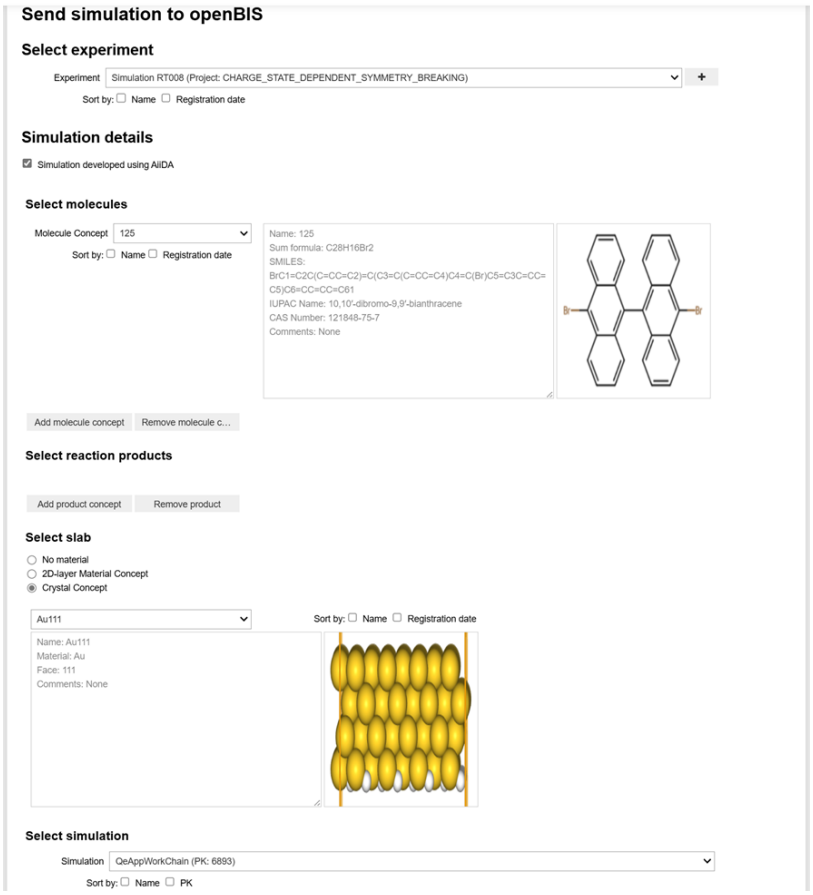


Figure 16.10: From the AiiDAlab GUI, users can select objects from the available inventory to send to openBIS.

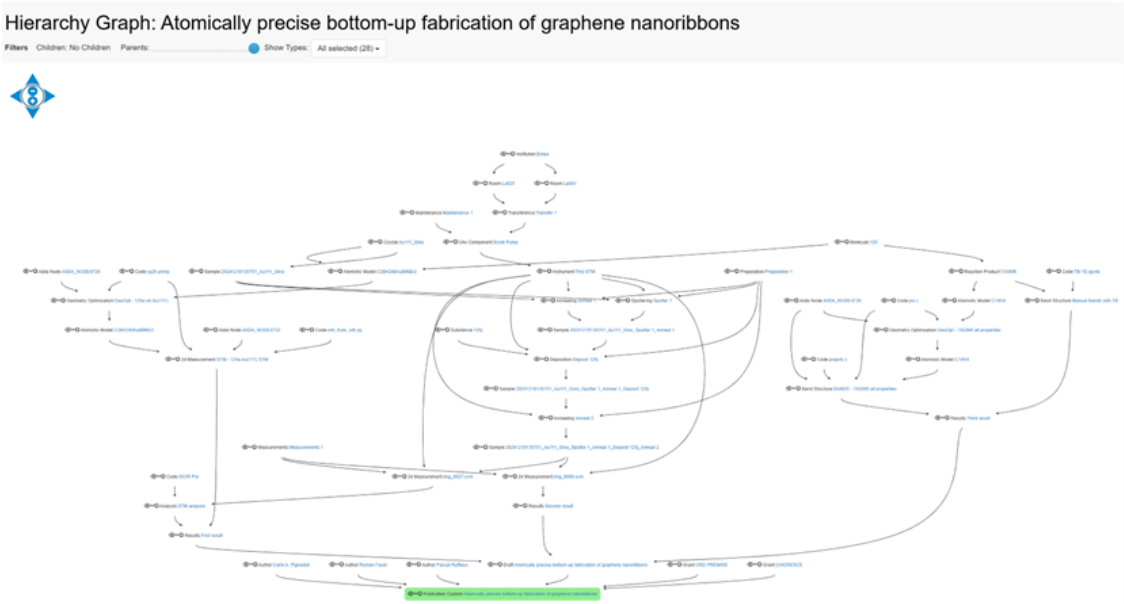


Figure 16.11: In openBIS, users can access the full provenance of the research study (in green, we highlight a "Publication" entry).

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

Platform-agnostic methods: Overcoming the N^2 problem

As stated above, dedicated one-to-one APIs, though powerful, come at an implementation cost repeated per interoperability partner. To mitigate, or perhaps avoid the cost altogether, it is desired to conceptualize a single API leveraging common protocols and standards for data exchange that can be implemented locally in each interoperating platform.

Such an API was the focus of a working group at the MADICES 2 workshop, sponsored by CECAM and the ORD-PREMISE project, for which a summary of results has been made available. The focus of the working group was to develop a foundation for the API in the form of specifications of transactional units. The efforts leveraged heavily on the use of semantic annotation and ontological concepts, experts of which were in attendance, to annotate data schemas. Briefly, the concept relies on semantic annotations to provide unique identifiers in the form of internationalized resource identifiers, or IRIs, to link schema fields to ontological concepts that can be made common across the interoperating spectrum, or at least subsections (or domains) of the spectrum. Through a set of tools to prepare, parse, and interpret semantically annotated schemas, we aim to form a single API to which each platform interested in data exchange can integrate following our protocols.

In the present stage, we are leveraging the JSON-LD-based RO-Crate specification to form a basis for our transactional unit. The crate manifest serves as a clear specification for the included data and metadata being exchanged, with the metadata semantically annotated following the JSON-LD specifications to convey the meaning of the data to be interpreted at the receiving platform, which can be consumed by platform-specific APIs. We provide a tutorial on our ORD-PREMISE YouTube channel explaining the concept and demonstrating a hypothetical use case using mock web apps developed at MADICES 2. A screenshot of the web app is also provided below.

To facilitate the integration of data and metadata in a receiving platform, the object (class) schema is also required. Here, JSON Schema can be used to achieve this. Alternatively, the metadata schema can be provided using the LinkML format, which provides means of combining both the object schema and semantic annotation. LinkML also provides a set of tools to convert between the various formats, including pydantic model classes for use in Python APIs. These formats are actively being explored and tested for their usability in data exchange and completeness w.r.t the various domains participating in the MADICES workshops.

In the upcoming MADICES 3 workshop to be held at the Paul Scherrer Institute (PSI) in October 2025, we will continue to build on the efforts and successes of MADICES 2 in establishing and publishing specifications for the protocols to be used in our API. For the upcoming workshop, we will have additional WFMS representation beyond AiiDA, including pyiron, Galaxy, and PerQueue. Lastly, we have identified members of the interoperability community involved in metadata annotation training and education, as well as in the application of AI/ML in collecting and annotating metadata. We believe expanding our community to include such efforts will accelerate development and adoption of our interoperability protocols and methods.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

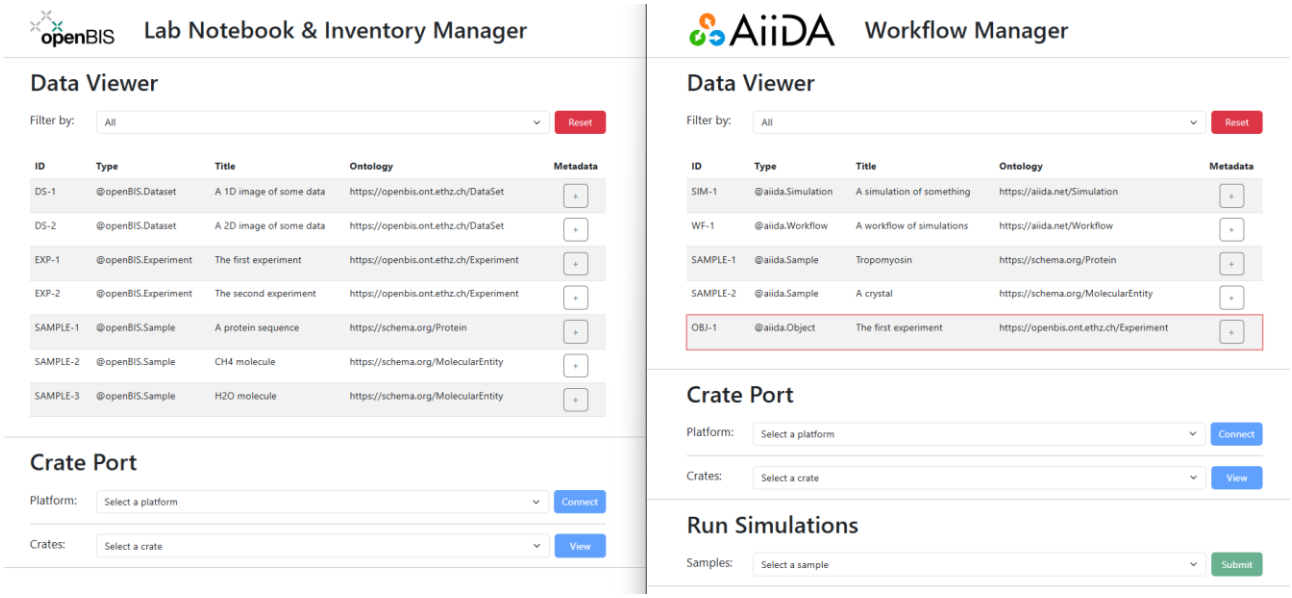
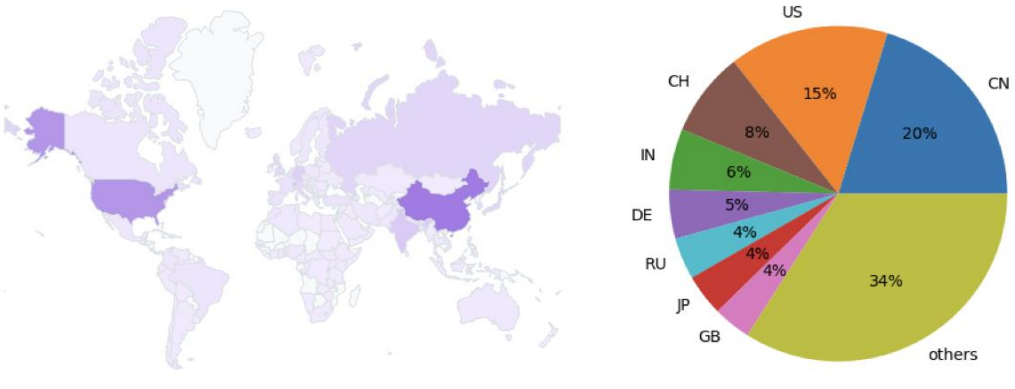


Figure 16.12: A screenshot of web apps mocking openBIS and AiiDAlab. The web apps provide means of exploring existing objects, querying the partner platform for objects, and exchanging objects. Annotated metadata schemas are included in the exchanged RO-Crates and can be viewed directly in the app. The schemas can be inspected against those of existing objects to verify full and accurate comprehension of the provided semantic annotations. See tutorial for more details. Here we highlight (in red) the OBJ-1 object obtained from the openBIS mock web app denoted by its unique ontological IRI (https://openbis.ont.ethz.ch/Experiment).

Open and FAIR Archiving of data

The Materials Cloud Archive (https://archive.materialscloud.org) has been developed as a moderated digital repository designed for storage and dissemination of research data in materials science. Launched to promote open access and enhance collaboration, the Archive provides a platform where scientists can share their datasets, simulations, and software. It is one of the two recommended repositories for materials data by the European Commission: https://open-research-europe.ec.europa.eu/for-authors/data-guidelines

The Materials Cloud Archive hosts more than 1100 records and has more than 1000 registered users. The repository is used worldwide, and has an average number of visits per month of ~2500.



This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

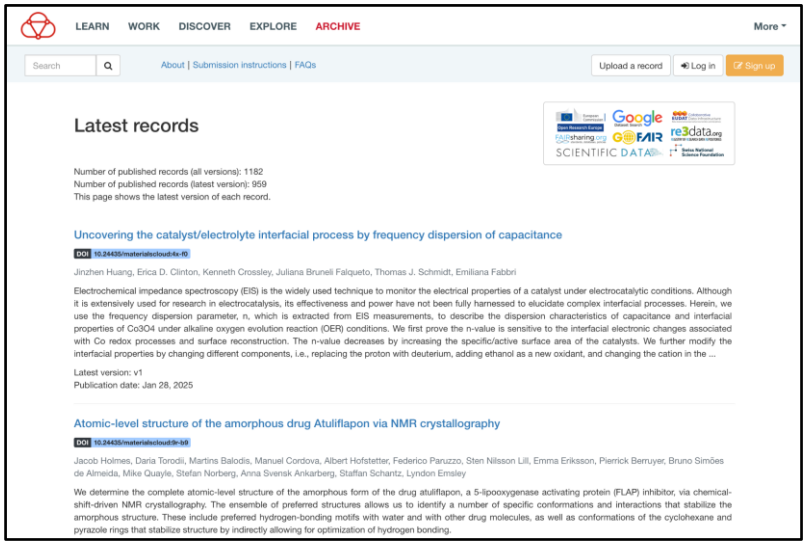


Figure 16.13: The Materials Cloud Archive is a moderated repository for computational materials science, used worldwide and hosting more than 1,100 records.

The first release of Archive consisted of an application written in-house using Jinja, Flask, and SQLAlchemy. Later, a new version was released, the Archive underwent a significant transformation, being completely rewritten with the Invenio (v3) framework, which is an open-source software created at CERN for building and managing digital repositories. Invenio is a framework composed of several packages (or modules) that can be used to build a repository. The packages offer an extensive array of tools and features for an effective storage, organization, and dissemination of digital content.
Currently, the Archive is undergoing a third major upgrade, transitioning from Invenio to InvenioRDM. In contrast to Invenio, which functions as a software library offering the framework and components for creating a digital repository, InvenioRDM is a complete repository application on its own. InvenioRDM enables most of its frontend and backend functionalities to be customised to meet the specific needs of the organisation or institution that manages the repository. This approach has the advantage of leveraging the framework's built-in features, minimizing development time while also providing the flexibility to customize the application.

The transition to InvenioRDM brings several new functionalities, including the release of a DOI (Digital Object Identifier) to represent all versions of a record, secret links to provide read or write access to a record before publication, usage statistics through tracking view counts and file downloads for each record, and importantly, APIs for all the repository operations. We describe below some of them in more detail.

The Archive already supports versioning of records and issues a DOI for each version of a record; the new Archive will also release a DOI to represent all versions of a record, known as a "canonical" DOI. The canonical DOI always resolves to the latest version of a record, making it useful for citing the record in a paper, particularly when regular updates and new versions are expected to be created.

Users often need to grant access to a record for review, feedback, or approval from colleagues before publishing it. The repository allows users to create links to records that can be distributed to

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

colleagues or referees. These links can provide either read or write access to the record and an expiration time can be set to make the links no longer accessible after a specified date.

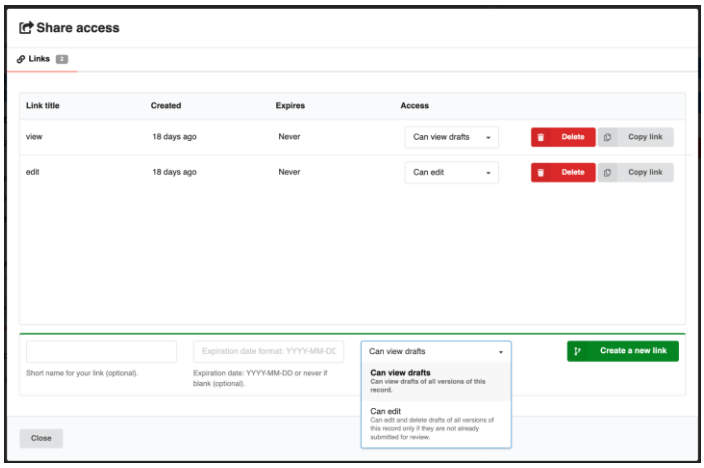


Figure 16.14: Links with tokens can be created by the owner of the record to grant either read or write access to the record before it is published.

InvenioRDM fully supports the DataCite Metadata Schema 4.5. Each metadata field of a record, such as title, description, authors, licenses, references, and funders, is recorded in the application database using a metadata schema compatible with DataCite. This enables the application to directly call the DataCite APIs to release the record DOI (and its canonical DOI). Once the DOI is released, the record metadata is accessible not only via the Archive application APIs but also through the DataCite APIs, making the record more easily findable.

By relying on standard metadata schemas to store the uploaded information, the Archive is easily interoperable with other systems, repositories, and applications.
The interoperability of the Archive is indeed a key concern, and significant efforts have been made to enable the use of datasets uploaded to the Archive in external applications. The Archive supports integration with some tools for data analysis and visualization, ensuring that researchers can not only store but also easily explore and use the shared resources.
Datasets uploaded to the Archive can be used by applications such as Renku, Chemiscope, and Optimade by simply uploading files in a specific format or with a designated filename. Once the record is published these external applications can be accessed with a simple click directly from the record.
Moreover, datasets and databases uploaded to the Archive can be used in other sections of Materials Cloud, such as the Discover or Explore sections, or by software tools in the Work section. As an example, the Materials Cloud MC2D and MC3D, which are part of the Discover and Explore sections, use databases uploaded to the Archive to allow users to easily browse and visualize crystal structures.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

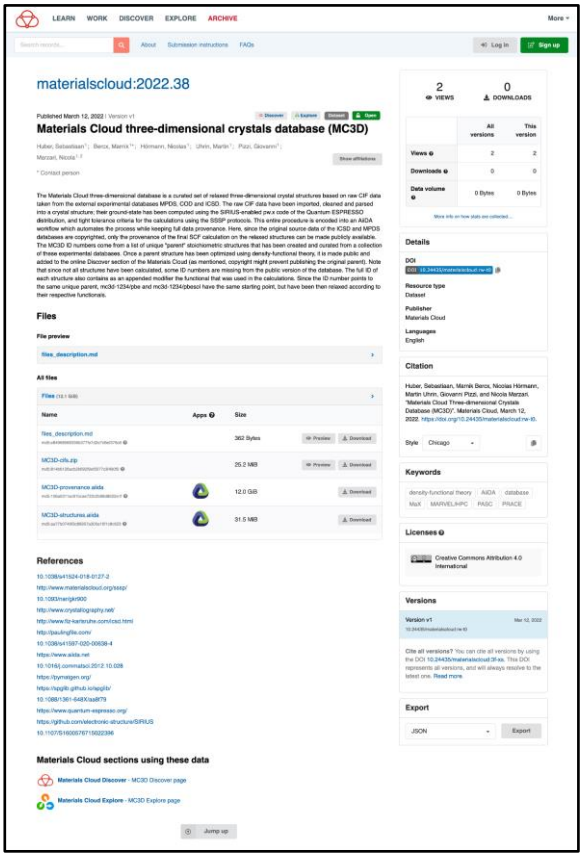


Figure 16.15: The record clearly displays the links to the Explore or Discover sections, or to other external applications, when the datasets of the record are used in these sections or applications, guaranteeing easy access to these resources.

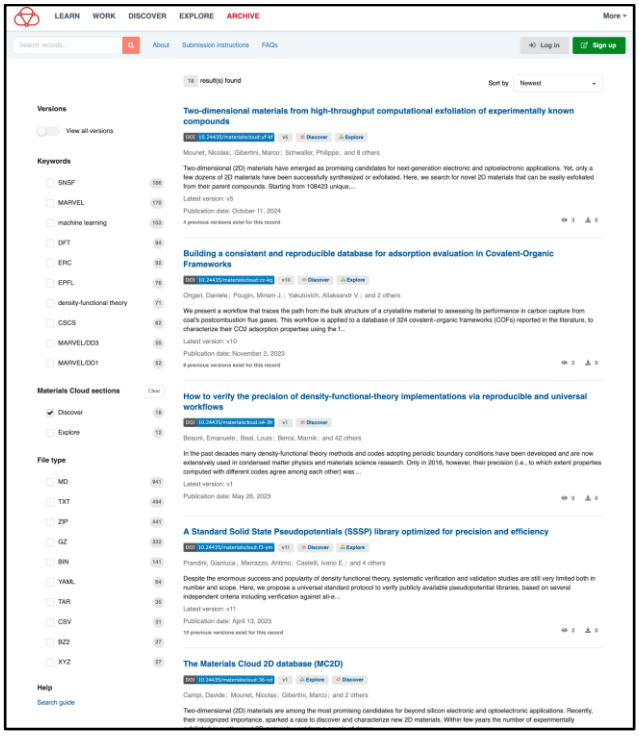


Figure 16.16: The powerful search engine of InvenioRDM, based on OpenSearch, allows users to easily search for records by words and filter results using facets to further refine their research.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project

As mentioned above the new version of Archive has APIs available for all operations on the repository. This will allow users to create, update, delete drafts and upload and download files not only via the user interface but also using scripts to access the application APIs. This is particularly important for automation, enabling users to integrate the Archive's functionalities into their workflows and facilitate the data and metadata upload and download without the need for manual intervention.

The transition from Invenio to InvenioRDM has taken several months. Efforts have been made to customize the upload form, the record landing page, the front page, and the user dashboard, and the process is now in its very final stages. The upload form has been cleared of metadata fields that are not required for the specific use of the Archive. Some functionalities have been disabled, such as the embargo feature and the ability to create communities. Other functionalities have been extended, such as the moderation of each version of a record. InvenioRDM comes by default with a moderation workflow that is limited to the first version of a record. When a record is created, the user must submit it for review to the Materials Cloud Archive moderators; the record will be published only after the corrections and changes suggested by the moderators have been made. By default, InvenioRDM allows the publication of any subsequent version of the record without requiring submission to the moderators. This feature has been modified to ensure that each version is moderated instead. These changes to the default functionalities required changes to the permissions settings on both the record and uploaded files.

Scripts have been written to migrate the metadata from the Invenio database to the InvenioRDM database, as the two differ in the way published and draft records are stored, and additional tables are present in the InvenioRDM database to manage the moderation workflow and the user requests to review records. The deployment of the application also required finalizing the Ansible scripts to deploy the application on virtual machines at the CSCS Castor computing cluster. Backup scripts were created to back up the database and OpenSearch indexes, ensuring that a copy of the uploaded files is available in locations other than the CSCS. Finally, work is ongoing to modify the setup of the current read-only failover server to adapt it to the new Archive using InvenioRDM. Importantly, the Materials Cloud Archive has nightly synchronization with the CINECA supercomputing center in Italy and the Juelich Supercomputing Center, in addition to a failsafe backup page at AWS, insuring safety and sovereignty of data.

This initiative has received funding from the EU's H2020 framework program for research and innovation under grant agreement n. 101007417, NFFA-Europe Pilot Project