



nffa.eu

PILOT 2021 2026

DELIVERABLE REPORT

WPN 16 WP Title JA6 - Implementing FAIR data approach within NEP

D16.3

Identification of good practices for data provenance

Due date

M18



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

16

JA6 - Implementing FAIR data approach within NEP

WORK PACKAGE LEADER

Dr. Giuseppe Piero Brandino (EXACT Lab SRL)

DELIVERABLE DETAILS

DELIVERABLE ID

DELIVERABLE TITLE

D – D16.3

Identification of good practices for data provenance

DELIVERABLE DESCRIPTION

Here we elaborate and implement FAIR-oriented procedures and recommendations to enforce data provenance in the NFFA scientific experiment's workflow, from data creation to data usage. The set of procedures is developed by taking into account needs coming from various communities within NEP. Close attention is paid to identify and tailor existing electronic lab notebook (ELN) and laboratory information management system solutions for describing sample processing workflows and (semi-) automated metadata recording during the experiments as initial steps for implementing FAIR by design datasets.

DUE DATE

ACTUAL SUBMISSION DATE

M18 (Month) 31/08/2022

01/09/2022



AUTHORS

Dr. Aliaksandr Yakutovich (EPFL), Dr. Giovanni Pizzi (EPFL), Prof. Nicola Marzari (EPFL), Dr. Rossella Aversa (KIT), Dr. Giuseppe Piero Brandino (EXACT Lab SRL), Dr. Mirco Panighel (CNR-IOM)

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



REPORT DETAILS

ACTUAL SUBMISSION DATE

01/09/2022

NUMBER OF PAGES

16

FOR MORE INFO PLEASE CONTACT

Nicola Marzari
 EPFL STI IMX THEOS
 Station 9
 1015 Lausanne

email: nicola.marzari@epfl.ch

| VERSION | DATE | AUTHOR(S) | DESCRIPTION / REASON FOR MODIFICATION | STATUS |
|---------|------------|---------------|---------------------------------------|---------------|
| 1 | 01/08/2022 | N. Marzari | First draft | Draft |
| 2 | 01/08/2022 | A. Yakutovich | ELN comparison | Draft |
| 3 | 01/08/2022 | G. Pizzi | Simulation tools | Draft |
| 4 | 31/08/2022 | All | Finalizing the document | Final version |

CONTENTS

| | |
|--|----|
| Introduction: Why data provenance? | 5 |
| ELN vs LIMS | 5 |
| Comparison criteria | 6 |
| Comparison of available tools | 7 |
| OpenBis | 7 |
| Cheminfo | 8 |
| Labfolder | 9 |
| eLabJournal | 9 |
| Rspace | 10 |
| eLabFTW | 11 |
| AiiDA | 12 |
| Integration of computational simulations | 13 |
| Conclusions | 15 |



INTRODUCTION: WHY DATA PROVENANCE?

Scientific discoveries are always made based on previous findings. It is impossible to move further without a good foundation. Herein, it is of prime importance to deliver a clear and complete description of an experiment, letting others repeat it, test it and evolve it further. Until recently, experimental research was recorded in paper laboratory notebooks, making it not searchable, hard to back up, and hard to share. Electronic laboratory notebooks do not have those flaws, and add even more features on top: simple embedding of accompanying data such as images and videos, worldwide access anytime, and the ability to check things from a smartphone.

In scientific publications, scientists tend to report only positive results, discarding a considerable amount of unsuccessful attempts that remain in the laboratory's logbooks. But those data can be used to predict the outcome of similar experiments. For the machine-learning purposes negative results are of equal importance to positive ones. However, data management is often considered by scientists as unnecessary burden that brings them away from their research. As a result, it is often done very poorly as an afterthought. For solving this problem, automated data management needs to be implemented. Thus, employing ELNs can facilitate data organisation simplifying its further use in machine-learning applications.

Finally, it is good to highlight some practical aspects of the ELN use that make working in a lab more efficient. It saves a lot of time to decode handwriting, as it is not always obvious to read other people's writing. ELN can be integrated with the other tools ranging from lab equipment for automated experiment recording, to simulation platforms to pair experimental results with predictions from theory. The use of ELN makes it easier to write publication or thesis, as it is much easier and faster to retrieve the details of a specific experiment. Last but not least, the data security aspect: it is unlikely to lose the data managed by ELN thanks to the version control and traceability.

ELN vs LIMS

To go further, the difference between electronic laboratory notebooks (ELN) and laboratory management systems (LIMS) must be clarified. According to Dirnagl and Przesdzing, ELNs and LIMSES [<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722687.1/>] differ by their features. A typical ELN allows:

- Share data among users
- Attach files to notes
- Visualize the attachments
- Complex rights management
- Automated storage of the previous note versions



- Provide API for customization

LIMS include all the features of an ELN, but add on top:

- Inventory management
- Workflows for certain samples, tasks, experiments
- Direct link to laboratory equipment
- Analysis of raw data within the system
- Data mining (aggregate and cluster structured data)

Therefore, LIMS can be considered an extension of ELN. For simplicity, further in the text we use the term ELN that represents both ELN and LIMS.

Comparison criteria

Switching to ELN by institutions automatically imposes requirements on the software. It must be versatile enough to be adapted by different groups that have different requirements. There are several choices on the market since the amount of ELNs has grown dramatically in the past decade. In 2011 Rubacha et al. reported 35 products [<https://journals.sagepub.com/doi/full/10.1016/j.jala.2009.01.002>], while in 2017 Kanza et al. identified 72 [<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0221-3>]. A recent study by Higgins et al. [<https://www.nature.com/articles/s41596-021-00645-8>] shows that up to now there were 172 ELN products, where 96 are still active and 76 are defunct. This amount makes it rather hard to choose the most suitable product for the needs of a particular laboratory/institution. Fortunately, several online resources have already made such a comparison. Based on the literature research we have identified the following comparison criteria:

Price. Despite one might think that free solutions are preferable, they typically have a significant drawback - lack of support. To adopt such free solutions institutions are required to have a support team "in house". Some institutes go even further and invest in the development of their own ELN system, like openBIS [<https://openbis.ch/>] at ETHZ and cheminfo [<https://cheminfo.github.io/>] at EPFL. Herein, open source solutions cannot be considered free of charge due to their hidden costs. At the same time, most public institutes tend to keep an IT team that manages the IT infrastructure. It is quite likely that the free option is often preferred by them.

Data security/Hosting. This point is related to the price criteria but has a separate aspect. Despite commercial solutions typically come with better support, they have some risks that can't be ignored. Keeping the data in proprietary formats creates difficulties in exporting them and, possibly, migrating to other ELN solutions. It can cause significant financial and scientific losses, and, thus, must be considered with great caution. In addition to that, quite commonly the national funding agencies require institutions to keep the data on the servers within the country. It is important that research groups can freely choose where to keep the data.

Versatility. Most of the ELN systems available on the market are for chemistry, which makes them inapplicable for other fields of science. Ideally, there should be only one ELN that covers all the use cases. That would simplify maintenance, support, data exchange, and facilitate adaptation by different groups.

API. The extended and well-defined application programming interface is a key factor for integration with the other services/tools.



User interface. Clear and attractive interface is always a plus, leading to a quicker adaptation by the users.

Inventory management. Previously, inventory management was decoupled from ELNs, but many start to support them now. Inventory management is an essential feature for a system to be classified as LIMS.

Comparison of available tools

Taking into account the amount of the available tools in the market, it would be prohibitively complex to compare all of them in this report. Moreover, such comparisons do exist already [<https://zenodo.org/record/4723753#.Yu6hYOxBwdQ>, <https://www.data.cam.ac.uk/data-management-guide/electronic-research-notebooks/electronic-research-notebook-products>, https://en.wikipedia.org/wiki/List_of_electronic_laboratory_notebook_software_packages, <https://datamanagement.hms.harvard.edu/news/finding-right-electronic-lab-notebook-core-lab>]. Based on those we had selected 6 tools for the final run of comparison 3 open source and 3 commercial (2 of which support freemium model): openBIS [<https://openbis.ch/>], Cheminfo [<https://cheminfo.github.io/>], Labfolder [<https://www.labfolder.com/>], eLabJournal [<https://www.elabnext.com/products/elabjournal/>], RSpace [<https://www.researchspace.com/>], eLabFTW [<https://www.elabftw.net/>]. In the table reported in the Conclusion paragraph, we compare the features of the selected notebooks according to the criteria identified in the previous section.

OpenBIS

OpenBIS is an ELN developed by the Scientific IT team at ETHZ. It is in active development since 2007. OpenBIS is 100% open source and distributed under Apache License 2.0. It has all the features of a modern LIMS: inventory management, laboratory notebook, data management. It is fully compatible with FAIR data principles.

OpenBIS comes with Java and Javascript APIs. Developers also provide a pybis tool capable of doing requests to the server from a Python environment. The Scientific IT team that develops OpenBIS provides a good level of support to the institutions that use the software. OpenBIS has a plugin system that allows extending the core functionality of the platform tailored to the needs of a specific research field. The platform enables collaborations on different projects, managing access rights among different members of the team.



Figure 1. Front-end screenshot of openBIS

Cheminfo

The Cheminfo ELN integrates tools from the Cheminfo ecosystem into a fast and lightweight laboratory notebook that runs in browser. Currently, around 100 tools can be found in the user interface. Every tool operates on a sample allowing for chemical characterisation, NMR spectra prediction, image analysis and many more. Such a design allows to quickly adapt to new requirements, implement new tools, etc.

Figure 2. Front-end screenshot of Cheminfo



Cheminfo is primarily developed at EPFL and mostly serves the needs of chemistry labs. It is 100% open source, most of the tools are under MIT license. The data are stored according to the FAIR principles. Integration with the other Python-based tools can be done using the *cheminfopy* package. Cheminfo is not general purpose ELN, as it is developed for chemistry.

Labfolder

Labfolder is web-based ELN that was founded in 2012 in Berlin, Germany. The software is closed-source, but it has free option for the groups of up to 3 people. Customers can choose whether to use the cloud-based version managed by the company, or to deploy Labfolder on premises. It is fully compatible with FAIR data principles.

Labfolder provides RESTful API designed for data exchange with the Labfolder ecosystem. Labfolder is a fully featured LIMS system that also comes with a clean and easy-to-use design. The platform facilitates collaborations within the group with an extensive rights management system.

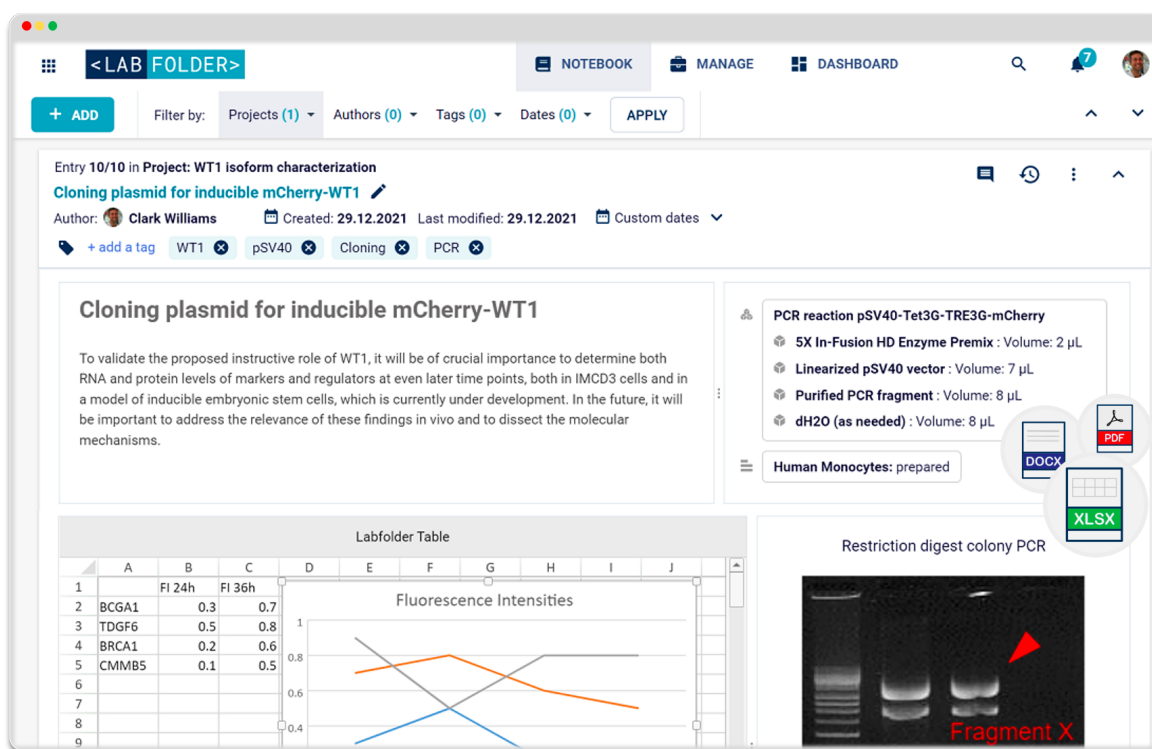


Figure 3. Front-end screenshot of Labfolder. This figure is taken from <https://www.labfolder.com/>.

eLabJournal

eLabJournal is a web software developed by eLabNext company, founded in 2010 in the Netherlands. The tool is closed source, with no free version available. By customer request eLabJournal can be installed on premises. It comes as multi-purpose, fully featured LIMS system. It has a clear and intuitive design. Inventory management is a strong feature of eLabJournal, as it allows tracking samples, managing protocols, centralize supplies ordering and book instrument access.

eLabJournal maintains a marketplace with a large variety of add-ons to extend the basic functionality of the platform. It also provides a fully-featured RESTful API allowing integration with the other tools.

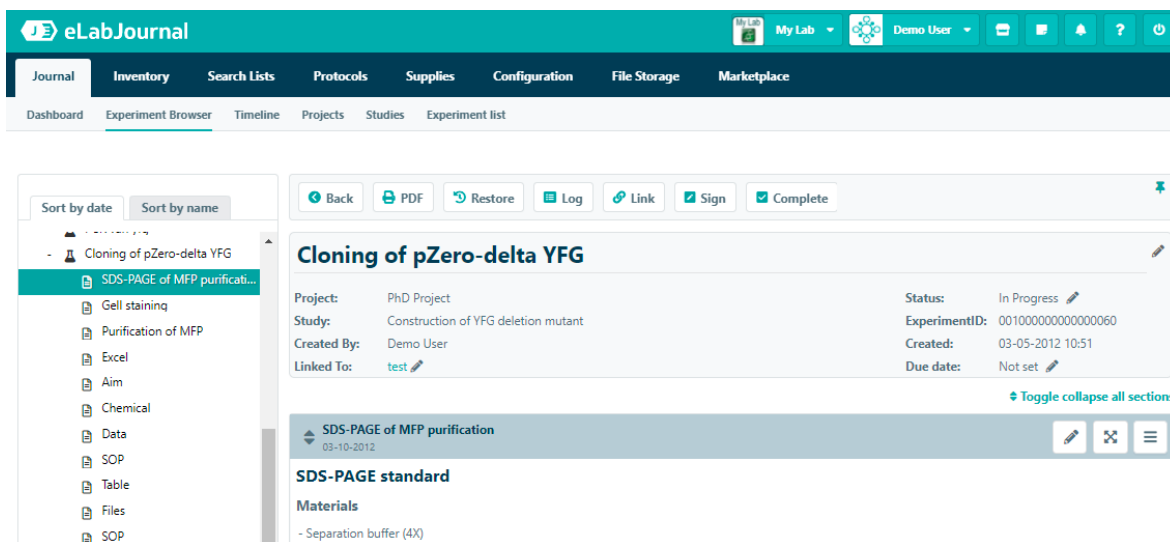


Figure 4. Front-end screenshot of eLabJournal. The figure is taken from <https://www.elabnext.com/>.

RSpace

RSpace was created in 2012 as a successor of eCAT. It is based in Edinburgh, Scotland. The original product was found in 2003 making which makes it one of the oldest ELNs on the market. It is closed source, but it offers a free "community edition" version. RSpace hosts data on cloud, but can be installed on premises if customers require so. RSpace offers out-of-the-box integrations with a lot of different tools including Jupyter, Slack, ChemDraw, Dropbox, and many more. RSpace has a clear and intuitive to use interface.

RSpace offers beta version inventory system for paid subscriptions, while the full production release is expected in December 2022. For easily integration with custom tools, it provides a RESTful API.



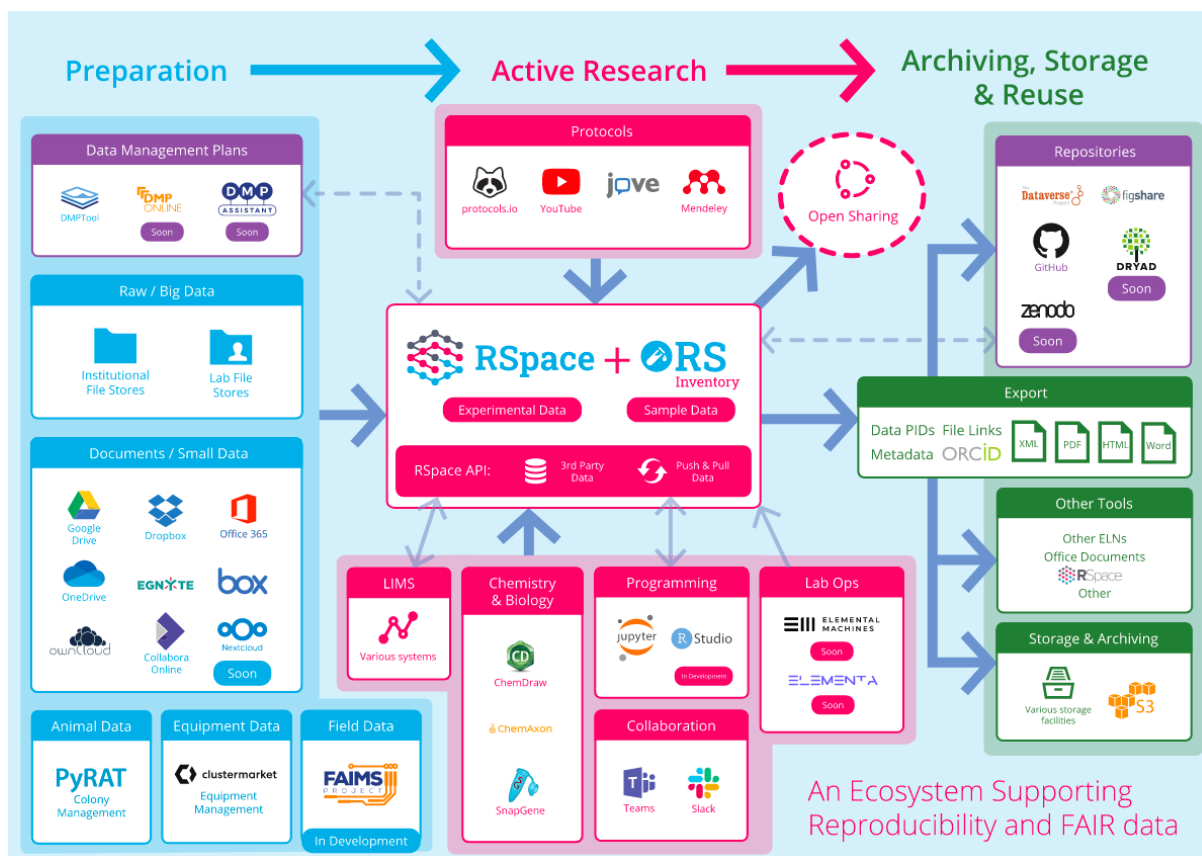


Figure 5. Front-end screenshot of RSpace. The figure is taken from <https://www.researchspace.com/>.

eLabFTW

The eLabFTW ELN is an open-source ELN that is in development since 2012. The main developer is Nicolas Carpi, the founder of Deltablot company that provides hosting and support options for users of eLabFTW. Moreover, eLabFTW is also a community-driven project with many contributions from external developers. The software is fully open source and is distributed under AGPL-3.0 license. eLabFTW is written in the PHP language. It comes with a support of the RESTful API to integrate with the other tools. In addition, a Python library named elabapy facilitates working with tools developed in Python language.

eLabFTW allows users to document experiments and link inventory needed for such an experiment. The experiment is described in a simple text field that can be further enhanced with pictures, drawings and molecular sketches. The inventory support of eLabFTW appears to be somewhat basic, as, for instance, it doesn't allow to follow the use of quantities of chemicals in different experiments.

While eLabFTW has been developed first with a target on biochemistry/medicine, its usage is actually very general and it can be applied to many domains. The inventory (called Database), while still basic, is versatile for different needs. Items can be differentiated into categories (e.g., in one case Instruments, Lab equipment, Substrates, Chemicals, but also as general as Publications of the group). All these items are linked between them and with experiments. The items in the Database support the addition of attached metadata to it (in JSON format), and eLabFTW itself is provided with APIs by which one can control several aspects of its various elements.



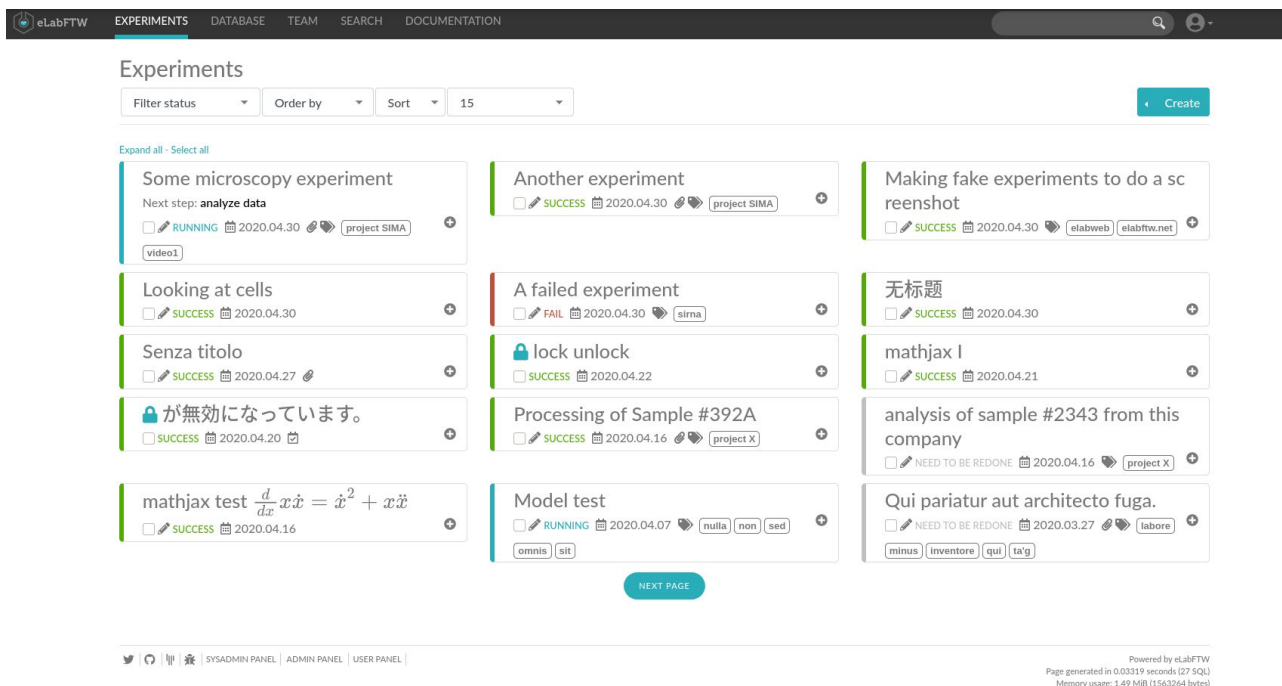


Figure 6. Front-end screenshot of eLabFTW. This figure is taken from <https://www.elabftw.net/>.

AiiDA

We emphasize that also for computational research and simulations data provenance is equally important. For simulations, data-provenance tracking is arguably easier than with experiments, since there are fewer external factors that can affect the results. Nevertheless, still many open challenges remain, and adoption of tools and practices to guarantee reproducibility and tracking of data provenance are not as widespread as one would like to. The problem has been though actively addressed and researched in the past few years, with solutions such as AiiDA [<https://www.aiida.net> and S.P.Huber et al., Sci. Data 7, 300 (2020), <https://www.nature.com/articles/s41597-020-00638-4>] providing automated tools to track data provenance, letting thus the researchers focus on the science rather than on the research data management (RDM). As a result, researchers using AiiDA will have not only the inputs and outputs of all simulations they ran automatically stored, but they will be organized in the form of a directed acyclic graph (DAG). In the DAG, inputs and outputs are linked to the respective simulations, but the DAG also natively highlights when outputs were used as inputs of subsequent calculations, thus providing an additional layer of metadata that can both convey information on the underlying workflows and how the final published result was obtained, and guarantee reproducibility not only of single calculations but of entire workflow sequences. As an example, to demonstrate the typical complexity of simulation research in computational materials science, automatically captured by tools such as AiiDA, we show in Figure 7 the provenance graph originating from a research project computing materials properties for about 500 materials. Even for such a relatively small-scale research project, the complexity of the resulting interconnections between calculations and data in the provenance DAG is clear, and it would be essentially impossible to regenerate manually even

by the scientist that performed the research. This highlights how the use of tools such as AiiDA is crucial to guarantee data accurate provenance tracking and, thus, reproducibility.

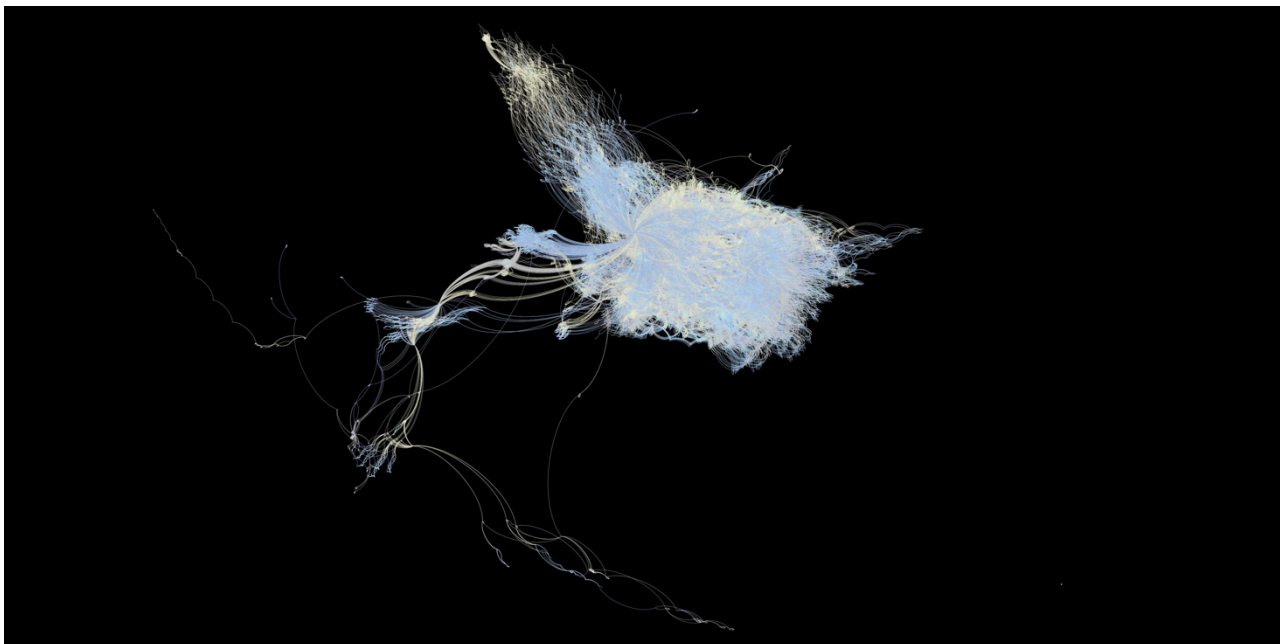


Figure 7. AiiDA DAG (direct acyclic graph) providing the entire provenance tree of complex computational workflows.

Integration of computational simulations

Nowadays, experimental science goes hand in hand with simulations. It is a good practice to support experimental findings with predictions based on computations. That is why, the obvious next step is integration of an ELN with a simulation platform, such as AiiDALab [<https://www.sciencedirect.com/science/article/pii/S092702562030656X>]. In the figure below we demonstrate a proof-of-concept for such an interaction developed in a laboratory from Empa [<https://www.empa.ch/>].

Typical tasks that experimental scientists have is to identify the adsorption configuration of a molecule on the surface. This is often done by comparison of the experimental STM images with the simulated ones. Since all the experiments are tracked withing the openBIS ELN, it makes sense to allow for an automated transfer of a molecule from ELN to AiiDALab simulation platform. Further, this molecule can be placed on the substrate and optimized. The optimised structure can be further used for simulating STM images, which can further be compared to the experimental ones (see Figure 8).

The integration is done employing the pybis tool [<https://pypi.org/project/PyBIS/>] maintained by the openBIS developers. It uses the openBIS API to enable data transfer between AiiDALab and openBIS instances. The data provenance is kept both in AiiDALab and openBIS, making it possible for the researcher to easily track what experiments and simulations were done to a specific material. In the figure below it is shown how user can find simulations done for molecule DBTP adsorbed on the Au[111] substrate. The search entry also shows an extra STM simulation that was done for DBTP.

The screenshot shows the AiiDALab interface. On the left is a navigation sidebar with categories like 'Lab Notebook', 'Inventory', 'Materials', 'Methods', 'Stock', 'Publications', 'User Profile', and 'Settings'. The main area is titled 'Collection: Molecules' and displays a table of molecules. The table has columns for Code, Preview, Name, SMILES, Identifier, ID, Type, Space, Parents, Children, Storage, Registrar, and Registration Date. Four molecules are listed: MCL14 (ID: 013b), MCL22 (ID: 223c), MCL18 (ID: 200a), and MCL19 (ID: 125f). Each entry includes a chemical structure and its SMILES string. To the right, a 'Submit geo opt for molecules, slabs and bulks' panel is open, showing a 3D model of a DBTP molecule (C18H12Br2) adsorbed on a gold slab (Au(111)). The panel includes options for 'Space' (Materials), 'Project' (Samples), and 'Object' (223c). It also features a 'Selection' tool, 'Edit Structure' options, and 'Process description' (223c on gold substrate). The 'Dispersion Correct' section is checked, and the 'Calculation Type' is set to 'Full DFT'.

Figure 8. Integration between openBIS and AiiDALab. A DBTP molecule (4,4''-dibromo-para-terphenyl) is imported into the AiiDALab application; once the structure is imported, the user can place it on a slab and further perform a variety of simulations that range from simple geometry optimization to scanning tunnelling microscopy.

Search AiiDA Database Slab optimizations

OLD workchains

System type:

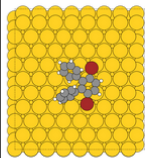
PKs:

Formulas:

Calculation Name:

Select the date range: From: To:

show comments

| PK | Creation Time | Formula | Calculation name | Energy(eV) | Structure | Extras |
|-----|------------------|-----------------------------------|------------------|--------------|---|--------|
| 301 | 2021-05-08 16:07 | C18H12Br2 at Au320 saturated: H80 | test open_bis | -294579.5129 |  | STM 1 |

Found 1 matching entries.

Figure 9. DBTP adsorbed on Au[111] can be found in the database of completed simulations. Please note that the STM image was simulated for the structure that was optimized in the previous AiiDALab workflow.

Once the STM link is opened a user is presented with an overview of the STM results. Those results can be further exported back to the ELN keeping the data provenance.

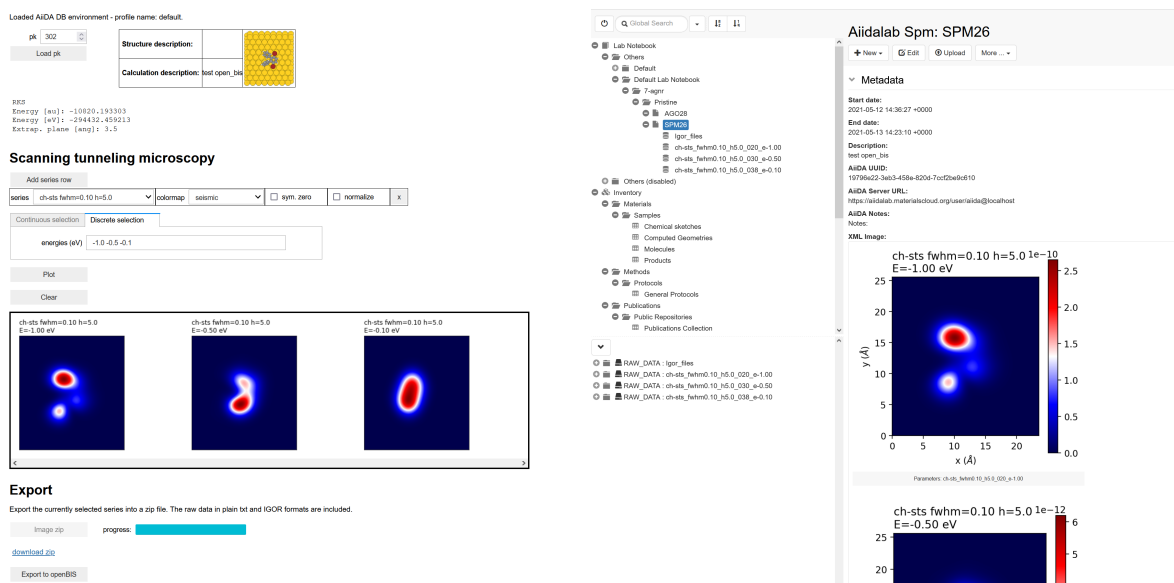


Figure 10. Simulated STM images for DBTP in Aiidalab (left) that got exported back to the openBIS (right).

Such openBIS-Aiidalab integration demonstrates how ELNs integrate with simulation environments offering a simple route for theory-experiment collaborations. Here we demonstrate one specific use case of STM image comparison, but the general idea remains the same and the approach can be transferred to other applications and scientific fields. With this approach scientists can run simulations in simple manner making them routine like any other types of measurements they do in the lab.

Conclusions

In the table below, we provide a summary of the collected information. Pricewise the tools range from free to up to 17 € per user per month. All the considered tools allow on-premises installation, which satisfies the requirement of keeping the data within the country. All ELNs except cheminfo are general and multipurpose, while cheminfo focuses on chemistry applications. All closed source ELNs offer REST API support, which is an advantage to openBIS and cheminfo.

What concerns the interface, all the paid ELNs are better at offering a more clean, intuitive and nice-looking interface. All the tools except cheminfo provide full support for inventory management.

Table 1. ELN solutions available on the market with respect to the criteria identified earlier

| Feature | openBIS | cheminfo | Labfolder | eLabJournal | RSpace | eLabFTW |
|---------|-------------------|-------------------|---|----------------------------|--|-------------------|
| Price | Free, open source | Free, open source | 17.0 € per user per month. Free for groups of | 12.95 € per user per month | 10 \$ per user per month. Free community | Free, open source |

| | | | | | | |
|-----------------------|---|---|-------------------------------|--------------------------------|--|---|
| | | | max 3 people. | | edition. | |
| Data security/Hosting | Manged by the IT specialists within institution | Manged by the IT specialists within institution | Allows on-premises deployment | Allows on-premises deployment | Allows on-premises deployment | Managed by the IT specialists within institution. Offers paid priority support and managed installation in a cloud. |
| Versatility | General | Chemistry | General | General | General | General |
| API | Java API, Javascript API, python API through pybis. | Python API through cheminfopy | RESTful API for data exchange | RESTful API support | RESTful API support | RESTful API support, elabapy tool for Python |
| User interface | Complex | Simple, sample-centered | Clean design, modern inteface | Clean design, modern interface | Clean design, modern interface | Clear and simple |
| Inventory management | Full support | Chemical inventory only | Full support | Full support | Full support available as Beta release | Basic inventory management |

After careful consideration of the ELNs, we concluded that openBIS and eLabFTW could be some the most appealing and forward-looking possibilities. For openBIS, while it is open source it comes with a satisfying level of support from the IT services of ETHZ. For eLabFTW, there is already significant uptake at the CNR Laboratories. Since in all cases it is required to have an instance of an ELN deployed per institute, it is unavoidable to allocate human resources to manage the installation - this makes free solutions more preferable, as they do not come with additional license costs. OpenBIS seems superior to cheminfo because of its multi-purposeness and better support of the inventory. At the same time, it would be recommended that developers invest in the development of RESTful APIs to simplify access from the other tools/hardware. Integration with simulation solutions such as AiiDALab are also appealing – and for what concerns simulations, tools such as AiiDA and the corresponding AiiDALab interface start to be adopted by the materials research community, and thus it becomes relevant (as discussed above) to also investigate how such tools can interact and couple with ELNs and LIMSEs. This is ever more relevance since data provenance is a complex production chain, and ELNs are just the beginning. A complete data provenance flow will be reported in D16.6 and will be discussed together with the rest of the JA6, in order to coordinate the strategy.

