Fifth FAIRmat Users Meeting

November 27, 2024

Center for the Science of Materials Berlin (CSMB)
Humboldt-Universität zu Berlin
Berlin, Germany

Overview Schedule

Wednesday, November 27
Introduction to FAIRmat *9:00-9:25; Room: 2.049*
User Talks *9:25-10:15; Room: 2.049*
Coffee Break *10:15-10:45; Room: Foyer*
FAIRmat Talk *10:45-11:10; Room: 2.049*
NFDI Consortium Talks *11:10-12:00; Room: 2.049*
Lunch *12:00-13:00; Foyer*
NOMAD I - On-boarding workshop *13:00-14:30; Room: 2.049*
Coffee Break *14:30-15:00; Room: Foyer*
NOMAD II - Tailoring RDM with NOMAD *15:00-16:30; Room: 2.049*
User support desks *16:30-18:00; Room: Foyer*

The detailed program can be found below and on the **webpage**.

FAIRmat Introduction Talk

FAIR Research Data Management with FAIRmat and NOMAD

Ahmed E. Mansour

In this talk, I will present the latest developments of FAIRmat and the NOMAD ecosystem, emphasizing the critical aspects of FAIR research data management (RDM) in condensed-matter physics. The NOMAD ecosystem, which includes platforms such as NOMAD, NOMAD Oasis, and NOMAD CAMELS, provides a complete solution for an RDM framework tailored to the needs of the condensed-matter physics and materials-science communities. I will demonstrate how these platforms facilitate the systematic acquisition and management of research data from diverse sources, including materials synthesis, advanced characterization techniques, and simulations. The presentation will also cover how NOMAD ensures that research data is not only securely stored, but also remains accessible and interoperable across different research domains. This in turn supports collaborative research efforts and accelerates innovation by ensuring that data is AI-ready.

Time: 9:00 – 9:25; Room: 2.049

User Talks

The FAIRification of PV research data

Eva Unger

The rapid expansion of research in many domains presents an increasing problem: results and information get distributed and hence fragmented across different scientific publications. This is in direct opposition to FAIR data principles and, as a consequence, research data is being underutilized despite the growing technical opportunities to acquire huge datasets of high-quality information. In addition, peer-reviewed publication often favors bias toward positive results, and a lot of solid research data is not being published in lieu of exciting scientific narratives that enable chasing high-impact publications. Out of desperation, we launched a collaborative initiative in 2019, collecting a unified database of perovskite solar cell data, now with information from over 46,000 individual solar cells, which is among the most expansive and comprehensive datasets in the field of PV. We are currently undertaking efforts to automatize data feed-in through LLM-based data mining (led by Kevin Jablonka). This initial effort taught us many important lessons on the challenge and importance of collecting cohesive datasets based on a unified data model. To enhance accessibility and adherence to FAIR principles, we migrated this dataset to the NOMAD data infrastructure and are now building a research data management platform for PV metadata as well as measurement data in collaboration with Helmholtz PV scientists. The goal is to generate a basis to enable the utilization of advanced machine learning methods in making better use of the collective research outcomes generated. This talk will outline our progress in data standardization, technical implementation, and the tailored adaptations developed with Helmholtz-Zentrum Berlin to meet the specific needs of photovoltaic research. We invite collaboration with AI and data management experts to optimize this resource for the research community, promoting a sustainable and globally interconnected data ecosystem.

Time: 09:25 – 09:50; Room: 2.049

Enabling high-throughput materials discovery of phosphosulfides by developments in FAIR data management, visualization and analysis in NOMAD

Lena Mittmann

High-throughput methods are increasingly employed to accelerate the discovery and investigation of new materials. In our group we produce combinatorial thin-film libraries with varying compositions in both X and Y direction. This drastically reduces the time spend on synthesis when mapping a compositional space and at the same time increases the need for detailed characterization and data management. This talk will focus on how NOMAD is used in our research and helps us conduct our daily research in the lab. Firstly, we track our samples all the way from cutting the substrates to the right size to annealing after deposition. NOMAD is also used to keep track of the usage and state of the instrument and targets used during the processes to better understand long term changes and contaminations. Lastly, we use NOMAD to organize and analyze our characterization data and to combine results from different mapping measurements done on one combinatorial sample. The new characterization techniques introduced to NOMAD are covering methods to determine compositional (EXD, XPS), structural (XRD) and optical properties (PL, ellipsometry).

Time: 09:50 – 10:10; Room: 2.049

FAIRmat Talk

Martignac: Computational workflows for reproducible, traceable, and composable coarse-grained Martini simulations

Tristan Bereau

Despite their wide use and far-reaching implications, molecular dynamics (MD) simulations suffer from a lack of both traceability and reproducibility. We introduce Martignac: computational workflows for the coarse-grained (CG) Martini force field. Martignac describes Martini CG MD simulations as an acyclic directed graph, providing the entire history of a simulation—from system preparation to property calculations. Martignac connects to NOMAD, such that all simulation data generated are automatically normalized and stored according to the FAIR principles. We present several prototypical Martini workflows, including system generation of simple liquids and bilayers, as well as free-energy calculations for solute solvation in homogeneous liquids and drug permeation in lipid bilayers. By connecting to the NOMAD database to automatically pull existing simulations and push any new simulation generated, Martignac contributes to improving the sustainability and reproducibility of molecular simulations.

Time: 10:45 – 11:10; Room: 2.049

NFDI Consortium Talks

Achieving semantic interoperability in materials science data and simulation workflows

Abril Azocar Guzman

The multiscale and multidisciplinary nature of materials science leads to complex scientific workflows and highly dimensional data. A lack of structured (meta)data hinders researchers' ability to find, access, interoperate, and reuse data [1] — all critical limitations for data-driven approaches and enhancing research sustainability. To address the manifold challenges in digitalization efforts within the materials science community, we develop ontologies with the goal of achieving semantic interoperability in the context of NFDI-MatWerk across various applications and use cases. In the field of atomistic simulations specifically, several challenges impair data reusability: (1) To facilitate the understanding and reuse of atomic structure data, well-described and harmonized metadata is essential. However, most existing approaches focus solely on perfect crystal structures, often overlooking defects. (2) Calculations frequently involve a combination of different software tools and diverse file formats, resulting in heterogeneous metadata that lacks semantic interoperability. (3) Workflow provenance detailing the processes used to set up digital samples is often absent. To tackle these challenges and facilitate data reuse, we have developed the Computational Materials Sample Ontology, an application-level ontology initially focused on describing structures at the atomistic level [2]. Its use is complemented by the development of domain-level ontologies that describe crystallographic defects [3] and atomistic simulation concepts [4]. To assist domain scientists in implementing ontologies in their research, the software tool atomRDF [3] enables users to annotate their data with ontologies automatically, creating application-level knowledge graphs. This enhances the querying and findability of research data. The combination of controlled vocabularies and software tools for generating linked open data promotes interoperability across file formats and software, while also offering potential for knowledge engineering and AI-ready data, which accelerates materials discovery.

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al., Sci Data, 2016, 3, 160018.
[2] https://purls.helmholtz-metadaten.de/cmso/
[3] https://github.com/OCDO/
[4] https://purls.helmholtz-metadaten.de/asmo/
[5] https://github.com/pyscal/atomRDF

Time: 11:10 – 11:35; Room: 2.049

An overview of NFDI4Cat´4 services and tools with a special focus on Voc4Cat the shared vocabularies for catalysis and related disciplines

David Linke

The talk will cover recent progress on services and tools developed in NFDI4Cat that could be useful beyond Catalysis. An overview will be presented to foster discussions on deepening the collaboration between FAIRMat and NFDI4Cat. The community represented by NFDI4Cat is characterised by a very broad range of topics from material science over chemistry and biology to chemical engineering. This results in large variety of data and complex metadata, which make achieving interoperability particularly challenging. Therefore, NFDI4Cat has devoted significant efforts to standard-compliant fundamental building blocks of FAIR RDM, such as ontologies, vocabularies or persistent identifiers.

Shared, machine-readable vocabularies are of utmost importance to FAIRly annotate data for machine consumption and to facilitate data reuse. In the field of catalysis, no such vocabulary existed which motivated the creation of Voc4Cat. For each concept (or "term"), Voc4Cat provides a unique resolvable persistent identifier and a carefully written textual definition, which reflects the community's shared understanding of the concept's meaning. The vocabulary is expressed in the SKOS standard. By using the identifiers provided by Voc4Cat to annotate data, ambiguity as to what is meant is avoided. Producing Voc4Cat-annotated data contributes to realizing the vision of machine-actionability described as the ultimate goal of the FAIR principles. Examples of (potential) applications where Voc4Cat can be integrated include its use in local electronic-lab-notebook-like tools (e.g. LARASuite, ADACTA, CaRMeN, LablMotion), data repositories (e.g. NOMAD, Repo4Cat), in education (e.g., RDM4Lab), and for a keyword catalogue in scientific publishing (e.g., collaboration with ChemCatChem).

Time: 11:35 – 12:00; Room: 2.049

NOMAD I - On-boarding Workshop

Siamak Nakhaie and Sarthak Kapoor

In this workshop, we will provide insights into NOMAD's approach to structuring data and metadata and continue with hands-on experience creating and managing ELN (Electronic Lab Notebook) entries for samples, measurements, and experiments. We will also introduce you to the advanced filters embedded in NOMAD and create interactive dashboards and visualizations to explore the vast amount of published data within NOMAD. You will need a laptop to participate in the interactive part of the workshop.

Time: 13:00-14:30; Room: 2.049

NOMAD II - Tailored RDM with NOMAD

Lauri Himanen and Markus Scheidgen

This workshop will explore how the NOMAD platform can be tailored to specific RDM needs. The workshop is built around a Jupyter Notebook, exploring how schemas, parsers, and apps can be built in Python to customize a NOMAD installation. The notebook will be available online; you only need a laptop and basic Python knowledge.

Time: 15:00-16:30; Room: Room: 2.049

User Support Desk

Talk to our Experts

Our experts and developers from various areas will be available to support you. Whether you need help with a specific question or need a consultation, our team will be around to help.

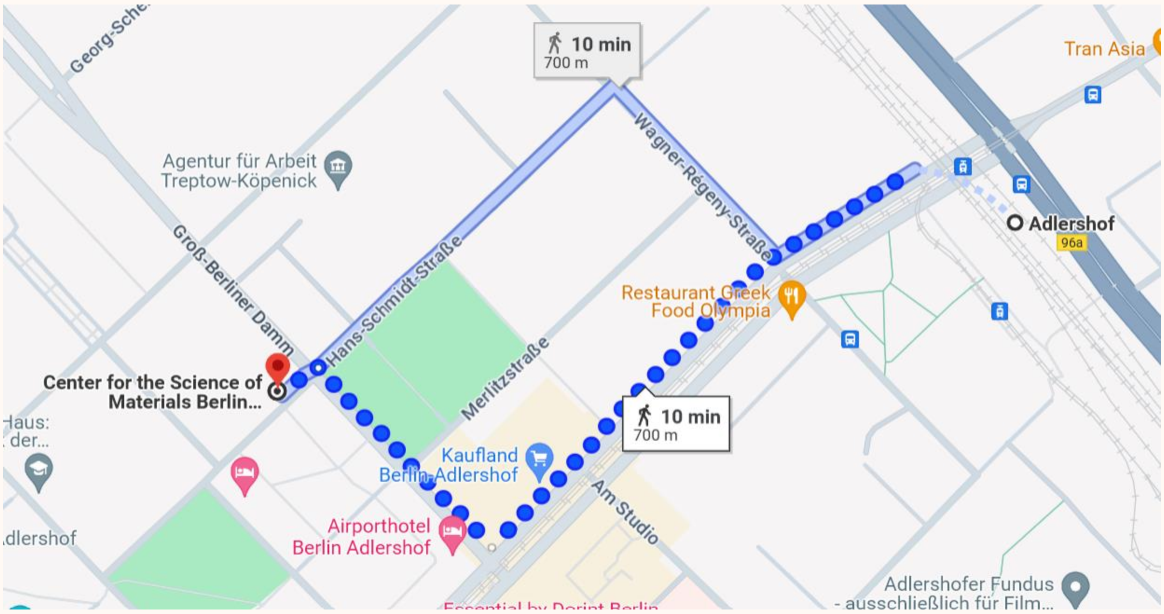
Time: 16:30-18:00; Room: Foyer

Venue

Center for the Science of Materials Berlin (CSMB)
Humboldt-Universität zu Berlin
Zum Großen Windkanal 2,
12489 Berlin, Germany

Public transport: S Adlershof (Lines: S8, S85, S9, S45, S46).

Walking from S Adlershof:



Contact

General organizational questions:

- Natalia Slepenko: natalia.slepenko@physik.hu-berlin.de
- Carolin Rehermann: carolin.rehermann@physik.hu-berlin.de