Symposium Pushing the Boundaries of Fair Data Practices for Condensed Matter Insights: From Workflows to Machine Learning (SYFD)

jointly organised by
the Surface Science Division (O),
the Chemical and Polymer Physics Division (CPP),
the Thin Films Division (DS), and
the Magnetism Division (MA)

Bridget Murphy
Christian-Albrechts-Universität zu Kiel
Leibnizstraße 19
24098 Kiel
murphy@physik.uni-kiel.de

Claudia Draxl
Humboldt-Universität zu Berlin
Zum Großen Windkanal 2
12489 Berlin
claudia.draxl@physik.hu-berlin.de

Frank Schreiber
Universität Tübingen
Auf der Morgenstelle 10
72076 Tübingen
frank.schreiber@uni-tuebingen.de

This symposium will highlight best practice in FAIR data and the development of streamlined workflows for open data and machine learning techniques. To be jointly organised by DAPHNE4NFDI and FAIRmat, topics will cover optimizing data collection methodologies and workflows, implementing electronic lab notebooks for efficient data recording, and integrating on-the-fly analysis techniques to enhance experimental outcomes. Moreover, the seminar will showcase cutting-edge advancements in data analysis methodologies, with a particular focus on on-the-fly analysis and machine learning techniques tailored to synchrotron and neutron data. Participants will discuss the integration of machine learning algorithms for data processing, analysis, and interpretation, thereby unlocking new avenues for scientific discovery and innovation. Furthermore, the DPG symposium will address the challenges associated with the storage and management of big data generated by modern data collection techniques and in particular, for condensed matter at synchrotron and neutron facilities. The symposium will provide the international state-of-the-art of the different disciplines and technology and give a platform to discuss future challenges and develop common solutions.

Overview of Invited Talks and Sessions

(Lecture hall H1)

Invited Talks

SYFD 1.1	Wed	9:30–10:00	H1	**Pushing the Boundaries of Fair Data Practices for Condensed Matter Insight** — •Astrid Schneidwind
SYFD 1.2	Wed	10:00–10:30	H1	**Establishing Workflows of Experimental Solar Cell Data into NOMAD** — Edgar Nandayapa, Paolo Graniero, Jose Marquez, Michael Götte, •Eva Unger
SYFD 1.3	Wed	10:30–11:00	H1	**Building up the EOSC Federation** — •Ute Gunsenheimer
SYFD 1.4	Wed	11:15–11:45	H1	**Data-Driven Materials Science for Energy-Sustainable Applications** — •Jacqueline Cole
SYFD 1.5	Wed	11:45–12:15	H1	**Machine Learning and FAIR Data in X-ray Surface Science** — •Stefan Kowarik

Sessions

SYFD 1.1–1.5	Wed	9:30–12:15	H1	**Pushing the Boundaries of Fair Data Practices for Condensed Matter Insights**

SYFD 1: Pushing the Boundaries of Fair Data Practices for Condensed Matter Insights

Time: Wednesday 9:30–12:15 Location: H1

Invited Talk SYFD 1.1 Wed 9:30 H1

Pushing the Boundaries of Fair Data Practices for Condensed Matter Insight — •Astrid Schneidwind — JCNS at MLZ Garching, FZ Jülich, Germany

The scientific impact of experiments using neutron, synchrotron and free-electron X-ray sources is drastically increasing, not least related to recent experimental and technical developments, which increase the demand for experiments, too. In parallel, new opportunities attract new researchers with less experience. The request for advanced computing opportunities - from data collection to robotics and AI-assisted experiments to the re-use of data and reproducible data analysis - is increasing accordingly. Adopting FAIR practices [1] opens further space for efficient, extended usage of the highly valuable data. Within the DAHNE4NFDI [2] initiative such workflows along the whole data pipeline are exemplarily developed, provided within use cases and systematically connected to prior and subsequent laboratory work. Standards are agreed on European level, as well as data repositories and reference data bases. ML-compatible data formats link the data to theoretical and computational approaches - closing the loop for increasing the efficiency of the experiments and exciting new outcomes.

[1] Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).

[2] https://doi.org/10.5281/zenodo.8040606.

Invited Talk SYFD 1.2 Wed 10:00 H1

Establishing Workflows of Experimental Solar Cell Data into NOMAD — Edgar Nandayapa[1], Paolo Graniero[1], Jose Marquez[2], Michael Götte[1], and •Eva Unger[1,3] — [1]Helmholtz-Zentrum Berlin für Materialen und Energie GmbH, HySPRINT Innovation Lab, Kekuléstraße 5, 12489 Berlin, Germany — [2]Humboldt University Berlin, FAIRmat Project, Zum Großen Windkanal 2, 12489 Berlin, Germany — [3]Humboldt University Berlin, Department of Chemistry and CSMB, Zum Großen Windkanal 2, 12489 Berlin, Germany

Materials for solar energy conversion are key enablers for the green energy transition. Perovskite Solar Cells (PSCs) are an excellent example of an emerging technology, where an intense and world-wide R&D activities has enabled a very fast improvement in the reported power conversion efficiencies. The monthly output of new reports in the peer-reviewed literature is in the hundreds to thousands and it is at this point neither effective nor possible to still make effcient use of the research results reported. Considering data reported in the peer-reviewed literature, this just represent the "tip of the iceberg" of research data that is actually being measured in labs around the world.

Out of desparation, our team started a fairly manual "data digging" initiative in 2019 to compile data from the then published data in the peer-reviewed literature into a single database based on a very rudimentary and single-metric representation of the actual research data resulting in the Perovskite Database (www.perovskitedatabase.com). In close collaboration with the FAIRmat project, we are now taking steps towards transferring the literature dataset into NOMAD and are creating NOMAD platforms to capture, store, analyse and share the actual experimental research data within and beyond our research community. The goal is to, both, initiate community driven data sharing platforms that can be used to directly share and disseminate experimental datasets to adhere to FAIR data principles, and make photovoltaic research data AI-ready to enable the utilization of modern ML-tools to facilitate a further acceleration of the technological exploitation of new materials.

Invited Talk SYFD 1.3 Wed 10:30 H1

Building up the EOSC Federation — •Ute Gunsenheimer — EOSC Association

The European Open Science Cloud (EOSC) envisions a unified system in Europe to enable researchers to store, share, process, and reuse FAIR data and services across disciplines and borders. Central to this vision is the EOSC Federation, a network of interconnected nodes, that will provide seamless access to scientific data and resources. With the public launch of the EOSC EU Node in October 2024, the development of the EOSC Federation has gained significant momentum. To realise its full potential, the EOSC Federation requires the enrolment of additional EOSC Nodes. The nodes will act as entry points to the Federation.

To enable the establishment of such a distributed system, many questions remain to be addressed, including defining minimum requirements for EOSC Nodes, establishing enrolment rules, ensuring effective governance, and developing financial mechanisms to support resource sharing. Answering these questions is critical to building a robust, distributed infrastructure that drives Open Science and innovation in Europe.

By the time of the conference, the first wave of EOSC Nodes will have been identified, representing a diverse range of thematic and national communities across Europe. At the same time, active work on the enrolment process will have been kicked-off.

This talk will explore the progress, challenges, and future steps in shaping the EOSC Federation.

15 min. break

Invited Talk SYFD 1.4 Wed 11:15 H1

Data-Driven Materials Science for Energy-Sustainable Applications — •Jacqueline Cole — Cavendish Laboratory, University of Cambridge, Cambridge, UK

Data-driven materials discovery is coming of age, given the rise of 'big data' and machine-learning (ML) methods. However, the most sophisticated ML methods need a lot of data to train them. Such data may be custom materials databases that comprise chemical names and their cognate properties for a given functional application; or data may comprise a large corpus of text to train a language model. This talk showcases our home-grown open-source software tools that have been developed to auto-generate custom materials databases for a given application. The presentation will also demonstrate how domain-specific language models can now be used as interactive engines for data-driven materials science. The talk illustrates the application of these data-science methods using case studies from the energy sector. The talk concludes with a forecast of how this 'paradigm shift' away from the use of static databases will likely evolve materials science.

Invited Talk SYFD 1.5 Wed 11:45 H1

Machine Learning and FAIR Data in X-ray Surface Science — •Stefan Kowarik — Phys. Chemistry, Univ. of Graz, Austria

Synchrotrons are among the world's largest producers of scientific data, yet many experiments fail to contribute adequately to databases. Publishing raw data without comprehensive metadata fails to align with the "Findable" and "Reusable" principles of FAIR data, which are essential to unlocking the full potential of these datasets.ML not only benefits from large FAIR datasets but also facilitates their creation. Our recent work highlights live ML-based analysis of X-ray reflectometry (XRR) for thin-film characterization, enabling adaptive experimentation with a fourfold increase in speed. Additionally, we demonstrate automated crystal structure solutions from grazing-incidence X-ray diffraction (GIXD) of thin films. These advancements lay the foundation for self-driving laboratories, where integrated ML algorithms can control thin-film deposition processes, enhancing precision and throughput.Importantly, live ML analysis generates metadata, such as unit cell parameters in textured thin films, improving data findability and reusability. While XRR requires standardized structural model formats*efforts championed by groups like ORSO*GIXD leverages established crystallographic formats for database integration.*In the future, these advancements could culminate in expansive, standardized databases for surface science, encompassing thin-film crystal structures, surface reconstructions, and thin film material properties, analogous to established bulk crystallographic databases.