TT 68: Focus Session: Making Experimental Data F.A.I.R. – New Concepts for Research Data Management II (joint session O/TT)

Time: Friday 9:30–12:45 Location: WIL A317

Topical Talk TT 68.1 Fri 9:30 WIL A317
FAIRifying ARPES: a Route to Open Data & Data Analytics — •Ralph Ernstorfer[1,2], Tommaso Pincelli[1,2], Patrick R. Xian[2], Abeer Arora[2], Florian Dobener[3], Sandor Brockhauser[3], and Laurenz Rettig[2] — [1]TU Berlin, Germany — [2]Fritz-Haber-Institut Berlin, Germany — [3]HU Berlin, Germany

While angle-resolved photoemission spectroscopy (ARPES) is the most direct probe of crystals' electronic structure, the globally collected ARPES data have not been merged into an open experimental electronic structure database in equivalence to well-established atomic structure databases. We discuss a data format based on NeXus [1] as a concept for unifying the data structure for all types of photoemission experiments including time-, spin-, and time-resolved ARPES [2]. The aim is to immediately enable preprocessed data and metadata shareability according to FAIR data principles, employing existing public storage and archiving research data infrastructures such as Zenodo, OpenAIRE, and Nomad/FAIRmat. Ultimately, the multidimensional photoemission spectroscopy (MPES) format is designed to allow high-performance automated access, providing experimental databases for high-throughput material search [3]. References: [1] https://www.nexusformat.org/ [2] https://mpes.science/; https://fairmat-experimental.github.io/nexus-fairmat-proposal/ [3] R. P. Xian et al., Scientific Data 7, 442 (2020); R.P. Xian et al., Nat. Comp. Sci, in print, arXiv:2005.10210

TT 68.2 Fri 10:00 WIL A317
A FAIR data infrastucture for photoemission experiments — •Marten Wiehn[1], Tobias Eul[1], Benjamin Stadtmüller[1,2], and Martin Aeschlimann[1] — [1]Department of Physics and Research Center OPTIMAS, TU Kaiserslautern, Germany — [2]Institute of Physics, Johannes Gutenberg University Mainz, 55128 Mainz, Germany

Recent trends toward data-driven, high-tech experimental research and the growing volumes of data associated with it show the increasing importance of comprehensive data acquisition and management. We present an automated workflow for well-described photoemission data from experiment to archive and publication. Utilizing a powerful experiment control software to capture essential metadata for each measurement enables the collection of FAIR-ready data (Findable, Accessible, Interoperable, Reusable). In addition, using an electronic lab notebook pushes our lab further toward a FAIR data infrastructure that supports researchers in their daily work.

TT 68.3 Fri 10:15 WIL A317
Multi-Dimensional Photoemission Spectroscopy: a concept for FAIR photoemission data — •Florian Dobener[1], Tommaso Pincelli[2,3], Abeer Arora[2,4], Steinn Ymir Augustsson[5], Dmytro Kutnyakhov[6], Michael Hartelt[7], Laurenz Rettig[1], Martin Aeschlimann[2], Ralph Ernstorfer[7], and Sandor Brockhauser[2,3] — [1]Department of Physics, HU Berlin, Germany — [2]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany — [3]Institut für Optik und Atomare Physik, TU Berlin, Germany — [4]FU Berlin, Fachbereich Physik, Berlin, Germany — [5]Institut für Physik, Johannes Gutenberg-Universität Mainz, Germany — [6]DESY, Hamburg, Germany — [7]Department of Physics and OPTIMAS, University of Kaiserslautern, Germany

The complexity and size of photoemission data is rapidly increasing as new technological breakthroughs have enabled multidimensional parallel acquisition. However, most of the community is currently using heterogeneous data formats and workflows. We propose a new data format based on NeXus, a hierarchically organized hdf5 structure. This multidimensional photoemission spectroscopy format is designed to allow high-performance automated access, enabling experimental databases for high-throughput material search. Our approach involves reaching out to the community using a website with extensive documentation of our proposed standard. As a demonstrator of the potential of our approach we present a workflow and data pipeline integrated into the NOMAD research data management solution, which provides powerful analysis and search functionalities.

TT 68.4 Fri 10:30 WIL A317
Towards an Infrastructure for FAIR Synthesis Data — •Sebastian Brückner[1,2], Andrea Albino[1], Jose Marquez[1], Florian Dobener[1], Hampus Näsström[1], Markus Scheidgen[1], Claudia Draxl[1], and Martin Albrecht[2] — [1]HU Berlin, Zum Großen Windkanal 2, 12489 Berlin — [2]IKZ Berlin, Max-Born-Straße 2, 12489 Berlin

A data infrastructure based on the FAIR (findable, accessible, interoperable and reusable) principles promises a new way of sharing and exploring data by using highly efficient data analysis and artificial intelligence tools. This also applies to data related to sample synthesis. At present, most synthesis data are not structured comprehensively or not even stored digitally but in handwritten lab books. There hardly exists any data standards in synthesis, which is in contrast to data from characterization techniques. The FAIRmat project (https://FAIRmat-NFDI.eu) is building a FAIR data infrastructure for condensed-matter physics and the chemical physics of solids. In FAIRmat's Area A, we focus on synthesis data to make sample synthesis reproducible, accelerate the development of novel materials, and make characterization data of synthesized materials assessable. Here we summarize our ongoing work and progress including: providing a general data model for synthesis which is harmonized with data from measurements and theory (ontologies); implementation of our data model in use cases and electronic laboratory notebooks; developing tools for data acquisition and analysis; data governance guidelines to enable a sustainable change of research data management at the institute/university level.

TT 68.5 Fri 10:45 WIL A317
FAIRifying Material Synthesis with the NOMAD Electronic Laboratory Notebook (ELN) — •Andrea Albino[1], Hampus Näsström[1], Florian Dobener[1], Jose Marquez Prieto[1], Lauri Himanen[1], David Sikter[1], Mohammad Nakhaee[1], Amir Golparvar[1], Sebastian Brückner[1], Martin Albrecht[2], Markus Scheidgen[1], and Claudia Draxl[1,3] — [1]Physics Department and IRIS Adlershof, Humboldt-Universität zu Berlin, Berlin, Germany. — [2]Leibniz-Institut für Kristallzüchtung, Berlin, Germany. — [3]The NOMAD Laboratory at the Fritz Haber Institute of the Max Planck Society, Berlin, Germany.

Approaching the era of big data-driven materials science, one crucial step to collecting, describing, and sharing experimental data is the adoption of ELNs. The project FAIRmat (fairmat-nfdi.eu) is offering such tools by developing and operating the open-source software NOMAD. The NOMAD ELN aims at offering a secure environment to protect the integrity of both data and metadata, whilst also affording the flexibility to adopt new synthetic processes or changes to existing ones without recourse to further software development.

We find that to promote an early adoption, it is important to adapt to a single user's needs and workflows. An inductive approach, going from a particular set of experiments to a general description of the similarities recurring in each of them, led us to adopt a common data structure as a standard. The state-of-the-art ELN features for a synthetic process will be shown in the talk, highlighting the development of both data modeling and specific implementation solutions.

TT 68.6 Fri 11:00 WIL A317
FAIR Data Infrastructure for Computation: Classical Simulations and Multiscale Modeling — •Joseph F. Rudzinski[1,2], José M. Pizarro[1], Nathan Daelman[1], Luca M. Ghiringhelli[1], Karsten Reuter[3], Kurt Kremer[2], Silvana Botti[4], and Claudia Draxl[1] — [1]Institut für Physik, Humboldt-Universität zu Berlin — [2]Max-Planck-Institut für Polymer Forschung, Mainz — [3]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin — [4]Institut für Festkörpertheorie und Optik, Friedrich-Schiller-Universität Jena

The emergence of the (big-)data-centric techniques as a fundamental paradigm of science calls for the development of infrastructure for ensuring FAIR—findable, accessible, interoperable, reusable—data management. The FAIRmat consortium aims to build extensive infrastructure for a wide variety of materials-science data, including soft matter simulations [1], by expanding upon the success of the NOMAD Laboratory—a repository for atomistic calculations in materials science [2]. Both the large volume and heterogeneous nature of classical molecular-dynamics simulation data presents a number of distinct challenges. In this talk, we present FAIRmat's progress in developing infrastructure for molecular-dynamics simulations, including metadata for molecular topologies and tools for workflow management. We will also discuss the need for standardization of metadata schemas and ontologies within the community, and planned collaborations with other open science initiatives and software developers.

[1] Scheffler, M. et al. Nature 2022, 604, 635-642.

[2] Draxl, C.; Scheffler, M. JPhys Materials 2019, 2, 036001.

Topical Talk TT 68.7 Fri 11:15 WIL A317

Electronic Lab Notebooks in Teaching and Implications on Science — •Michael Krieger — Lehrstuhl für Angewandte Physik, Department Physik, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

In our department, we have recently introduced Electronic Lab Notebooks (ELN) in the obligatory electronic lab course in the 4th semester of the physics curriculum. Immediate advantages for the students are obvious: all data, raw data and metadata including experiment description and experimental observations, are digitally stored at the same place. Moreover, student teams share and actively work in their group ELN with access at the university as well as at home, and script-based evaluations can be performed directly in the ELN.

The introduction of ELNs in teaching has also implications on science. Students carry their ELN experience and data competences into all research groups. There, however, modern research data management is much more complex. According to the FAIR principles, it requires structured, machine-readable data using open formats and vocabularies that meet community standards. The development of such standards is in many cases still to be done and is the core of the Nationale Forschungsdateninfrastruktur (NFDI). Here, ELNs in teaching provide a sandbox with short learning and innovation cycles for testing structured schemas. The experience helps to develop and establish sustainable and FAIR documentation of research workflows in science.

TT 68.8 Fri 11:45 WIL A317

A step towards predicting synthesis conditions of metal-organic frameworks — •Dinga Wonanke[1], Thomas Heine[2], and Christof Wöll[1] — [1]Institut für Funktionelle Grenzflächen (IFG), Karlsruher , Germany — [2]Faculty of Chemistry and Food Chemistry,Dresden, Germany

The process of synthesising metal-organic frameworks (MOFs) falls under a branch of chemistry known as reticular chemistry. Here, well defined crystalline compounds are synthesised from a well thought out design principle by linking predefined building blocks under specific conditions. Although, this approach appears to be intuitive, the synthesis of any novel MOF still follows the conventional approach that begins with a thorough literature survey to explore reagents, calculations of aliquots and finally a series of time consuming and stressful trial-and-error syntheses. Consequently, although millions of stable hypothetical MOFs with interesting properties have been predicted, only approximately 100 thousand crystal structures of MOFs currently exist in the Cambridge Structural Database (CSD). Indicating a significant bottleneck in the intelligent design of novel stable MOFs with targeted properties.

In this talk, we will present an overview of our journey to design a new machine learning algorithm for predicting the synthesis condition of existing and hypothetical MOFs. We will discuss our experiences and challenges in mining and curating the MOF subset in the CSD. Finally, we will present our new MOF database that maps every MOF to its experimental synthetic conditions.

TT 68.9 Fri 12:00 WIL A317

Deep learning surface scattering data analysis for processing large synchrotron datasets — •Vladimir Starostin, Valentin Munteanu, Linus Pithan, Alexander Gerlach, Alexander Hinderhofer, and Frank Schreiber — Institute of Applied Physics, University of Tübingen, Germany

In situ real-time surface scattering experiments such as grazing-incidence wide-angle X-ray scattering (GIWAXS) produce large amounts of data, frequently exceeding the capabilities of traditional data processing methods. Here we demonstrate an automated pipeline for the analysis of GIWAXS images, based on a machine learning architecture for object detection, designed to conform to the specifics of the scattering data [1]. Our pipeline enables real-time GIWAXS analysis and is designed to be employed at synchrotron facilities. We also present FAIR data strategies and traceable data resources from the raw data to the corresponding scientific publication and vice versa [2] including intermediate processing steps.

We demonstrate our method on real-time tracking of lead halide perovskite structure crystallization processes, which are relevant for hybrid solar cell applications. However, our approach is equally suitable for other crystalline thin-film materials by design. In general, the solution substantially accelerates the analysis process of GIWAXS images, potentially boosting the speed of scientific discoveries in material science.

[1] V. Starostin et al. *npj Comput Mater* **8**, 101 (2022)

[2] V. Starostin et al. *Synch Rad News* **13**, 31–37 (2022)

TT 68.10 Fri 12:15 WIL A317

FAIR Data Infrastructure for Computation: Mapping out the Space of Density Functionals — •Nathan Daelman[1], Joseph F. Rudzinski[1,2], José M. Pizarro[1], Luca M. Ghiringhelli[1], Miguel A. L. Marques[3], Silvana Botti[4], and Claudia Draxl[1] — [1]Institut für Physik und IRIS-Adlershof, Humboldt-Universität zu Berlin, Berlin — [2]Max-Planck-Institut für Polymer Forschung, Mainz — [3]Institut für Physik, Martin-Luther-University Halle-Wittenberg, Halle — [4]Friedrich Schiller Universität Jena, Jena

The NOMAD Laboratory [1] holds over 135 million computational results, the vast majority of which stem from density-functional theory (DFT). The platform provides adequate querying and data analytics tools (e.g., machine-learning modelling) for processing such Big Data. However, the exchange-correlation (xc) functional with which the data was generated, limits the analysis scope of most thermodynamic and kinetic properties. Here, we present a strategy rooted in semantics for extending method interoperability. We will showcase our map of the entire xc-functional space that, in the context of the FAIRmat consortium [2], is built to be widely accessible and facilitate findability. Lastly, we will discuss the integration of this xc-functionals map into the NOMAD data platform, as well as its publication in ontology format as an effort towards a community-wide vocabulary standard.

[1] C. Draxl and M. Scheffler, MRS Bulletin 43, 676-682 (2018).

[2] M. Scheffler, M. et al., Nature 604, 635-642 (2022).

TT 68.11 Fri 12:30 WIL A317

OpenSemanticLab: Towards Open Semantic Research — •Simon Stier and Matthias A. Popp — Fraunhofer ISC, Neunerplatz 2, 97082 Würzburg, Germany

In materials science, complex relationships exist between the properties of materials and their composition and processing. Therefore, digital transformation and acceleration in this domain represents a particularly important challenge. Although it is generally agreed that data must be linked by means of semantics and ontologies to form holistic data spaces, there is still a lack of suitable tools for integrating the necessary structures into the everyday work of scientists.

Fraunhofer ISC addresses this challenge with a broad-based strategy that closely links activities at all relevant levels. The goal hereby is the development towards Lab 4.0, the machine-readable documentation of scientific processes and the harmonization of data structures in accordance with international standards.

Core of the resulting OpenSource solution architecture is the central web data platform OpenSemanticLab [1] that links people (knowledge), machines (data) and algorithms (AI) equally. As an open system, this platform is easily adaptable even without programming knowledge and without losing the uniform structure. In this way, OpenSemanticLab enables us as scientists to contribute individually and yet in a standardized fashion to future digital materials research.

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