TT 56: Focus Session: Making Experimental Data F.A.I.R. – New Concepts for Research Data Management I (joint session O/TT)

Data have been identified as major resource of the 21st century, unlocking great potential if refined and processed in the right way. In scientific research, particularly modern data science concepts like machine learning or neural networks enable novel types of data analysis with often strong predictive power. This Focus Session aims at providing a framwork for presenting and discussing novel concepts, tools and platforms for managing experimental research data related to surface science and solid-state physics. In particular, in light of the German NFDI initiative, where several consortia are actively working on tackling the imminent challenges of research data management in experimental solid-state physics, this Focus Session will offer an ideal environment for exchange among researchers, and bringing these novel developments into the labs. The intended topics include the description of experimental data and meta data generation workflows, meta data schemas and file formats, electronic lab notebooks, novel tools for handling and analyzing scientific research data, as well as sharing and searching platforms according to F.A.I.R. principles.

Organizers: Martin Aeschlimann (TU Kaiserslautern), Laurenz Rettig (FHI Berlin) and Heiko Weber (U Erlangen)

Time: Thursday 15:00–18:30 Location: WIL A317

Topical Talk TT 56.1 Thu 15:00 WIL A317

Introducing a FAIR research data management infrastructure for experimental condensed matter physics data — •Christoph Koch — Humboldt-Universität zu Berlin, Department of Physics & IRIS Adlershof, Berlin, Germany

Digitization and an increase in complexity and price of experimental materials characterization techniques, an increase in accuracy and system size of computational solid state physics (or computational materials science), and the maturation of machine learning tools to extract patterns from large amounts of very diverse (annotated) data promise an acceleration of materials development by synergistically combining research data from many sources. While some labs start to upload their (raw) research data to data repositories, this is only a first but not sufficient step in leveraging the above-mentioned potential, since such repositories are typically either specific to a very particular technique or agnostic to the content of the data being uploaded. In both cases the research data cannot easily, and definitely not without significant human effort, be compared to and integrated with experimental data from other sources, or numerical predictions. In this talk I will report on recent progress of the FAIRmat NFDI consortium in extending the novel materials discovery laboratory (NOMAD), the world's largest data base for ab-initio computational materials data, to ingest experimental research data on the synthesis and characterization of materials in a machine-accessible manner, i.e. annotated with well-defined and interoperable metadata, achieved by establishing links between related (experimental and computational) quantities.

TT 56.2 Thu 15:30 WIL A317

Introducing an electronic laboratory notebook in a collaborative research center — •Sebastian T. Weber[1], Aneta Daxinger[1], Philipp Pirro[1], Marek Smaga[1], Christiane Ziegler[1], Mathias Kläui[2], Georg von Freymann[1], Baerbel Rethfeld[1], and Martin Aeschlimann[1] — [1]Department of Physics and Research Center OPTIMAS, RPTU Kaiserslautern-Landau — [2]Institute of Physics, Johannes Gutenberg University Mainz

The basis of a FAIR data management is a well-described and detailed documentation of every single step of the experiment and data analysis. In recent decades, however, the focus has shifted from analog measuring instruments and analytical calculations to computer-based experiments and simulations. This has lead to a large increase in the numbers of measurements and observed quantities and therefore in the amount of data generated. Consequently, traditional paper lab notebooks have reached their limits. Electronic lab notebooks (ELNs) are better suited for storing, indexing, searching and retrieving a large amount of entries. In particular, the automated filling-in of meta data can lead to a reduction in the workload of the scientists in the long term.

We present the lessons learned on challenges and advantages with the introduction of a joint electronic lab notebook within our collaborative research center CRC/TRR173 *Spin+X*. We report on our experiences in the daily work of the scientists and in education in student labs.

TT 56.3 Thu 15:45 WIL A317

An efficient workflow for processing single event dataframes. — •Steinn Ýmir Ágústsson[1], M. Zain Sohail[2,3], David Doblas Jiménez[4], Dmytro Kutnyakhov[3], and Laurenz Rettig[5] — [1]Aarhus University, DK — [2]RWTH, Aachen — [3]DESY, Hamburg — [4]Eu-XFEL, Schenefeld — [5]FHI, Berlin

Single event resolved data streams measured by delay-line-detectors allow to correlate each measured photoelectron with the state of the experimental apparatus. This allows corrections and calibrations to be applied on a shot-to-shot basis and a flexible investigation of correlations between various measurement parameters.

We are developing an open-source python package[1], where highly optimized dataframe management and binning methods enable leveraging the full potential of event-resolved data structures. The flexible design of the pipeline allows processing any event-resolved data stream.

With momentum microscopy as the primary target application, we developed axis calibration and artifact correction methods designed to be agnostic to the experimental apparatus. These methods are tested on data generated by microscopes at FELs (HEXTOF@FLASH) as well as at HHG sources (FHI), but are easily exended to other endstations using similar detection techniques.

Our aim is to provide tools for the community which will reduce the development time for each end station, as well as an open and accessible data processing pipeline, built around the FAIR data principles.

[1] github.com/openCOMPES/sed

TT 56.4 Thu 16:00 WIL A317

FAIR Data Infrastructure for Computation: Advanced many-body methods. — •José M. Pizarro[1], Nathan Daelman[1], Joseph F. Rudzinski[1,2], Luca M. Ghiringhelli[1], Roser Valentí[3], Silvana Botti[4], and Claudia Draxl[1] — [1]Institut für Physik und IRIS-Adlershof, Humboldt-Universität zu Berlin — [2]Max-Planck-Institut für Polymer Forschung, Mainz — [3]Institut für Theoretische Physik, Goethe University Frankfurt am Main — [4]Institut für Festkörpertheorie und Optik, Friedrich-Schiller-Universität Jena

Big-data analyses and machine-learning approaches have recently emerged as a new paradigm to study and predict properties of materials. In order to perform these analyses, materials data have to be structured in a FAIR (findable, accessible, interoperable, and reusable) format [1]. While most of the current databases deal with density-functional-theory (DFT) calculations, there is a clear need for developing FAIR-data schema for methodologies going beyond DFT. Methods such as the *GW* approximation, dynamical mean-field theory, and time-dependent DFT allow to calculate excited- and many-body-states properties beyond DFT, thus having a direct quantitative comparison with experiments. In this talk, we will introduce the achievements and challenges undertaken within the FAIRmat consortium towards fully structuring the (meta)data of all these techniques. We demonstrate how users can analyze the data and compare with angle-resolved photoemission spectroscopy.

[1] M. Scheffler et al., Nature **604**, 635 (2022).

TT 56.5 Thu 16:15 WIL A317

Electronic Laboratory Notebooks for FAIR Data Management; Evaluation and Recommendations for Solutions at Research Infrastructures — •Philipp Jordt[1], Wiebke Lohstroh[2], and Bridget Murphy[1] — [1]IEAP, Kiel University, Germany — [2]MLZ, Technische Universität München, Germany

Electronic Laboratory Notebooks (ELN) are the digital counterpart to the classical handwritten paper notebook and play a vital role in the implementation of FAIR data standards. Modern ELN solutions range from simple note taking applications to integrated tools, combining documentation, inventory management, progress tracking and more. Nowadays, ELNs are becoming more prominent in research laboratories around the world, replacing paper notebooks. This evaluation of basic needs was carried out in the context of the DAPHNE4NFDI consortia. Of special interest is the view on ELNs for combined use at large scale facilities and in the home laboratory. Thus, the requirements regarding implementation, deployment, authentication, etc., may differ from those for single or laboratory use at universities. An overview of different concepts and existing solutions is given. Multiple ELNs have been evaluated during test runs at large scale facilities and a survey on existing solutions was held. From these results, a list of ELN specifications is presented, ranging from useful to necessary. These insights may serve as a guideline for evaluating or implementing ELNs in the future.

TT 56.6 Thu 16:30 WIL A317

Ontology for Experimental Data — •Sandor Brockhauser — Center for Materials Science Data, Humboldt-Universität zu Berlin, Germany

Ontology is the scientific field of formal knowledge representation. This field contributes to Data Science and helps Experimentalist to properly annotate their data and metadata on a FAIR way. During the last decades several different ways have been developed in the field of Ontology for describing knowledge as a set of information and their relationships. These include Information mapping, Concept maps, Topic maps, Mind maps, Knowledge graphs, BORO (Business Objects Reference Ontology), RDF (Resource Description Framework), OWL (Web Ontology Language), ORO (Object-Role Modeling), UML (Unified Modeling Language), ISO 15926 (standard for data sharing), OLOG (mathematical framework for knowledge representation), GELLISH (ontology language for data storage and communication), etc. For describing experimental facts, we suggest using an ontology in OWL which is derived from the NeXus community standard. It represents all the concepts developed for explaining experiments and experimental data, just like the relationships between them. Such representation allows connecting the concepts defined in NeXus also to other ontologies. Additionally, any data management systems, like NOMAD which accepts experiment data provided in the NeXus standard, can immediately link the data and metadata to the ontology and make them interoperable.

15 min. break

Topical Talk TT 56.7 Thu 17:00 WIL A317

Open Research Data for Photons and Neutrons: Applications in surface scattering and machine learning — •Linus Pithan — Universität Tübingen, Institut für Angewandte Physik - DAPHNE4NFDI

Open (F.A.I.R.) research data is becoming a key ingredient for data driven machine learning (ML) applications that requires access to existing data of preceding experiments - which goes well beyond data collected in the context of one's own experiments which one might keep in a secret drawer. We will discuss current possibilities as well as future opportunities and challenges with special emphasis on surface scattering. Embedded in the DAPHNE4NFDI (DAta from PHoton and Neutron Experiments) consortium we present efforts on how data catalogs may serve as backbone for F.A.I.R. datasets provided by synchrotron and neutron sources or through community efforts. Besides suitable metadata collection also the harmonization of data- and metadata formats are issues still to be tackled especially for systematic access to fully analyzed, experimental datasets (e.g. by adopting NeXus community conventions). After a broader overview and shining light on the SciCat meta-data catalog system,[1] we discuss as application examples efforts in the field of reflectometry (XRR, NR) [2,3] and X-Ray scattering and diffraction (WAXS, GIWAXS and XPCS).[1,4]

[1] V. Starostin, L. Pithan et al. 2022, SRN, Vol. 35, No. 4
[2] A. Greco et al. 2022, J. Appl. Cryst. 55 362
[3] L. Pithan et al., Refl. dataset, 10.5281/zenodo.6497438
[4] V. Starostin et al. 2022, npj Comp. Mat. 8, 101

TT 56.8 Thu 17:30 WIL A317

NOMAD OASIS as a Tool for Electron and Atom Probe Microscopists — •Markus Kühbach — Department of Physics, Humboldt-Universität zu Berlin, Germany

Embracing the FAIR principles for sharing data and knowing how to work with different tools in research data management systems is becoming an invaluable skill in a scientist's daily life. Embracing such systems of tools, one of which is offered with NOMAD OASIS, allows you to start organizing your research data locally. Learning such tools will train you to understand what schemes and electronic lab notebooks are and how the data and metadata are processed by these tools. Example implementations of specific workflows can give you ideas where to start from and how to customize these tools for the needs of your own research and colleagues. Thereby, you can provide feedback which supports the evolution and improvement of the research data management system.

NOMAD OASIS offers you many examples which show now also how data and metadata of specific experiments can be parsed into a standardized representation. These examples teach users through detailing how data can be entered, viewed, and organized with customizable schemes in NOMAD. Furthermore, the examples suggest strategies for how the information in NOMAD can be accessed for generic or domain-specific data analytics tools.

In my talk, I will go through one or two of these examples specific to electron microscopy (orientation imaging microscopy or spectroscopy).

TT 56.9 Thu 17:45 WIL A317

FAIR Data Infrastructure for Computation: Introducing the parsers for Quantum Monte Carlo and ALF — •Jonas Schwab[1], José M. Pizarro[2], Jefferson Stafusa E. Portela[1], Luca M. Ghiringhelli[2], and Fakher F. Assaad[1] — [1]Institut für Theoretische Physik und Astrophysik and Würzburg-Dresden Cluster of Excellence ct.qmat, Universität Würzburg, 97074 Würzburg, Germany — [2]Institut für Physik und IRIS-Adlershof, Humboldt-Universität zu Berlin

DFT calculations lead to low-energy effective models. A modern example would be Kitaev types spin Hamiltonains for $RuCl_3$. Once the model is specified, many different many-body calculations can be carried out. Here, we will concentrate on the ALF [1] implementation of the auxiliary field quantum Monte Carlo algorithm that can deal with general models that includes the ones produced by DFT calculations. Being a Monte Carlo method, the ALF library produces stochastic time series. We will discuss how to implement this workflow in the NOMAD Repository & Archive (https://nomad-lab.eu) and concentrate on a FAIR meta-data scheme. The first challenges are to define the models in a searchable way as well as standards for the Monte Carlo time series. In this talk we will discuss the present state of this project for the special case of the ALF-library and how users can exploit the benefits of the NOMAD repository to find, compare and resuse our QMC data.

[1] F. F. Assaad et al., SciPost Phys. Codebases **1** (2022).

TT 56.10 Thu 18:00 WIL A317

CAMELS - A Configurable Instrument Control Software for FAIR Data — •Alexander Fuchs[1,3], Johannes Lehmeyer[1,3], Michael Krieger[1,3], Heiko B. Weber[1,3], Patrick Oppermann[2,3], and Heinz Junkes[2,3] — [1]Lehrstuhl für Angewandte Physik, Friedrich-Alexander-Universität Erlangen-Nürnberg, — [2]Fritz-Haber-Institut der Max-Planck-Gesellschaft (FHI), Berlin — [3]FAIRmat, Humboldt-Universität zu Berlin, Berlin, Germany

We are developing a configurable measurement software (CAMELS), targeted towards the requirements of experimental solid-state physics. Here many experiments utilize a multitude of measurement devices used in dynamically changing setups. CAMELS [1] will allow to define instrument control and measurement protocols using a graphical user interface (GUI). This provides a low entry threshold enabling the creation of new measurement protocols without programming knowledge or a deeper understanding of device communication. The GUI generates python code that interfaces with instruments and allows users to modify the code for specific applications and implementations of arbitrary devices if necessary. Even large-scale, distributed systems can be implemented. CAMELS is well suited to generate FAIR-compliant output data. Nexus standards, immediate NOMAD integration and

hence a FAIRmat compliant data pipeline can be readily implemented. [1] https://github.com/FAU-LAP/CAMELS

TT 56.11 Thu 18:15 WIL A317

OpenSemanticLab: Usecase Device Repository — •Matthias A. Popp and Simon Stier — Fraunhofer ISC, Neunerplatz 2, 97082 Würzburg, Germany

Fully automated experiments provide benefits regarding precision, repeatability, as well as data quality and therefore gain more and more popularity. However, setting them up can be time consuming, especially when computer interface information has to be manually transferred from device manuals to software.

In order to implement FAIR principles precise descriptions of measurement setups and instrumentation are necessary. Currently, this results in extra workload for experimental scientists.

With our framework OpenSemanticLab, we address this shortcoming by providing a central metadata repository for scientific instruments. The ontology-based repository meets both human (GUI) and machine requirements (APIs). A device ontology helps finding and classifying devices. In an associated Python package, abstract device drivers and concrete device metadata can be combined into executable workflows. Overall, this approach not only strengthens the transparency of research according to FAIR principles, but also significantly reduces the implementation effort for complex setups.