CPP 66: Focus Session: Big Data in Aquisition in ARPES (joint session O/CPP)

Due to the advancement of both electron detectors and light sources, ARPES data is increasing in volume and complexity. This applies to ARPES performed at 3rd and 4th generation light sources as well as lab-based sources. We have reached a point where data handling, workflow management, visualization and analysis is a severe challenge and potentially become the bottleneck in our workflows rather than data acquisition itself. Currently there exist mainly isolated, i.e. lab- or facility-specific, solutions for data acquisition and file formats, metadata definitions, data-processing workflows, and analysis approaches. A community-wide ARPES (meta)data schema in the quest for reproducible, scalable and transparent data analysis is not yet established. This focus session aims to reveal the great potential for speeding up our progress by attacking certain challenges in joint efforts.

Organized by: Ralph Ernstorfer (FHI Berlin), Michael Hartelt and Martin Aeschlimann (TU Kaiserslautern)

Time: Wednesday 10:30–13:15 Location: REC C 213

Invited Talk CPP 66.1 Wed 10:30 REC C 213

Towards FAIR experimental data — •Claudia Draxl — Humboldt-Universität zu Berlin — Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany

Knowledge and understanding of materials is based on their characterization in terms of a variety of properties and functions. Surprisingly though, for only a very small number of materials this information exists. Making materials data available, opens avenues for data-driven research in terms of re-purposing (using materials for a different purpose than intended by the original work), detecting candidate materials for a given application, and finding descriptors by approaches of artificial intelligence. Prerequisite for all this is a FAIR (findable, accessible, interoperable, reusable) data infrastructure. In computational materials science, the NOMAD Laboratory (https://nomad-coe.eu) has set the stage for FAIR data [1], by offering services like free upload to the NOMAD Repository, the NOMAD Archive, the NOMAD Encyclopedia, and the NOMAD Analytics Toolkit. In this talk, I will address our concepts and first steps towards extension of this open-science platform towards experimental data and sample synthesis. Here, for instance, data volume and velocity are big issues for many measurement techniques, while large uncertainties may come from (often incompletely known) sample quality, instrumental resolution, or measurement conditions. These challenges are tackled within the non-profit association FAIR-DI (https://fairdi.eu) and FAIRmat (https://fairdi.eu/fairmat), a proposed consortium for the NFDI.

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

CPP 66.2 Wed 11:00 REC C 213

NXarpes, the Data File Standard for ARPES in NeXus — •Moritz Hoesch[1], Pavel Dudin[2], and Tobias Richter[3] — [1]DESY Photon Science, Hamburg, Germany — [2]Synchrotron Soleil, Gif-sur-Yvette, France — [3]European Spallation Source, Lund, Sweden

NeXus is a common data format for neutron, x-ray, and muon science. It is being developed as an international standard by scientists and programmers representing major scientific facilities in order to facilitate greater cooperation in the analysis and visualization of neutron, x-ray, and muon data (cited from [1]). Diamond Light Source has adopted the NeXus standard, including for the instruments HR-ARPES and nano-ARPES on beamline I05 [2]. The specific NXarpes format, deliberately focusing on the essentials and thus expandable without deviation from the standard is available to the community [3]. In this presentation I will show examples of NXarpes data files and discuss the reception of this format by the community of ARPES users.

[1] https://www.nexusformat.org; [2] https://www.diamond.ac.uk/I05 ; [3] http://download.nexusformat.org/sphinx/classes/applications/NXarpes.html

CPP 66.3 Wed 11:15 REC C 213

Handling Big Multidimensional Experimental Data on Small Desktop Computers — •Michael Hartelt, Benjamin Frisch, Tobias Eul, Eva Prinz, Marten Wiehn, Benjamin Stadtmüller, and Martin Aeschlimann — Department of Physics and Research Center OPTIMAS, TU Kaiserslautern, Germany

In the pursuit of discovering new phenomena, photoemission experiments have evolved to capture ever more information about the electronic properties of materials. Major progress was made in the parallel detection of more degrees of freedom, which can include real- or k-space, energies and spin states of the emitted electrons. Additional dimensions of the parameter space are opened up by state-of-the-art experimental techniques that vary the sample temperature, the photon energy of the light source, or the time-delay between ultrashort laser pulses. As a result, experimental datasets of a single experiment can nowadays be 4-dimensional or even more. This makes the analysis of experimental data a non-trivial task, both conceptually and computationally.

We present our approach for the easy handling of these multi-dimensional, bigger-than-memory datasets on a conventional office computer. Using the Python programming language gives us access to powerful open-source packages like h5py, opencv, numpy, pint, and pycuda. Taking advantage of these, we integrated them into a toolbox package, which manages storage of large datasets for optimized I/O performance. The user is provided with an interface based on physical context, to perform data evaluation procedures with high efficiency.

CPP 66.4 Wed 11:30 REC C 213

Data Acquisition and Treatment on a Scientific and Industrial Level — •Stefan Böttcher, Christian Fleischer, and Thorsten Kampen — SPECS Surface Nano Analysis GmbH, Voltastrasse 5, 13355 Berlin

The recent developments in angular resolved photoemission and momentum microscopy let arise scientific instruments which produce enormous amount of raw data, easily exceeding several Tb of file size. Solutions or attempts of standard data- or transfer-formats are present in many fields, such as XPS or SPM. Here we present our approach on the data acquisition and processing in the acquisition and analysis software. We show the classes of metadata available to the experiment and the routes to export the data into usable formats. The data handling and the storage of transformation is a critical aspect with the present discussion. Finally an outlook into possibilities for DIN/ISO standardization can be given.

CPP 66.5 Wed 11:45 REC C 213

Challenges in data collection at a modern cw laser-driven spin-ARPES system — Tristan Heider[1], Peter Baltzer[2], Claus M. Schneider[1], and •Lukasz Plucinski[1] — [1]FZ Jülich PGI-6, Jülich, Germany — [2]MBS AB, Uppsala, Sweden

At PGI-6 in Jülich we operate a laser-driven angle- and spin-resolved photoemission (spin-ARPES) system, based on the *A*1 hemispherical analyzer with the lens deflector (MBS AB) and a single k-point *Ferrum* spin detector (Focus GmbH). The details of the system are described in a separate talk [1].

The typical data sets at our laboratory are 3D k_x vs. k_y vs. E_{kin} spin-integrated ARPES and 2D k_x vs. k_y spin maps. The size of a typical 3D dataset is approx. 100 MB, and for a complete measurement with the 6eV *cw* laser we take 4 of such sets, two with linear and two with circular light polarizations. The angular range of a single set is approx. 35°, therefore often more than one sample position needs to be measured, and a daily dataset exceeds 1 GB.

We will discuss the data collecting techniques, the data format, the data plotting, and the data storage and backup. We typically use MATLAB for data evaluation, however, we will discuss other options that might be more efficient for quick scanning through the large 3D datasets. We will use datasets from Fe-based superconductor and from 3D topological insulator as examples, and discuss challenges in evaluation of such data.

[1] T. Heider et al. this conference.

CPP 66.6 Wed 12:00 REC C 213

Single event data processing for multidimensional photoemission spectroscopy — •Steinn Ymir Agustsson[1], Rui Patrick Xian[2], Yves Acremann[3], Maciej Dendzik[2], Kevin Bühlmann[3], Davide Curcio[4], Dmytro Kutnyakhov[5], Federico Pressacco[6], Michael Heber[5], Shuo Dong[2], Philip Hofmann[4], Martin Wolf[2], Wilfried Wurth[5], Jure Demsar[1], Laurenz Rettig[2], and Ralph Ernstorfer[2] — [1]JGU Mainz — [2]FHI Berlin — [3]ETH Zurich — [4]Aarhus University — [5]DESY Photon Science, Hamburg — [6]Uni-Hamburg

The advent of novel electron detectors has opened up the field of photoemission spectroscopy to the single event detection regime. This significantly extends the accessible multidimensional parameter space for data acquisition, but also drastically increases the output of data from such experiments to the tens of MB/s regime. Handling such data therefore requires new approaches for data treatment, but also presents the opportunity for more advanced post-processing and analysis techniques. We present a distributed workflow for processing multidimensional photoemission data into an open source unified data structure. This allows, when combined with open source analysis algorithms, to directly apply such routines on data sets obtained from different experimental setups, from large scale facilities to table-top systems.

Invited Talk CPP 66.7 Wed 12:15 REC C 213

Reproducible data analysis with Snakemake — •Johannes Köster — Algorithms for reproducible bioinformatics, Genome Informatics, Institute of Human Genetics, University of Duisburg-Essen, Hufelandstr. 55, 45147 Essen Germany

Data analyses usually entail the application of many command line tools or scripts to transform, filter, aggregate or plot data and results. With ever increasing amounts of data being collected in science, reproducible and scalable automatic workflow management becomes increasingly important. Snakemake is a workflow management system, consisting of a clean, human-readable, text-based workflow specification language and a scalable execution environment, that allows the parallelized execution of workflows on workstations, compute servers, clusters and the cloud without modification of the workflow definition. Snakemake is hugely popular and was used to build analysis workflows for numerous high impact publications. With about 350 citations in the last two years, it is one of the leading frameworks for reproducible data science. This talk will show how Snakemake can be used to easily document, execute, and reproduce data analyses.

CPP 66.8 Wed 12:45 REC C 213

Concept for Handling of Photoemission Data at European XFEL — •Markus Scholz[1], Dmytro Kutnyakhov[2], Michael Heber[2], Manuel Izquierdo[1], Hans Fangohr[1], Yves Acremann[4], Kai Rossnagel[3], Anders Madsen[1], and Serguei Molodtsov[1] — [1]European XFEL Facility, Holzkoppel 4, 22869 Schenefeld, Germany — [2]Deutsches Elektronen-Synchrotron DESY, 22607 Hamburg, Germany — [3]Ruprecht-Haensel-Labor, Christian-Albrechts-Universität zu Kiel and Deutsches Elektronen-Synchrotron DESY, 24098 Kiel and 22607 Hamburg, Germany — [4]Laboratorium für Festkörperphysik, ETH Zürich, 8093 Zürich, Switzerland

European X-ray Free Electron Laser (EuXFEL) is currently the world's biggest, brightest and highest repetition rate XFEL providing up to 27000 pulses/second. The planned open port named "Soft X-ray Port" (SXP), will allow time-resolved X-ray photoelectron spectroscopy (TR-XPES) experiments. In this contribution I will present how near-online analysis of photoemission data based on Jupyter notebooks could be realized and embedded in the EuXFEL software framework. For compute-intensive notebooks, it is possible to allocate dedicated nodes with user-specified hardware configuration from the Maxwell computer cluster to a running JupyterHub session. This is of particular value due to the size of data sets and the the remotely accessible analysis.

CPP 66.9 Wed 13:00 REC C 213

Processing workflow for band structure reconstruction from multidimensional photoemission data — R. Patrick Xian[1], Vincent Stimper[2], Shuo Dong[1], Maciej Dendzik[1], Samuel Beulieu[1], Bernhard Schölkopf[2], Martin Wolf[1], Stefan Bauer[2], Laurenz Rettig[1], and •Ralph Ernstorfer[1] — [1]Fritz Haber Institute of the Max Planck Society, Berlin, Germany — [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

Recent advances in photoelectron detectors and light sources result in an increase of size and dimensionality of photoemission data, leading to new challenges in data preprocessing and analysis: the large number of adjustable parameters of modern electron optics require experiment-specific calibration and artifact correction [1]; visual inspection of multidimensional ARPES data may be hampered by the varying levels of contrast [2]. We discuss a workflow for conditioning volumetric three- and four-dimensional momentum microscopy data for band structure mapping from single-electron events to calibrated, reusable data [3].

[1] Xian et al., Ultramicroscopy 202, 133 (2019).
[2] Stimper et al., IEEE Access 7, 165437 (2019).
[3] Xian et al., arXiv 1909.07714.