ISBN: 978-3-95450-221-9 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2021-WEBL01

FAIRMAT - A CONSORTIUM OF THE GERMAN RESEARCH-DATA INFRASTRUCTURE (NFDI)

H. Junkes*, P. Oppermann, R. Schlögl, A. Trunschke, Fritz Haber Institute, Berlin, Germany
M. Krieger, H. Weber, Universität Erlangen-Nürnberg, Erlangen, Germany

Abstract

The FAIRmat project, was selected on Friday, July 2, 2021 by the German Joint Science Conference (Gemeinsame Wissenschaftskonferenz – GWK) in a multi-stage competition of the National Research Data Infrastructure (NFDI). The project will receive funding to set up an infrastructure that helps making materials–science data FAIR: Findable, Accessible, Interoperable, and Reusable. This will enable researchers in Germany and beyond to store, share, find, and analyze data over the long term. During the five-year term, a total of 60 project leaders from 34 German institutions will work together in the FAIRmat consortium[1].

FAIRMAT CONSORTIUM

When applying for funding for the FAIRmat project, the following Objectives, work program and research environment were described in the letter of intent:

> Prosperity and lifestyle of our society are very much governed by achievements of condensed-matter physics, chemistry, and materials science, because new products for the energy, environment, health, mobility, and IT sectors, for example, largely rely on improved or even novel materials. The enormous amounts of research data produced every day in this field, therefore, represent the treasure trove of the 21st century. This treasure trove is, however, of little value, if these data are not comprehensively characterized and made available. How can we refine this feedstock, in other words, turn data into knowledge and value? For this, a FAIR data infrastructure is a must.
>
> Here is where FAIRmat ("FAIR Data Infrastructure for Condensed-Matter Physics and the Chemical Physics of Solids") comes in. By building a FAIR research-data infrastructure for the noted fields, the consortium will lift the treasure trove of materials data, and therewith contribute to a disruptive change in the way science and R&D are conducted. Within FAIRmat the acronym FAIR is interpreted in a forward-looking way: Research data should be Findable and Artificial-Intelligence Ready. This new perspective will advance scientific culture and practice. This will not replace scientists, but scientists who use such FAIR infrastructure may replace those who don't.

Projekt Plan

FAIRmat will install a FAIR [1] data infrastructure for the wider area of condensed-matter physics and the chemical physics of solids. This represents a very broad range of different communities that can be characterized by either different classes of condensed matter (e.g. semiconductors, metals and alloys, soft and biological matter, etc.), by different techniques (e.g. ranging from crystal-growth and synthesis to experimental and theoretical characterization by a multitude of probes), or by functionality (exemplified here by battery materials, optoelectronics, catalysts, etc.). As a consequence, the data produced by the community are enormously heterogeneous and diverse in terms of the 4V of Big Data

- Volume (amount of data)
- Variety (heterogeneity of form and meaning of data)
- Velocity (rate at which data may change or new data arrive)
- Veracity (uncertainty of data quality).

Also note that many research data produced today may appear irrelevant in the context they have been produced. Being regarded as *waste*, they are not published. However, they may turn out highly valuable for other purposes.

So, the R in FAIR (reusability) also means "store, share, and recycle the waste!" To cope with all the diversity and complexity, a bottom-up approach that satisfies the needs of the different sub-communities is a must to foster acceptance by the community and participation of a large number of individual researchers and laboratories. FAIRmat sets out to tackle this challenge by a user-driven approach to develop easy-to-use tools and an infrastructure towards FAIR data processing, storage, curation, sharing, and future use of materials data. For the latter, a major goal of FAIRmat is making data artificial-intelligence (AI) ready.

Data obtained by a certain experimental technique for a specific sample of a selected material are only worth keeping if the sample is fully characterized and apparatus and measurement conditions as well as the measured quantity are described in detail. Likewise, computed data are only meaningful when method, approximations, code and

* junkes@fhi.mpg.de

[1] FAIRmat press release

ISBN: 978-3-95450-221-9 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2021-WEBL01

code version, as well as all computational parameters are known. In essence, we need an extensive annotation, i.e. a systematic *metadata* catalogue, also covering ontologies. This also includes the description of provenance and data quality (their usefulness for a given context).

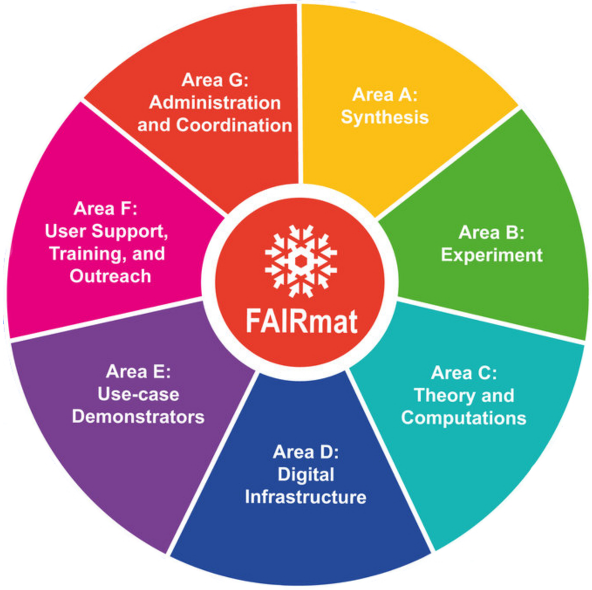

Figure 1: FAIRmat task areas.

To address all these aspects and to support basic science and individual researchers, we have identified several task areas (in short called Areas, see Fig. 1) that are sketched in the following:

- Area A – Synthesis – is dedicated to the full characterization of samples and the corresponding synthesis and growth processes. Without this information, reproducibility of materials and their properties with given quality will be hampered. The specific tasks will consider various synthesis routes, i.e., from the gas, liquid, and solid phases, and by assembly.
- Area B – Experiment – covers the microscopic characterization of materials by a broad variety of measurement techniques. Each of them comes with specific challenges concerning processing, curation, and storage, owing to differences in volume, velocity, data formats, etc. FAIRmat will exemplify its approach in a first phase by a representative selection of experimental probes.
- Area C – Theory and Computation – deals with numerical techniques to compute materials properties. Such techniques differ in the theoretical concepts, approximations, and numerical recipes, depending on the employed software. Our general approach to tackle this diversity relies on the concepts developed by the NOMAD Laboratory [2]. The tasks of Area C concern extensions towards excitations and strongly correlated materials, as well as towards classical (particle-based) simulations and multi-scale modeling.
- Area D – Digital Infrastructure – will be a common brace to all other areas. Specific tasks will be dedicated to processing and decentral storage, creating a network of data hubs at different locations. The FAIRmat Portal, a web-based graphical user interface, including a Materials Encyclopedia [3], will allow for searching, accessing, and inspecting (meta)data from all over Germany (and worldwide). FAIRmat will develop and provide tools, from processing to post-processing, including analysis by artificial intelligence, and it will provide guidance and advancements of electronic lab-books (ELN), laboratory information management systems (LIMS), etc. Tight collaborations with HPC centers will ensure the embedding into the overall NFDI landscape.
- Area E – Use-Case Demonstrators – will highlight how the tools developed in the above areas will benefit different scientific communities and demonstrate hand-shakes and potential synergies with other consortia. The specific tasks of this area will cover in the first phase use cases on battery materials, heterogeneous catalysis, optoelectronics, spintronics&magnetism, metalorganic frameworks, biological physics applications, and artificial intelligence.
- Area F – User Support, Training&Outreach – will reflect our concept for how to engage with the community, to allow researchers to make use and handle the FAIRmat tools. A variety of workshops and other training opportunities will be offered and support for connecting data hubs to the overall data infrastructure.
- Area G – Administration and Coordination – will deal with all coordination and management issues and the embedding into the Overall NFDI. As such, it will address synergies with other consortia and the interaction with the NFDI Directorate.

FAIRmat represents a broad community of numerous researchers from universities and leading institutions in Germany. It builds on extensive experience with the worldwide biggest data infrastructure in computational materials science, the

Novel Materials Discovery (NOMAD) Laboratory and the association FAIR-DI e.V. FAIRmat aims at covering the full breadth of the Condensed Matter Section of the German Physical Society (DPG) with its 12 divisions, and is further supported by the Chemistry, Physics, and Technology Section of the Max Planck Society, the Bunsen Society for Physical Chemistry, and more. It is fully embedded internationally, e.g., in the Research Data Alliance, the European Open Science Cloud, GO FAIR, etc. and has signed Memoranda of Understanding with leading institutions worldwide, for example NIST (USA), Shanghai University (China), and CSC (Finland). FAIRmat will continue to raise awareness and acceptance of a FAIR research-data infrastructure in Germany, Europe, and beyond. [4, 5]

Task D5

A universal and easy-to-configure software environment for measurement data acquisition and documentation is to be developed in Task D5 of the FAIRmat consortium.

In the field of Applied Physics, measurement setups with numerous specific measurement devices are often required – in each case adapted to the experimental problem. The diversity requires adaptable and easy-to-configure software for experiment control and data acquisition. But it's not just about the raw data. The experiment description, including all settings of the laboratory equipment used, the *metadata*, is also required. Only in this way, the experiment is documented completely and FAIR, and the valuable measurement data can be re-used by other scientists.

A prototype software developed at the chair of the Department of Physics at the Friedrich-Alexander-Universität (FAU) Department for experiment control with uniform and documented data output has been using for years [6]. In the FAIRmat project this successful concept will be put together with the open-source Experimental Physics and Industrial Control System (EPICS) [7].

CONFIGURABLE EXPERIMENTAL CONTROL SYSTEM (CECS)

An existing system developed at the Department of Physics (FAU) has been built based on LabVIEW [8].

This system has proven itself at the department in research and teaching (Fig. 2). However, it does not satisfy the FAIR rules. By using LabVIEW, the system is:

- not open source,
- not operating system independent,
- not platform independent,
- is very poorly scalable.

Because of these missing features, the new CECS will be based on EPICS to control the scientific and measurement equipment [9].

Although EPICS is also suitable for small experiments [10], there is a persistent opinion that it can only be used by experts and larger teams of developers. This may also have to do with the fact that the aspect of data acquisition is rarely taught at universities nowadays. One aspect in the development of the new system is the possibility to introduce the users to the data acquisition step by step and transparently.

The system shall provide an easy to use graphical experiment configuration interface like the LAP Measurement before. We decided to implement this configuration interface in python. The packages from Python for Scientific Computing (Fig. 3) are used.

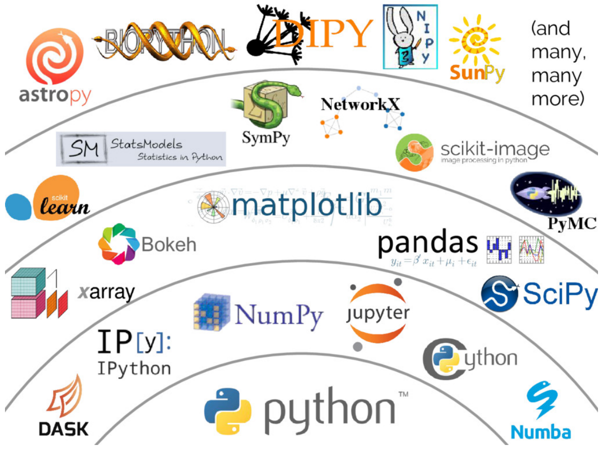

Figure 3: Python for Scientific Computing.

Python has now established itself as a programming language, especially in scientific subjects, and is taught at universities. The premise of CECS is the ability to configure a measurement system graphically without having to have programming knowledge. However, the GUI itself should not yet represent a direct implementation of a measuring program but generate a human (and machine) readable output which then serves as a *recipe* for a measuring program. This gives the possibility to let the configuration system generate a template (python), which one can refine and adapt to special needs. This makes it possible to get deeper and deeper into the system and to adapt it. The experience has shown that the scientist is willing to improve such templates if everything is presented transparently. The configuration system can also be used to specify general structures (e.g. storage location, file format, etc.) and thus achieve a standardized data recording. These *recipes* are excellent for documenting the workflows and can be linked to the documentation system (ELNs, etc). This also makes it possible to save these work steps in a versioning system. This makes it possible to repeat measurements.

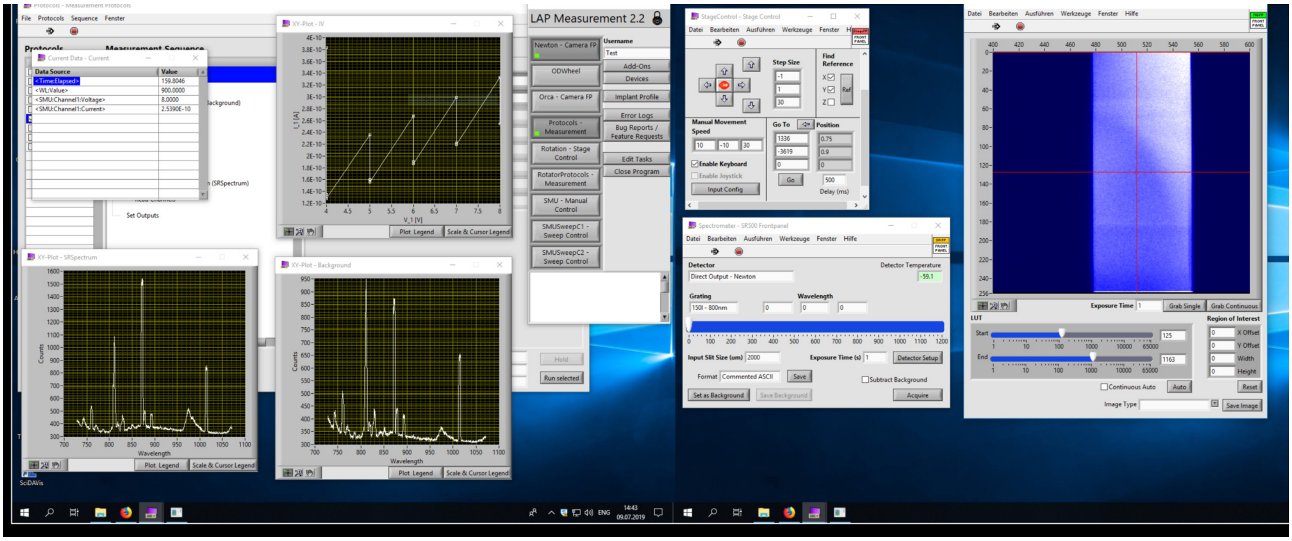

Figure 2: Example of the LAP Measurement User Interface.

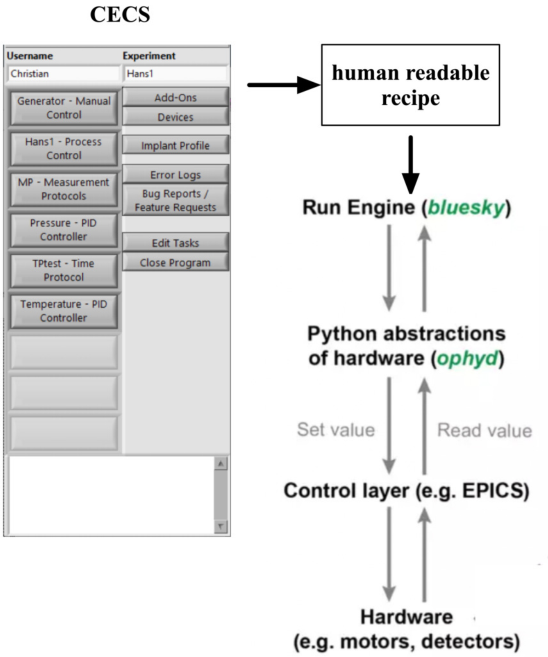
CECS
Username
Experiment
Christian
Hans1
Generator - Manual Control
Hans1 - Process Control
MP - Measurement Protocols
Pressure - PID Controller
TPtest - Time Protocol
Temperature - PID Controller
Add-Ons
Devices
Implant Profile
Error Logs
Bug Reports / Feature Requests
Edit Tasks
Close Program
human readable recipe
Run Engine (bluesky)
Python abstractions of hardware (ophyd)
Set value
Read value
Control layer (e.g. EPICS)
Hardware (e.g. motors, detectors)

Figure 4: CECS recipe for Bluesky.

With the help of this *recipe*, a plan and virtual devices for Bluesky [11] are created. Bluesky stands on the shoulders of EPICS, and provides additional capabilities such as live visualization and data processing tools, and can export data into nearly any file format in real time. Bluesky was developed using python. That will make Bluesky simple for future scientists to modify, and to implement to new experiments.

EPICS

Complex systems EPICS is a set of software tools and applications which provide a software infrastructure for use in building distributed control systems to operate devices such as Particle Accelerators, Large Experiments and major Telescopes. Such distributed control systems typically comprise tens or even hundreds of computers, networked together to allow communication between them and to provide control and feedback of the various parts of the device from a central control room, or even remotely over the internet.

High bandwidth, soft realtime networking applications EPICS uses Client/Server and Publish/Subscribe techniques to communicate between the various computers. Most servers (called Input/Output Controllers or IOCs) perform real-world I/O and local control tasks, and publish this information to clients using robust, EPICS specific network protocols Channel Access and pvAccess. These protocols are designed for high bandwidth, soft real-time networking applications that EPICS is used for, and is one reason why it can be used to build a control system comprising hundreds of computers.

Flexible and scalable E.g. at the Advanced Photon Source national laboratory in the United States, EPICS is used extensively within the control system for the accelerator (see Fig. 5) and many of the experiments. There are about hundreds of IOCs that directly or indirectly control almost every aspect of the machine operation, while 40 workstations and servers in the control room provide higher level control and operator interfaces to the systems, and perform data logging, archiving and analysis.

Roadmap EPICS is developed through a collaborative open-source process where anyone is free to contribute to the EPICS family of software. In addition to this, heavy EPICS users – typically large scientific facilities – gather together in an open EPICS council to define a roadmap for the future direction of EPICS.

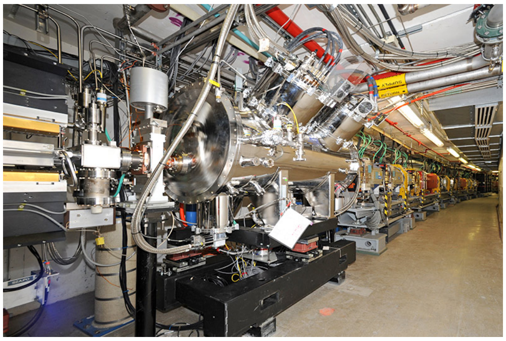

Figure 5: Beampipe at APS.

License EPICS is provided under an open source license called the EPICS Open License, which is similar to the BSD license. The EPICS licensing page [12] gives details and some history.

Devices supported The IOC, the Input/Output Controller is the I/O server component of EPICS. Almost any computing platform that can support EPICS basic components like databases and network communication can be used as an IOC. One example is a regular desktop computer, other examples are systems based on real-time operating systems like vxWorks or RTEMS and running on dedicated modular computing platforms like MicroTCA, VME or CompactPCI. EPICS IOC can also run on low-cost hardware like RaspberryPi or similar.

IOCs can support any number of records and record types. Similarly, record support does not contain device-specific knowledge, so each record type can have any number of independent device support modules. If the method for accessing hardware is more complicated than device support, then a device driver can be developed.

Currently, there are already a number of modules available to support instruments that are also commonly used in many experiments, such as temperature controllers (Eurotherm) and mass flow controllers (Bronkhorst) [liste]. These individual instruments can be connected to the IOC via serial or Ethernet interfaces.

Bluesky

Bluesky [11] is a mini-ecosystem of co-developed but individually useful Python libraries for experiment control and data acquisition, management, and access. The project is developed and maintained by a multi-facility collaboration. The core includes a high-level hardware abstraction above EPICS, an experiment orchestration engine, a formally-defined schema for streaming data and metadata, and data access tools integrated with the open source scientific Python stack (see Fig. 6). It emphasizes the following virtues:

- Live, Streaming Data: Available for inline visualization and processing.
- Rich Metadata: Captured and organized to facilitate reproducibility and searchability.
- Experiment Generality: Seamlessly reuse a procedure on completely different hardware.
- Interruption Recovery: Experiments are *rewindable*, recovering cleanly from interruptions.
- Automated Suspend/Resume: Experiments can be run unattended, automatically suspending and resuming if needed.
- Pluggable I/O: Export data (live) into any desired format or database.
- Customizability: Integrate custom experimental procedures and commands, and get the I/O and interruption features for free.
- Integration with Scientific Python: Interface naturally with numpy and Python scientific stack.

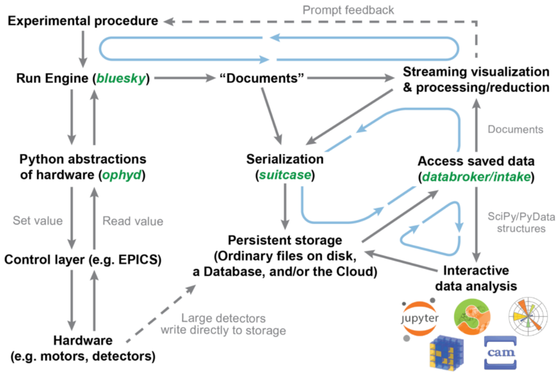
Prompt feedback
Experimental procedure
Run Engine (bluesky)
"Documents"
Streaming visualization & processing/reduction
Documents
Python abstractions of hardware (ophyd)
Serialization (suitcase)
Access saved data (databroker/intake)
Set value
Read value
SciPy/PyData structures
Control layer (e.g. EPICS)
Persistent storage (Ordinary files on disk, a Database, and/or the Cloud)
Interactive data analysis
Large detectors write directly to storage
Hardware (e.g. motors, detectors)
jupyter
cam

Figure 6: Bluesky project.

APPLICATION AREAS

Another example of the ideal use of such a low-threshold data acquisition system is in catalysts research.

The challenges of the energy transition can only be solved with the help of innovative catalyst technologies. In catalysis research, a number of empirical concepts exist and predictions of theory are based on first-principles-based models that are subject to a variety of assumptions [13]. For simple reactions, material-function relationships have been predicted very successfully in this way. However, this is much more difficult for complex reactions and catalyst systems that are subject to high dynamics. Energy-relevant reactions, such as the hydrogenation of carbon dioxide, or important catalytic conversions for chemical hydrogen storage, are such complex reactions. Here, catalyst dynamics, i.e. the change in the state of the catalyst depending on the history of the catalyst and the reaction conditions, complicates the application of Big Data analyses and data mining methods because experiments performed according to different workflows can lead to inconsistent data sets. On the other hand, the complex relationships between catalyst structure and functional properties of the catalyst can probably only be decoded with reasonable effort using artificial intelligence methods. For this purpose, reliable, reproducible data sets with high diversity are required. Handbooks in which the characterisation of catalysts and the determination of kinetic data are precisely prescribed serve to generate such data. These handbooks should specify what minimum data set should be generated for each catalyst and how the measurements should be performed. An example of a workflow in catalysis research is shown in Figure Fig. 7.

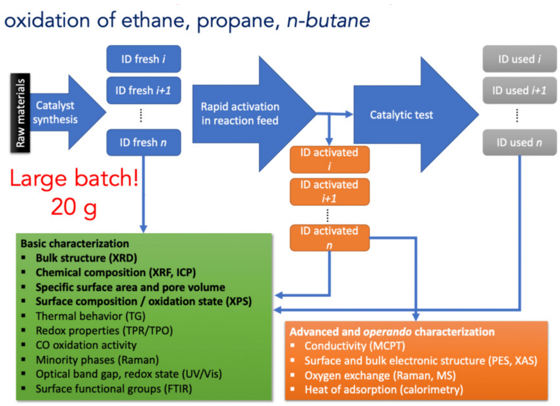
oxidation of ethane, propane, *n-butane*
Raw materials
Catalyst synthesis
ID fresh *i*
ID fresh *i+1*
ID fresh *n*
Large batch!
20 g
Rapid activation in reaction feed
ID activated *i*
ID activated *i+1*
ID activated *n*
Catalytic test
ID used *i*
ID used *i+1*
ID used *n*
Basic characterization
Bulk structure (XRD)
Chemical composition (XRF, ICP)
Specific surface area and pore volume
Surface composition / oxidation state (XPS)
Thermal behavior (TG)
Redox properties (TPR/TPO)
CO oxidation activity
Minority phases (Raman)
Optical band gap, redox state (UV/Vis)
Surface functional groups (FTIR)
Advanced and *operando* characterization
Conductivity (MCPT)
Surface and bulk electronic structure (PES, XAS)
Oxygen exchange (Raman, MS)
Heat of adsorption (calorimetry)

Figure 7: Workflow catalyst investigation.

These handbooks can then be converted into readable (human and machine) formats which can then be e.g. read as *recipes* by Bluesky.

For a typical setup in catalayse research see Fig. 8.

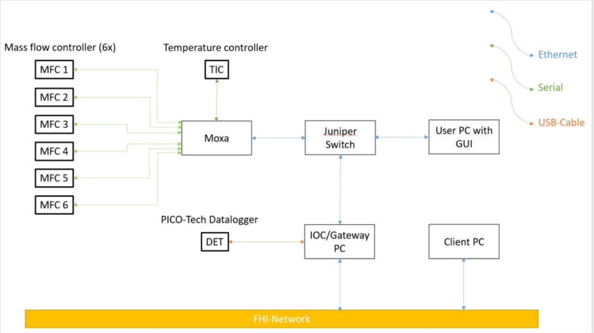

Figure 8: Block diagram of a reactor for testing catalysts.

A gateway computer (see Fig. 9) sets up its own experiment subnet and collects all the data. It is equipped with two Ethernet interfaces and up to 20 serial interfaces. In addition, custom device drivers were written to communicate with analog-to-digital converters, the Modbus protocol is used for this purpose.

ACKNOWLEDGEMENTS

Alexander Fuchs and Johannes Lehmeyer (FAU) are implementing the new CECS system as part of the FAIRmat project. Abdulrhman Moshantaf (FHI) implements the control system for the CatLab reactor systems.

Figure 9: Jetway IOC.

REFERENCES

[1] FAIR principles, https://www.go-fair.org/fair-principles/.

[2] NOMAD-Laboratory, https://nomad-lab.eu/.

[3] NOMAD materials encyclopedia, https://www.nomad-coe.eu/index.php?page=materials-encyclopedia

[4] FAIRmat-LoI, https://www.fair-di.eu/uploads/documents/FAIRmat_LoI_2020_Website.pdf

[5] FAIRmat-Template for Submission of LoI, https://www.fair-di.eu/uploads/documents/FAIRmat_Abstract.pdf

[6] C. Ott "LAP Measurement v2.2: Talk", Private communication, LAP_Measurements_Kaffeeseminar_2019-11-25

[7] EPICS: Collaboration website, about EPICS, https://epics-controls.org/about-epics/.

[8] LabVIEW: Company website, https://www.ni.com

[9] FAIRmat: Lifting the treasure trove of materials data https://www.physics.nat.fau.eu/2021/07/05/fairmat-lifting-the-treasure-trove-of-materials-data

[10] H. Junkes, "EPICS Also for Small and Medium Sized Experiments", in *Proc. ICALEPCS'19*, New York, NY, USA, Oct. 2019, pp. 1269–1272. doi:10.18429/JACoW-ICALEPCS2019-WEPHA075

[11] Bluesky: Project description, https://blueskyproject.io/.

[12] EPICS: License information, https://epics-controls.org/licensing/.

[13] A. Trunschke *et al.*, "Towards Experimental Handbooks in Catalysis", *Topics in Catalysis*, vol. 63, pp. 1683–1699, 2020. doi:10.1007/s11244-020-01380-2