ISBN: 978-3-95450-221-9 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2021-WEBL05

FAIR MEETS EMIL: PRINCIPLES IN PRACTICE

Gerrit Günther*[1], William Smith[1,2], Markus Kubin[1], Marcus Bär[1–4], Nico Greve[1], Rolf Krahl[1], Simone Vadilonga[1], Regan G. Wilks[1,2], Oonagh Mannix[1],
[1] Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Lise-Meitner-Campus, Hahn-Meitner-Platz 1, 14109 Berlin, Germany
[2] Energy Materials In-Situ Laboratory Berlin (EMIL), HZB, Albert-Einstein-Str. 15, 12489 Berlin, Germany
[3] Helmholtz Institute Erlangen-Nürnberg for Renewable Energy (HI ERN), Albert-Einstein-Str. 15, 12489 Berlin, Germany
[4] Friedrich-Alexander Universität Erlangen-Nürnberg (FAU), Egerlandstr. 3, 91058 Erlangen, Germany

Abstract

Findability, accessibility, interoperability, and reusability (FAIR) form a set of principles required to ready information for computational exploitation. [1] The Energy Materials In-Situ Laboratory Berlin (EMIL) at BESSY II [2], with its unique analytical instrumentation in direct combination with an industrially-relevant deposition tool, is in the final phase of commissioning. It provides an ideal testbed to ensure workflows are developed around the FAIR principles; enhancing usability for both human and machine agents. FAIR indicators [3] are applied to assess compliance with the principles on an experimental workflow realized using Bluesky. [4] Additional metadata collection by integrating an instrument PID [5], an electronic laboratory book, and a sample tracking system is considered along with staff training. Data are collected in Nexus format and made available in the ICAT repository. This paper reports on experiences, problems overcome, and areas still in need of improvement in future perspectives.

INTRODUCTION

The FAIR principles – findability, accessibility, interoperability and reusability – are well known and serve as guidelines for data curators and data management stewards along with IT professional; people who manage rather than generate data. Here, we apply high-level data management concepts contained in the FAIR principles on the level of granularity of a complex, beamline-based materials research infrastructure – the EMIL infrastructure at BESSY II, that is in its final phase of commissioning. In doing so we clarify the relationship between data generator, research infrastructure, and data infrastructure; demonstrating how the FAIR principles relate to each actor.

To report the present state of work, the remainder of this paper is organized as follows: First, we sketch the complex scientific infrastructure of EMIL at BESSY II focusing on its end station SISSY I (**S**olar energy material **I**n-**S**itu spectroscopy at the **Sy**nchrotron). Since the FAIR principles are independent of each other [1] we continue with a discussion of their implementation following the sequence reusability, interoperability, accessibility, and findability. By balancing user effort and data accuracy we proceed to focus on the sequence of measures taken to find an appropriate workflow. Finally, we apply FAIR indicators and assess the SISSY I end station.

Throughout the paper, the convention of [3] is adopted to distinguish between data and metadata. Here, data refers to the primary output of detectors or other objects of outstanding scientific interest while metadata belongs to information that helps to analyze the primary data such as the sample or the sample's properties. If both types of data are addressed the term (meta)data is used.

OUTLINE OF EMIL INFRASTRUCTURE

The Energy Material In-Situ Laboratory Berlin (EMIL) [2,6,7] is a newly implemented, complex research infrastructure at BESSY II which combines state-of-the-art synthesis and deposition systems on industry-relevant scale with sophisticated in system/in situ/operando x-ray spectroscopy capabilities, outlined in Fig. 1. By focusing on study of the growth of materials and the formation of functional interfaces EMIL tracks the evolution of material properties through successive processing steps. A fully automated ultra-high vacuum multi-chamber system built by PREVAC [8] transfers samples between various production sites, analysis tools, and the SISSY I end station. As a result samples are changing both their state and location within EMIL and may be placed at in SISSY I end station for characterization.

The SISSY I end station houses an electron analyzer able to detect electrons with kinetic energies between 50 eV and 10 keV. Exploiting the full energy range of EMIL's two-color beamline, photon energies between 80 and 10000 eV – delivered by the two canted undulators (UE48 and U17) – can be used to probe the sample properties by soft and hard x-ray photoelectron emission electron spectroscopy. [9] Currently, the SISSY I end station is in the final stages of commissioning and so provides an excellent test bed to implement the FAIR principles during the transition to user operation.

The instrument and beamline are controlled with EPICS [10]. Devices are accessed in a Python environment through

* gerrit.guenther@helmholtz-berlin.de

ISBN: 978-3-95450-221-9 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2021-WEBL05

the abstraction layer Ophyd [11]. The Python library Bluesky [4] uses this abstraction layer to operate the instrument and beamline hardware. Measurement data and metadata are saved together in a mongoDB database [12], which is not publicly accessible. A more detailed outline of the architecture is found in [13].

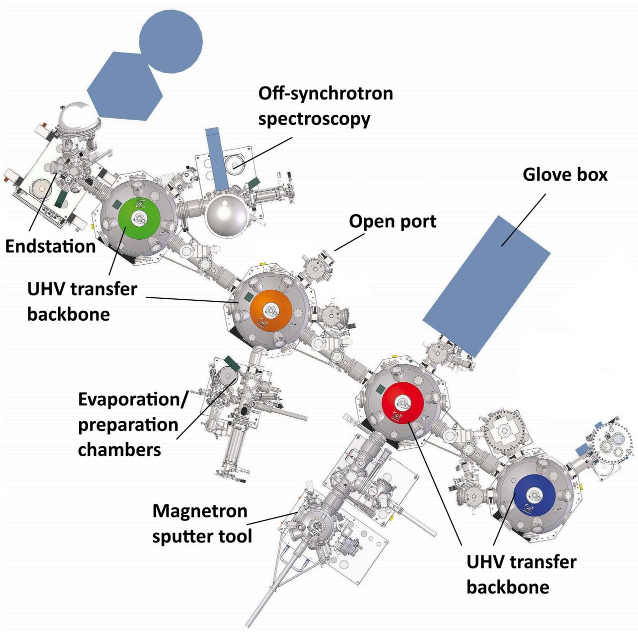

Figure 1: Schematic presentation of the SISSY@EMIL's setup connecting various processing and analysis chambers, among them the SISSY I endstation.

REUSABILITY – RICHNESS OF (META)DATA

The SISSY I operation is envisioned to be highly automated. It provides well-defined experimental environments enabling the automatic collection of detailed (meta)data. Unfortunately, not all critical metadata can be collected automatically and crucial information such as sample name, calibration measurements, special setups, or notes during the experiment may be buried in a paper log book and run the potential risk of being permanently lost.

Sample Information

In most cases sample information cannot be automatically collected but is critical for data reuse. For that reason, we implement the collection of sample information on the instrument level by defining a sample pseudo-axis in the Bluesky control software of SISSY I to enter the sample name, chemical formula, InChi key [14], or another sample identifier. The sample axis is treated in the same way as any motor position which would allow to define the sample by a script running for a lengthy period for example, when a sample holder with several samples is used.

At SISSY I all samples are transferred through the EMIL vacuum transport system, a highly automated environment in which samples may undergo successive steps of physical treatments and analysis. Each operation is stored in a SQL database and therefore the entire history of a sample is accessible. We implemented a readout routine to update the sample-pseudo axis on demand. If necessary, this can be corrected or enriched by the user during the experiment. It automatically identifies the sample holder currently at SISSY I through the sample database by using the date and time of the measurement. Once the sample holder is identified, the sample information is extracted from the sample database and former entries of the sample are collected in a chronological list of applied processes and measurements. In case the initial state of the sample when entering EMIL corresponds to an industrial standard (such as a silicon wafer) the sample would be fully defined.

For other cases we provide workarounds to add information manually. We implemented a graphical user interface to help describing the sample and enrich (meta)data by additional files such as photos, notes, or documents after the measurement. Further, we implemented a routine to trawl pdf documents from the HZB's proposal system GATE for samples which can be selected from a drop-down menu to add their information. However, direct access to the proposal system would be advantageous and has to be established.

Electronic Logbook – Another Kind of Metadata

Paper logbooks are still used at most BESSY II beamlines and contain useful or, even essential information for scientific reproducibility. Details of sample preparation, other analytical results, or even the experimental methodology can be found in these log books; additional qualitative information such as the scientists thoughts at the time of measurement, or classification in a wider context can also be found. Such intellectual interpretation of data belongs to another type of metadata that cannot be harvested automatically – but that can be stored electronically. To provide an electronic option of paper notebooks an instance of the ESRF e-logbook [15] was installed as a BESSY II-wide service. The logbook can include comments from the experimental team, as well as comments written automatically by the machine. Information about the current plan being executed is written by Bluesky in a callback.

Since Jupyter notebooks enjoy great popularity in the community to combine code, notes, and visualizations of results within a single file [16], and they are native to the Python Bluesky environment, their integration in a reproducible way is under discussion.

Persistent Identifiers

The provided information should be unambiguous; using globally unique persistent identifiers ensures clarity when considering important metadata. For example, an ORCID [17] distinguishes multiple individuals clearly even if they share the same name. For this reason, at SISSY I ORCIDs are automatically assigned to members of the experimental team if they are identified through name and facility.

ISBN: 978-3-95450-221-9 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2021-WEBL05

Moreover, globally unique instrument PID [5] identify the instrument SISSY I [18], the soft [19] and hard [20] x-ray EMIL beamlines, and other parts of EMIL [21,22] to clearly state from where the data originates and where the instrument belongs. In the future, more PIDs will be included to precisely categorize important metadata, not least a semi-automated, globally unique sample identification system.

INTEROPERABILITY – SPEAKING THE SAME LANGUAGE

SISSY I stores metadata parameters acquired by Bluesky in a Mongo database. The names of these parameters are specific to this facility, beamline, and endstation. So that this metadata is eventually understandable to a wider audience we implemented a routine that converts the database content of a measurement to the NeXus file format [23] which is a well-known standard of the x-ray and neutron community. NeXus uses HDF5 [24] as physical file format and applies the NeXus Definition Language (NXDL) to clearly define the nomenclature and arrangement of information while being versatile enough to meet the demands of a wide range of complex scientific techniques.

The file format is suitable for high-performance data processing [25, 26] and is supported by capable analysis software such as Dawn [27] which provides a Python interface following the spirit of the Bluesky control software.

ACCESSIBILITY – TWENTY-FOUR/SEVEN

Well-curated data is useful when revisiting work after some time or when continuing another's work, but unleashes its full potential when accessible to co-workers, instrument users, and collaborators from around the world. That is why NeXus files are stored in HZB's central data repository [28] which is an instance of the ICAT software [29, 30] also in use at other Photon and Neutron facilities, including ISIS, Diamond Light Source, and ESRF.

ICAT's authentication system allows the experimental team to access the files associated with its proposal through a web interface using the HTTP standard protocol. According to the HZB data policy [31] raw data and associated metadata become openly accessible after a embargo period of normally five years.

However, the HZB data repository stores data on a tape library which has drawbacks concerning accessibility in general. In order to access the data, a request must be made in the metadata catalogue. The data will then be staged to disk automatically and a download link will be provided once the staging is complete. This might cause accessibility barriers to machine-agents, but there is currently no rigid procedure to deal with the situation.

FINDABILITY – 'HELLO, WORLD'

SISSY I's local (meta)data storage procedure using Mongo database is more advanced than most current BESSY II instruments that store data and metadata in various files. The Mongo database is completely searchable and returns measurements and metadata without laborious scans of a (potentially) confusing file structure. The same advantage applies to the ICAT's metadata catalogue which stores selected (meta)data of ingested files in a separate database and is also searchable. The difference is that ICAT is publicly accessible and centrally managed.

To increase visibility we need higher level services than those of HZB to connect (meta)data and the world – we need globally unique persistent identifiers (PID) to give measured raw data an unambiguous identity. While for (processed) data publications a manual procedure for adding PIDs exists, ingested raw data files in ICAT thus far lack this feature. However, the automatic assignment of PID's to single NeXus files containing raw data is a prerequisite to be harvested by higher level catalog services such as B2Find [32] and is currently in the implementation phase. This and the Harmonization of the metadata schema with regard to B2FIND is work in progress and ensure that (meta)data collected at SISSY I will be globally findable once the embargo period has ended. Such high-level search engines are crucial hubs in the world-wide data network and make the instrument's work public to visible to an community beyond that of the repository.

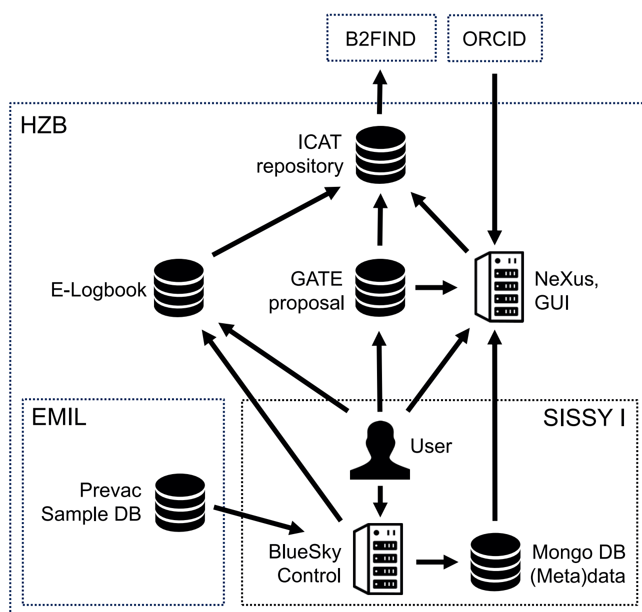

Figure 2: Visualization of the data infrastructure connecting the SISSY I instrument with services inside and outside the HZB.

WORKFLOW

In the previous sections, we presented tools and services that are implemented or under consideration. We purposefully did not define a unique sequence of executed steps. Since different experiments involve different procedures, we adapt the workflow to the needs of each experiment.

ISBN: 978-3-95450-221-9 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2021-WEBL05

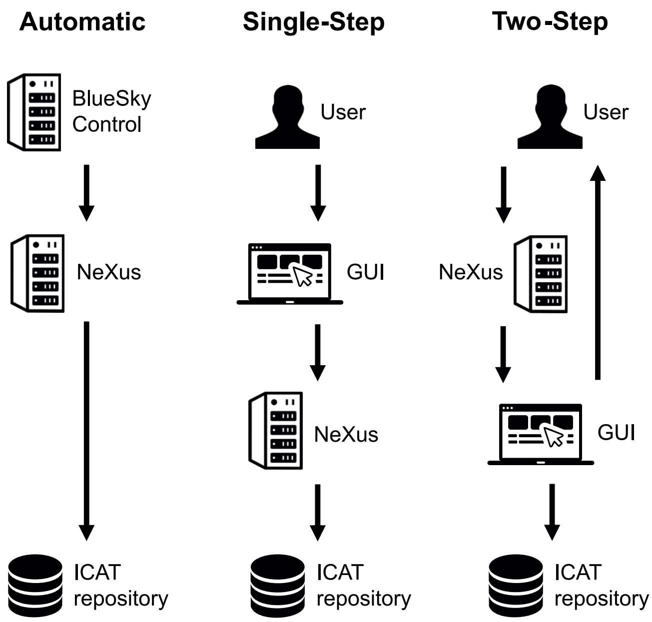

Figure 3: Provided workflows for the SISSY I instrument including automatic and manual processes.

Metadata may become available at different times of the experiment or even after its completion and thus the data management process may become stuck somewhere requiring human input to make data meaningful.

However, general principles guide the implementation of each workflow. Whenever possible procedures are outsourced to a physically separated server to keep the impact of the data management process on the experiment to a minimum. The parallel, mostly independent infrastructure is suitable to update running instruments without interrupting the production of results which is a major problem in fragmented software landscapes such as BESSY II [33].

The central data conversion server runs (i) the readout routine to get (meta)data from SISSY I's Mongo database, (ii) optionally the readout routine to collect sample information from the sample database, (iii) the routine trying to assign ORCIDs to members of the experimental team, and (iv) the NeXus writer routine that creates a NeXus file and feeds the file to the ICAT repository (see Fig. 2). All routines are written in Python, to harmonize with SISSY I's Bluesky process and, also to be transparent to the instrument scientist who is encouraged to maintain the code of the instrument-specific part. A training course is provided to ensure the scientist is supported in this.

Assuming that detailed sample information is available SISSY I's (meta)data are considered to be rich enough to start the data conversion process automatically when a measurement ceases (see Fig. 3). The data conversion server works in the background, only accessing SISSY I's Mongo database once, in the beginning, to get the (meta)data of the run. The subsequent enrichment of (meta)data, the creation of the NeXus file and the ICAT ingest process doesn't affect the SISSY I instrument.

For samples not originating from the EMIL infrastructure there are two manual processes implemented on the data conversion server, shown in Fig. 3, which can be initiated by the instrument scientist of SISSY I. First, the single-step process provides a graphical user interface (GUI) through which essential metadata such as the proposal number, sample information and contacts to the experimental team can be provided. Then the experiment scientist starts the NeXus creation process which proceeds in the same way as the automatic workflow but adds the provided GUI metadata.

Second, for maximum control of ingested data a manual two-step process is offered. Here, the NeXus file is created in the beginning and the GUI displays the selected content that which can be adjusted to the experiment scientist's needs. Changes are written to the NeXus file which can be inspected by the favored software until the content meets the requirements.

In general, instrument scientists and data curators have to balance between automatic (meta)data processing offering low effort but bear the risk of wrong or missing (meta)data and a manual curation accompanied by increased workload but potentially more accurate information. Manual workflows would be the method of choice for example, when the instrument scientist must differentiate between data that can or cannot be made generally accessible, which is a concern when the beamline is shared between scientific users and users from industry whose results are confidential.

Currently, a coherent interface between various parts such as the e-logbook, repository, and the proposal system is missing and represent hurdles to the workflow, as additional user authentications are required. We are building bridges between services but an IT environment avoiding such problems would be desirable.

RESULTS AND DISCUSSION: FAIR ASSESSMENT

At SISSY I, we built a pure Python environment for beamline control, (meta)data acquisition, metadata enrichment, and conversion. This was made possible by the Python package Bluesky. We improve all fields of FAIR principles to increase data quality for long-term storage. To assess the FAIR status of the SISSY I instrument we apply the FAIR indicators of the FAIR maturity model [34] where results are summarized in table 1 following the presentation in [35]. Most of the time we restrict the discussion to indicators which are classified as 'essential' and therefore focus on the most important indicators to be fulfilled.

Reusability represents mainly the richness of metadata and, thus, we take measures on the instrument level to collect additional information. Since we were able to include sample information, electronic notebook and metadata identifiers such as ORCID or the instrument PID, we consider (meta)data to be sufficiently enriched that it can be reused. Moreover, it is presented in the machine-readable community standard NeXus. Currently, the method of including

ISBN: 978-3-95450-221-9 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2021-WEBL05

FAIR	ID	Indicator	SISSY I
		Findability	
F1	RDA-F1-01M	Metadata is identified by a PID	3
F1	RDA-F1-01D	Data is identified by a PID	3
F1	RDA-F1-02M	Metadata is identified by a globally unique ID	3
F1	RDA-F1-02D	Data is identified by a globally unique ID	3
F2	RDA-F2-01M	Rich metadata is provided to allow discovery	4
F3	RDA-F3-01M	Metadata includes the identifier for the data	4
F4	RDA-F4-01M	Metadata can be harvested and indexed	3
		Accessibility	
A1	RDA-A1-01M	Metadata contains information to get access to the data	3
A1	RDA-A1-02M	Metadata can be accessed manually	4
A1	RDA-A1-02D	Data can be accessed manually	4
A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	3
A1	RDA-A1-03D	Data identifier resolves to a digital object	3
A1	RDA-A1-04M	Metadata is accessed through standardized protocol	4
A1	RDA-A1-04D	Data is accessible through standardized protocol	4
A1	RDA-A1-05D	Data can be accessed automatically	3
A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	4
A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	4
A1.2	RDA-A1.2-01D	Access protocol supports authentication and authorization	4
A2	RDA-A2-01M	Metadata remains available after data is no longer available	4
		Interoperability	
I1	RDA-I1-01M	Metadata uses knowledge representation in standardized format	4
I1	RDA-I1-01D	Data uses knowledge representation in standardized format	4
I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	4
I1	RDA-I1-02D	Data uses machine-understandable knowledge representation	4
I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	3
I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	3
I3	RDA-I3-01M	Metadata includes references to other metadata	4
I3	RDA-I3-01D	Data includes references to other data	4
I3	RDA-I3-02M	Metadata includes references to other data	2
I3	RDA-I3-02D	Data includes qualified references to other data	4
I3	RDA-I3-03M	Metadata includes qualified references to other metadata	4
I3	RDA-I3-04M	Metadata include qualified references to other data	2
		Reusability	
R1	RDA-R1-01M	Plurality of accurate and relevant attributes are provided	4
R1.1	RDA-R1.1-01M	Metadata includes information about the reuse license	3
R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse license	3
R1.1	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse license	3
R1.2	RDA-R1.2-01M	Metadata includes provenance information in community specific standard	4
R1.2	RDA-R1.2-02M	Metadata includes provenance information in cross community language	2
R1.3	RDA-R1.3-01M	Metadata complies with a community standard	4
R1.3	RDA-R1.3-01D	Data complies with a community standard	4
R1.3	RDA-R1.3-02M	Metadata is in machine-understandable community standard	4
R1.3	RDA-R1.3-02D	Data is in machine-understandable community standard	4

0 – not applicable, 1 – not being considered yet, 2 – under consideration, 3 – in implementation phase, 4 – fully implemented
Essential FAIR indicators are indicated by gray background.

Table 1: FAIR maturity model indicators for SISSY I in its current state of the commissioning phase.

the reuse license, in this case the HZB data policy, is under discussion.

Interoperability is closely connected to the way the knowledge is presented. Rather than the technical file format here the NeXus Definition Language NXDL is considered. Although no indicator is essential, NXDL is assumed to fulfill most requirements. However, to our understanding the NeXus vocabulary is not fully FAIR-compliant since the terms are not resolvable to let machines understand their meaning, but this work is up to the NIAC [36] which maintains the NeXus format. According to the assessment summarized in table 1 major deficiencies due to missing connection to other data, e.g., calibration measurements, must be addressed later in the commissioning phase.

By moving (meta)data to HZB's central repository ICAT we address the issue of accessibility. ICAT allows members of the experimental team to access (meta)data through a web interface. Missing the required identifier, (meta)data records

ISBN: 978-3-95450-221-9 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2021-WEBL05

are not yet resolvable and, thus, access through machines is limited.

Like the accessibility, the findability depends mainly on features of the repository. To be globally visible (meta)data must be harvested by higher level services such as B2FIND and the repository itself must be known outside the HZB. A prerequisite to interact with global services is the use of globally unique PIDs that are automatically assigned to (meta)data and which will soon be implemented.

As a result, we expect to realize the essential FAIR indicators at the SISSY I end-station in the near future. However, please note that the assessment seems to be a rather subjective process since some FAIR indicators are ambiguous and need further clarification (e.g., the 'richness' of metadata).

CONCLUSION

We built a Python environment for automatic data workflow ranging from the sample handling to the instrument control software over the NeXus writing procedure to the repository ingest. Most of the required infrastructure is centralized to have a minimum impact on the instrument's work and share resources as e.g., the major part of writing NeXus files is independent of the scientific technique [15].

Important measures were implemented on the instrument level to increase the reusability significantly such as the automatic collection of sample information, and these have an impact on the workflow at the instrument. The early implementation of FAIR principles during the commissioning phase results in mutually induced workflow changes of instrument control and data management without running risk of increased downtime.

ACKNOWLEDGMENTS

The authors would like to thank Heike Görzig for constructive advice on software developments and Roland Mueller for helpful discussions.

This publication was supported by the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative.

GG, WS, OM conceptualized the project. GG and WS designed the workflow methodology and wrote software. WS and SV integrated Bluesky and metadata collection. GG and MK performed the investigation leading the FAIR assessment. MB and RW provided resources in the form of the EMIL laboratory and SISSY I end station. RK and NG provided software in the form of the e-logbook. RK worked on the instrument PID and provided the ICAT. GG wrote the original draft. All authors contributed to the review and editing process.

REFERENCES

[1] Mark D. Wilkinson, Michel Dumontier, and IJsbrand Jan Aalbersberg *et al*. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. https://doi.org/10.1038/sdata.2016.18.

[2] Klaus Lips, David Starr, and Marcus Bär *et al*. Emil: The energy materials in situ laboratory Berlin. *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)*, 11:0698–0700, 2014. https://doi.org/10.1109/PVSC.2014.6925017.

[3] FAIR Data Maturity Model Working Group. FAIR Data Maturity Model. Specification and Guidelines. *Zenodo*, 9:2651–2659, 2020. https://doi.org/10.15497/rda00050.

[4] Bluesky Project. https://blueskyproject.io/. Accessed: 2021-09-10.

[5] Markus Stocker, Louise Darroch, Rolf Krahl, Ted Habermann, Anusuriya Devaraju, Ulrich Schwardmann, Claudio D'Onofrio, and Ingemar Häggström. Persistent identification of instruments. *Data Science Journal*, 19(1):18, 2020. http://doi.org/10.5334/dsj-2020-018.

[6] Klaus Lips, Tim Schulze, and David Starr *et al*. EMIL: the energy materials in-situ laboratory Berlin – a novel characterization facility for photovoltaic and energy materials. *31st European Photovoltaic Solar Energy Conference and Exhibition*, 9:0698–0700, 2015. https://doi.org/10.4229/EUPVSEC20152015-1AO.2.1.

[7] Stefan Hendel, Franz Schäfers, Michael Hävecker, Gerd Reichardt, Michael Scheer, Johannes Bahrdt, and Klaus Lips. The EMIL project at BESSY II: Beamline design and performance. *AIP Conference Proceedings*, 1741:030038, 2016. https://doi.org/10.1063/1.4952861.

[8] Prevac. https://www.prevac.eu. Accessed: 2021-10-05.

[9] Rolf Follath, M. Hävecker, G. Reichardt, Klaus Lips, Johannes Bahrdt, Franz Schäfers, and P. Schmid. The energy materials in-situ laboratory Berlin (EMIL) at BESSY II. *Journal of Physics: Conference Series*, 425:212003, 2013. https://org.doi/10.1088/1742-6596/425/21/212003.

[10] EPICS. https://epics-controls.org/. Accessed: 2021-09-11.

[11] Ophyd. https://nsls-ii.github.io/ophyd/. Accessed: 2021-09-11.

[12] MongoDB. https://www.mongodb.com/en-us. Accessed: 2021-09-11.

[13] Will Smith *et al*. Status of Bluesky deployment at BESSY II. this conference.

[14] S. R. Heller, A. McNaught, and I. Pletnev *et al*. InChI, the IUPAC international chemical identifier. *J Cheminform*, 7:23, 2015. https://doi.org/10.1186/s13321-015-0068-4.

[15] R. Dimper, A. Götz, A. de Maria, V. A. Solé, M. Chaillet, and B. Lebayle. ESRF data policy, storage, and services. *Synchrotron Radiation News*, 32(3):7–12, 2019. https://doi.org/10.1080/08940886.2019.1608119.

[16] Joao Felipe Pimentel, Leonardo Murta, Vanessa Braganholo, and Juliana Freire. Understanding and improving the quality and reproducibility of Jupyter notebooks. *Empirical Software Engineering*, 26(4), 2021. https://doi.org/10.1007/s10664-021-09961-9.

[17] Declan Butler. Scientists: your number is up. *Nature*, 485:564, 2012. https://doi.org/10.1038/485564a.

[18] SISSY I endstation. https://doi.org/10.5442/NI000020.

ISBN: 978-3-95450-221-9 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2021-WEBL05

[19] Soft x-ray EMIL beamline. https://doi.org/10.5442/NI000018.

[20] Hard x-ray EMIL beamline. https://doi.org/10.5442/NI000019.

[21] OASE – operando absorption and emission spectroscopy at EMIL. https://doi.org/10.5442/NI000021.

[22] SISSY-off SR. https://doi.org/10.5442/NI000022.

[23] Mark Könnecke, Frederick A. Akeroyd, Herbert J. Bernstein, Aaron S. Brewster, and Stuart I. Campbell *et al.* The nexus data format. *Journal of Applied Crystallography*, 48(1):301–305, 2015. https://doi.org/10.1107/S1600576714027575.

[24] Francesco De Carlo, Doğa Gürsoy, Federica Marone, Mark Rivers, and Dilworth Y. Parkinson. Scientific data exchange: a schema for HDF5-based storage of raw and analyzed data. *Journal of Synchrotron Radiation*, 21(6):1224–1230, 2014. https://doi.org/10.1107/S160057751401604X.

[25] U. K. Pedersen, N. Rees, M. Basham, and F. J. K. Ferner. Handling high data rate detectors at diamond light source. *Journal of Physics: Conference Series*, 425:062008, 2013. https://doi.org/10.1088/1742-6596/425/6/062008.

[26] N. P. Rees, M. Basham, F. J. K. Ferner, U. K. Pedersen, T. S. Richter, and J. A. Thompson. High speed detectors: problems and solutions. *Proceedings of ICALEPCS2013*, pages 1016–1019, 2013. paper WECOBA07, ISBN 978-3-95450-139-7.

[27] Mark Basham, Jacob Filik, Michael T. Wharmby, Peter C. Y. Chang, and Baha El Kassaby *et al.* Data Analysis WorkbeNch (DAWN). *Journal of Synchrotron Radiation*, 22(3):853–858, 2015. https://doi.org/10.1107/S1600577515002283.

[28] HZB Data Repository. https://data.helmholtz-berlin.de/. Accessed: 2021-09-11.

[29] Stephen M. Fisher, Frazer Barnsley, Wayne Chung, Sylvie da Graça Ramos, and Alex de Maria *et all*. The growth of the ICAT family. *NOBUGS 2016 Proceedings*, 9:2651–2659, 2008. https://doi.org/10.17199/NOBUGS2016.45.

[30] ICAT. https://icatproject.org/. Accessed: 2021-09-11.

[31] HZB Data Policy. https://hz-b.de/datapolicy. Accessed: 2021-09-11.

[32] B2FIND. http://b2find.eudat.eu/. Accessed: 2021-09-11.

[33] R. Müller, A. Balzer, P. Baumgärtel, O.-P. Sauer, and G. Hartmann *et al.* Modernization of experimental data taking at BESSY II. *Proceedings of ICALEPCS2019*, pages 65–69, 2019. https://doi.org/10.18429/JACoW-ICALEPCS2019-MOCPL02.

[34] FAIR Data Maturity Model. Specification and Guidelines. https://doi.org/10.15497/rda00050. Accessed: 2021-09-11.

[35] Christophe Bahim, Carlos Casorrán-Amilburu, Makx Dekkers, Edit Herczog, and Nicolas Loozen *et all*. The FAIR data maturity model: An approach to harmonise FAIR assessments. *Data Science Journal*, 19(41):1–7, 2020. https://doi.org/10.5334/dsj-2020-041.

[36] NeXus International Advisory Committee. https://www.nexusformat.org/NIAC.html. Accessed: 2021-09-11.