FAIR DATA OF PHYSICAL AND DIGITAL BEAMLINES

Gerrit Günther*, Simone Vadilonga, Oonagh Mannix, Ovsyannikov Ruslan
Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB),
Lise-Meitner-Campus, Hahn-Meitner-Platz 1, 14109 Berlin, Germany

Abstract

Simulations play a crucial role in instrument design, as a digital precursor of a real-world object they contain a comprehensive description of the setup. Unfortunately, this digital representation is often neglected once the real instrument is fully commissioned. To preserve the symbiosis of simulated and real-world instrument beyond commissioning we connect the two worlds through the instrument control software. The instrument control simultaneously starts measurements and simulations, receives feedback from both, and directs (meta)data to a NeXus file – a standard format in photon and neutron science. The instrument section of the produced NeXus file is enriched with detailed simulation parameters where the current state of the instrument is reflected by including real motor positions such as incorporating the actual aperture of a slit system. As a result, the enriched instrument description increases the reusability of experimental data in sense of the FAIR principles. The data is ready to be exploited by machine learning techniques, such as for predictive maintenance applications as it is possible to perform simulations of a measurement directly from the NeXus file. The realization at the Aquarius beamline at BESSY II in connection with the Ray-UI simulation software and RayPyNG API[1] serves as a prototype for a more general application.

INTRODUCTION

Synchrotron beamlines are complex instruments which often comprise customized parts to enable scientific investigations that were not feasible before. The life cycle of such unique instruments starts with a simulation in which components and their arrangement to each other are optimized to achieve maximum performance. During the commissioning phase the physical beamline is compared with its digital precursor to make sure that the newly build instrument is constructed according to the simulation and meets the expected performance. Unfortunately, once the beamline is in operation, usually the simulation is neglected and the physical beamline evolves independently of its digital counterpart.

As the central authority over the physical beamline, the instrument control system orchestrates data taking and storage and, thus, is a suitable element to keep a connection to the digital beamline beyond commissioning. By combining the information of physical and digital beamline, the FAIRness in sense of the FAIR principles (findability, accessibility, interoperability, reusability) [1] of experimental and simulation data is mutually improved since they complement each other. As found before [2, 3], the interoperability and reusability aspects of FAIR are inherent in the data and, in this case, concern experimental and simulation data as well as their relation. In this work, we report on our efforts to store beamline data in a meaningful way in that sense that the relation between experimental and simulation data is visible and distinguishable to human and machine agents. The approach is part of a wider framework which explores the next-generation experiment control and (meta)data software at HZB [3–8].

Throughout the paper, the convention of [9] is adopted to distinguish between data and metadata. Here, data refer to the primary output of physical and simulated detectors or other objects of outstanding scientific interest while metadata belong to information that helps to analyze the primary data such as the description of beamline components. If both types of data are addressed the term (meta)data is used.

PHYSICAL BEAMLINE

The Aquarius beamline at BESSY II is currently in commissioning and, thus, represents an ideal testbed to explore (meta)data workflows between the physical instrument and its digital counterpart. Aquarius employs a next-generation experiment control system based on BlueSky which is currently developed and tested at different beamlines [5]. When starting a measurement, the experiment control system collects (meta)data of various devices and stores them in a locally accessible Mongo database which provides advanced machine-readability and search options (Fig. 1). For long-term storage and (meta)data publication, the content of the Mongo database is converted to NeXus files which are a standard data format in photon and neutron science [10]. NeXus is physically a HDF5 file format [11] whose content is arranged according to the NeXus Definition Language (NXDL), a semantic framework of predefined structure and naming convention. NeXus files are particular suitable for storing experimental (meta)data with a dedicated instrument section to define instrument details and, thus, clearly state where (meta)data originate from.

DIGITAL REPRESENTATION

Beamline simulations are sophisticated programs to compute beam properties along the path of photons through the instrument. They rely on appropriate methods catching the main physics and a digital representation of the beamline using characteristic numbers to describe the instrument. For example, there could be a certain method to deselect parts of the beam when passing a slit and numbers which detail the width, position and material properties of the aperture, deter-

* gerrit.guenther@helmholtz-berlin.de

[1] raypyng 0.0.20, `https://pypi.org/project/raypyng/`

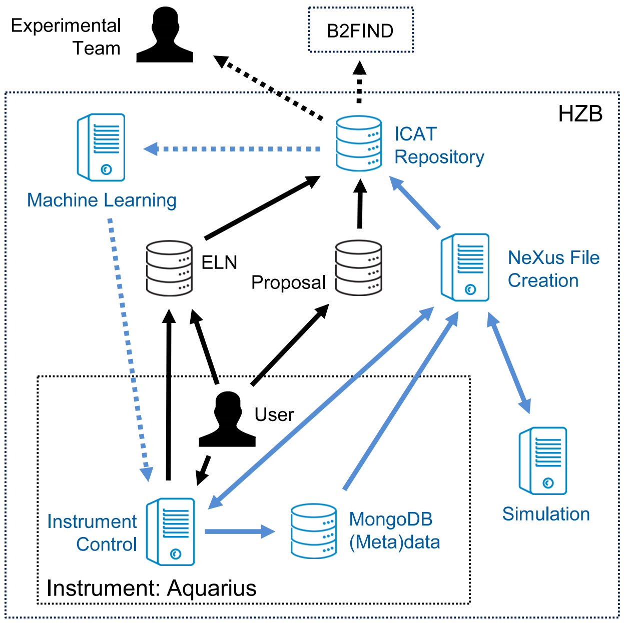

Figure 1: Schematic (meta)data streams establishing connections between the Aquarius beamline, HZB infrastructure, and the outside world (e. g. higher-level services such as B2Find [12]); the part used in this work is marked in blue.

mining absorption and reflection of the beam. A part of the parameters is usually measured by the instrument, such as a motor encoder giving the aperture width, and is available through the instrument control software. Other parameters, e. g. the aperture's material properties or position, must be derived from an external source since they are assumed to remain fixed during operation of the beamline and are not tracked by the instrument control system. These fixed parameters may be found in various documents which are created during the construction of the physical beamline, e. g. in a specification sheet, technical document, engineering software, or test report. However, the digital representation of a simulation collects this information in formal structure and semantics. Moreover, the simulation input restricts to performance-relevant parameters keeping the instrument description to an absolute minimum while being complete in the sense that the instrument performance is reflected. This renders the digital representation of a beamline suitable and efficient for instrument description.

The Ray-UI simulation software [13] employs an ASCII configuration file in the XML data format to describe the beamline. The file contains a comprehensive list of beamline components, their performance-relevant parameters and geometric information such as the component's position and facing. The order of the list matches the sequence in which the beam passes the components and reflects Ray-UI's modular structure which computes beam properties componentwise as the beam propagates through the beamline. Ray-UI's arrangement according to components matches the instrument description of NXDL and Ray-UI components can be translated straightforward to the base classes of NXDL. For example, a Ray-UI component of type *Slit* corresponds to the NXDL base class *NX_slit*. The same applies to the parameters of the Ray-UI components which are mapped to NXDL data fields. Some parameters have an exact counterpart in NXDL, such as the Ray-UI *totalWidth* could become *x_gap* in NXDL, and, thus, can be considered to be machine-readable. This also applies to the position and facing of components which require a transformation from the Ray-UI coordinate system to that of NXDL. Unfortunately, a large part of Ray-UI parameters are missing in NXDL and the corresponding NeXus data fields get rather arbitrary names mimicking NXDL convention (of using lower case characters and underscores) labeled by an attribute with the original Ray-UI name. However, this renders a large part of the instrument section ambiguous for humans and machines.

The NXDL instrument section describes the physical beamline in the state during the measurement and, thus, the change of parameters during beamline operation must be considered. To reflect the current state of the instrument, an additional mapping connects parameters of beamline components with motor values of the real world object which are stored in the NeXus file. The mapping could include a computation, e. g. to convert motor steps into units of mm. For example, a motor value connected to a slit system is used to define the value of the parameter *x_gap* in the instrument section. On a semantic level, the boundary between physical and digital beamline becomes blurred since NXDL lacks a rigid annotation to distinguish between experimental and simulation (meta)data. However, the beamline is detailed to the granularity which is required to conduct Ray-UI simulations and, thus, is fully described in the sense that a simulation of the beamline can be performed from the NeXus file.

There are ongoing efforts [5] to store the instrument description directly in the Ophyd abstraction layer of the BlueSky experiment control software to create the instrument description dynamically depending on the devices that are connected to BlueSky. This would allow to automatically track changes of the instrument setup and would tighten the connection between physical and digital beamline significantly.

DIGITAL BEAMLINE

Simulations can be useful to further increase the reusability of experimental data in sense of the FAIR principles as they can provide theoretical (meta)data of instrument components that are not accessible in the physical beamline, such as beam dimensions at the sample position. For comparison purposes, the output of physical and digital beamlines often share the same representation (e. g. data dimensions and units) which runs the risk of leaving experimental and simulation data indistinguishable. However, NXDL relies on structure and semantics as basic elements to make relations visible and reduce ambiguity.

To enrich instrument description, selected simulation results are added to the NeXus file where context is established by assigning the simulated (meta)data to the corre-

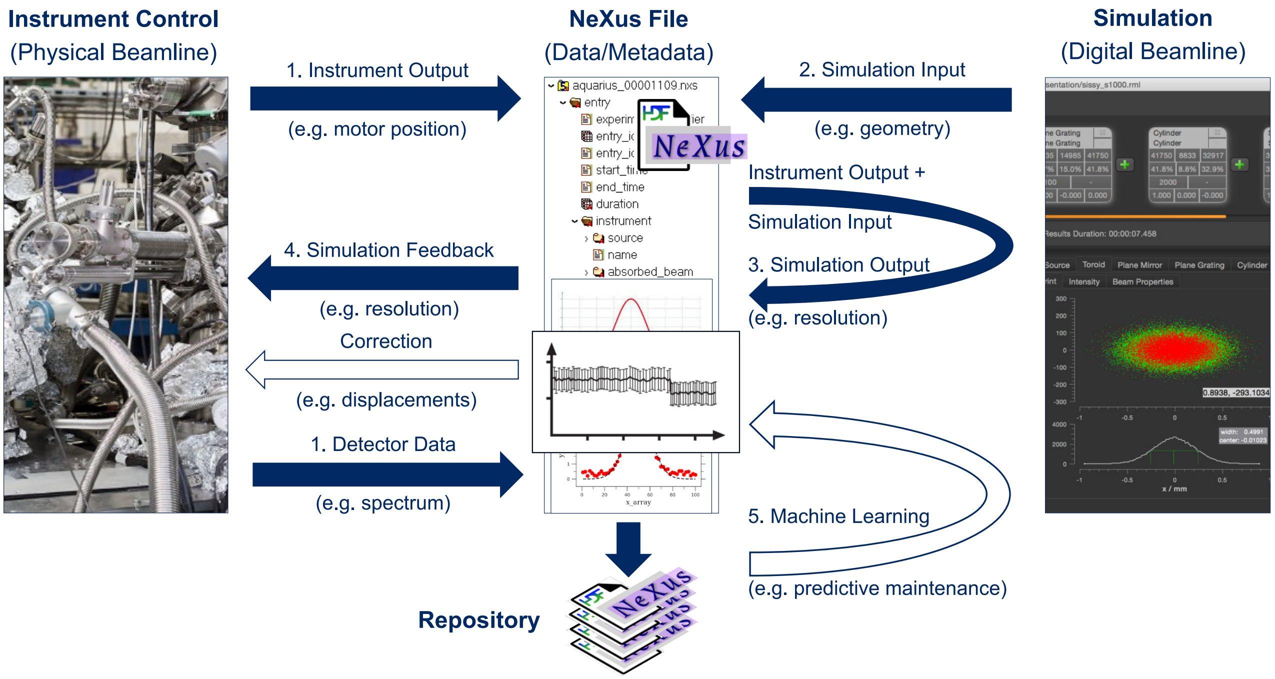

Figure 2: Schematic representation of (meta)data streams between physical beamline, NeXus file, digital beamline and repository which, finally, contains experimental and simulation data ready to be exploited by machine learning applications.

sponding beamline component. The simulation results can range from a single number, such as the FWHM of a resolution function, to more complex information such as a beam profile. Although humans are usually aware that the instrument section of a NeXus file is made from different sources, the (meta)data of physical and digital beamlines should be distinguishable for humans and machines. With a focus on experimental (meta)data, NXDL misses precise semantics for simulated (meta)data. To make their origin visible to humans, (meta)data are labeled by name (e. g. 'simulated_beam_profile') and an attribute (e. g. *generated_by* containing either *'simulation'* or *'measurement'*). However, an arbitrary name and a controlled vocabulary, which is not part of NXDL, are ambiguous for humans and rather unintelligible for machines, and, thus, a satisfying solution is still pending.

Moreover, the validity of simulation results must be considered since NXDL creates context for a measurement investigating a sample which is defined by the *NXsample* base class. Contrary, Ray-UI restricts to simulations of the beamline without sample. As a consequence, Ray-UI provides results for components up to the sample position but after that their meaning may change. For example, after passing the sample position the simulated beam could provide the instrument resolution function at a detector (if sample interaction is negligible). However, the simulation software would have to take into account the interaction between the sample and the beam in order to meaningfully complement experimental (meta)data measured after the beam has passed the sample.

(META)DATA WORKFLOW

The instrument control software of the physical beamline orchestrates various software processes to direct devices, such as motors or detectors, and, thus, is a natural choice to control the digital beamline. At the Aquarius beamline, the BlueSky instrument control software employs the ZeroMQ Python library to establish a connection to an external server which uses a Python script to manage the digital beamline (see Fig. 1). This setup is consistent with the Pythonic BlueSky software and outsources simulation as well as production of the NeXus file to have minimal impact on the measurement.

During the measurement process the NeXus file is written step-wise and provides a well-defined interface for subsequent processes making the procedure software-agnostic to a large extent (see Fig. 2). When starting a measurement, the instrument section of the NeXus file is enriched with (meta)data of the simulation and experiment where the current state of the instrument is reflected. A XML configuration file is created from the NeXus instrument section to start the Ray-UI simulation whose results are added to the NeXus file and partially returned back to the physical beamline to provide additional information to the instrument scientist. Once the measurement ceases, experimental detector data complement the NeXus file which is automatically ingested to the globally available ICAT repository [14, 15].

The accuracy of simulations is adjusted to establish a feedback to the beamline scientist during the measurement for comparison with experimental results. (Meta)data of the digital beamline can help to make decisions on the further course of the experiment and allow the beamline scientist to check the agreement of physical and digital beamline continuously. Various approaches are identified to reduce feedback time and increase simulation accuracy. The Ray-UI simulation could be outsourced to a more powerful computational infrastructure, such as a high performance computer cluster, to reduce computing time. Moreover, the Ray-UI software could be adapted to enable the storage of intermediate results along the beam which would allow to reuse simulation parts that remain fixed and, thus, save computing time. Since a beamline is usually operating in a standard configuration range, former simulation results could be reused to avoid repetition of calculations. A combination of the measures could increase simulation performance significantly and consolidate the connection between physical and digital beamline.

By combining (meta)data of the physical and digital beamline in the same NeXus file, a rigid connection between both worlds is established in the sense that (meta)data (i) are combined to increase reusability, (ii) create context by arranging them close to each other (e. g. in the same component group), (iii) use consistent semantics of NXDL identifying different terms that share the same meaning, and (iv) are stored in the same physical format allowing the use of the same software to read experimental and simulation (meta)data. This makes the (meta)data usable for machine learning algorithms. Since a NeXus file belongs to the beamline in a certain state at a certain time, a machine learning algorithm could observe the performance of the physical beamline over time by comparing experimental and simulation (meta)data of different files. This would allow, for example, to conduct predictive maintenance to identify the misalignment, decay or ongoing breaking of beamline components. However, the available (meta)data of the physical beamline are a limiting factor since the detector signal placed behind the sample position is usually governed by features of a sample, burying weak instrument traces, while detectors located before the sample position are missing. To increase the amount of comparable data to train a machine learning algorithm, additional hardware in form of detectors and sensors along the beamline, dedicated measurements without (or with a standard) sample, or a combination of both would be required.

CONCLUSION

By establishing (meta)data workflows between physical and digital beamlines, additional information is provided during the experiment and for later reuse. The close arrangement of physical and simulation data within the same NeXus file creates context between the two worlds and allows either simulation data to be considered as a complement to experiments, or to use experimental (meta)data to assess simulation quality. To further strengthen the connection between both worlds, physical and digital beamline can be optimized (i) on the simulation software level by reusing ’partial’ simulations of the beamline or include sample interaction, (ii) on the organizational level by reusing former simulations of the same state, (iii) on the infrastructure level such as integrating high performance computing, (iv) on the semantic level by extending NXDL with regard to simulations, and (v) on the physical level by adding auxiliary sensors to the real world object for comparison with the digital beamline. This could help to increase the FAIRness of (meta)data produced by physical and digital beamlines and optimize their usage for machine learning techniques.

ACKNOWLEDGMENTS

This project 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.

REFERENCES

[1] M. D. Wilkinson, M. Dumontier, and *et al*, “The FAIR guiding principles for scientific data management and stewardship”, *Sci. Data*, vol. 3, p. 160 018, 2016. doi:10.1038/sdata.2016.18

[2] G. Günther *et al.*, “IR of FAIR - principles at the instrument level”, presented at ICALEPCS 2023, Cape Town, South Africa, 2023, paper WE3BCO09, this conference.

[3] G. Günther *et al.*, “FAIR Meets EMIL: Principles in Practice”, no. 18, pp. 574–580, 2022. doi:10.18429/JACoW-ICALEPCS2021-WEBL05

[4] W. Smith *et al.*, “Experimental data taking and management at BESSY II and HZB: The upgrade process”, presented at ICALEPCS 2023, Cape Town, South Africa, 2023, paper MO2AO04, this conference.

[5] S. Vadilonga, G. Günther, S. Kazarski, R. Ovsyannikov, S. Sachse, and W. Smith, “Advancements in beamline digital twin at BESSY II”, presented at ICALEPCS 2023, Cape Town, South Africa, 2023, paper THMBCMO18, this conference.

[6] H. He, G. Preuß, S. Sachse, W. Smith, and R. Ovsyannikov, “Bluesky web client at Bessy II”, presented at ICALEPCS 2023, Cape Town, South Africa, 2023, paper TUPDP014, this conference.

[7] P. Wegmann *et al.*, “SECoP integration for the Ophyd hardware abstraction layer”, presented at ICALEPCS 2023, Cape Town, South Africa, 2023, paper THMBCMO10, this conference.

[8] W. Smith, M. Arce, M. Bär, M. Gorgoi, C. E. Jimenez, and I. Rudolph, “Using ArUco codes for beam spot analysis with a camera at an unknown position”, presented at ICALEPCS 2023, Cape Town, South Africa, 2023, paper THMBCMO30, this conference.

[9] FAIR Data Maturity Model: Specification and Guidelines, doi:10.15497/rda00050

[10] M. Könnecke, F. A. Akeroyd, H. J. Bernstein, A. S. Brewster, and *et al*, “The nexus data format”, *J. Appl. Crystalogr.* vol. 48, no. 1, pp. 301–305, 2015. doi:S1600576714027575

[11] F. D. Carlo, D. Gürsoy, F. Marone, M. Rivers, and D. Y. Parkinson, "Scientific data exchange: A schema for HDF5-based storage of raw and analyzed data", *J. Synchrotron Radiat.*, vol. 21, no. 6, pp. 1224–1230, 2014. doi:S160057751401604X

[12] M. Demleitner and C. Martens, "B2FIND – searching for research data across disciplines", in *E-Science-Tage 2021: Share Your Research Data*. heiBOOKS, 2022, pp. 196–207. doi:10.11588/heibooks.979.c13729

[13] P. Baumgärtel *et al.*, "RAY-UI: New features and extensions", *AIP Conf. Proc.*, vol. 2054, no. 1, p. 060 034, 2019. doi:10.1063/1.5084665

[14] S. M. Fisher, F. Barnsley, W. Chung, S. da Graça Ramos, and A. de Maria *et all*, "The growth of the ICAT family", in *Proc. NOBUGS 2016*, vol. 9, 2008, pp. 17–22. doi:10.17199/NOBUGS2016.45

[15] ICAT, https://icatproject.org/