ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE3BC009

IR OF FAIR: PRINCIPLES AT THE INSTRUMENT LEVEL

Gerrit Günther[1]*, Sebastian Baunack[2], Luigi Capozza[2], Oliver Freyermuth[3], Pau Gonzalez-Caminal[4], Boxing Gou[5], Johann Isaak[6], Sven Karstensen[7], Axel Lindner[7], Frank Maas[2], Oonagh Mannix[1], Andrew Mistry[8], Isabella Oceano[7], Christiane Schneide[7], Thomas Schörner-Sadenius[7], Kilian Schwarz[7], Vivien Serve[1], Lisa-Marie Stein[7], Stefan Typel[6], Malte C. Wilfert[2].

[1] Helmholtz-Zentrum Berlin für Materialien und Energie GmbH (HZB), Hahn-Meitner-Platz 1, 14109 Berlin, Germany

[2] Helmholtz-Institut Mainz, Johannes Gutenberg-Universität Mainz, Staudingerweg 18, D-55099 Mainz, Germany

[3] Rheinische Friedrich-Wilhelms-Universität Bonn, Nußallee 12, 53115 Bonn, Germany

[4] Fusion for Energy (ATG Science & Technology, S.L.), c/ Josep Pla 2, Blg. B3, 08019 Barcelona, Spain

[5] Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou 730000, China

[6] Technische Universität Darmstadt, Institut für Kernphysik, Schlossgartenstraße 9, 64289 Darmstadt, Germany

[7] Deutsches Elektronen-Synchrotron DESY, Notkestraße 85, 22607 Hamburg, Germany

[8] GSI Helmholtzzentrum für Schwerionenforschung GmbH, Planckstraße 1, 64291 Darmstadt, Germany

Abstract

Awareness of the need for FAIR data management has increased in recent years but examples of how to achieve this are often missing. Focusing on the large-scale instrument A4 at the MAMI accelerator, we transfer findings of the EMIL project at the BESSY synchrotron to improve raw data, i.e. the primary output stored on long-term basis, according to the FAIR principles. Here, the instrument control software plays a key role as the central authority to start measurements and orchestrate connected (meta)data-taking processes. In regular discussions we incorporate the experiences of a wider community and engage to optimize instrument output through various measures from conversion to machine-readable formats over metadata enrichment to additional files creating scientific context. The improvements were already applied to currently built next generation instruments and could serve as a general guideline for publishing data sets.

INTRODUCTION

In recent years the concept of FAIR data, addressing their Findability, Accessibility, Interoperability, and Reusability [1], evolved from a theoretical framework to realization where processes and infrastructure enable a data life cycle from creation over long-term storage to its re-use. Large-scale research instrumentation represents a starting point of this life cycle as they generate data from sophisticated measurement processes. As a consequence, the quality and precision of these raw data are crucial for the quality of future data that may originate from this basis. Here, raw data refers to the first version of experimental output that is stored on long-term basis and serves for the instrument scientists as original data source for subsequent analysis.

In previous work we have found that responsibility for the findability and accessibility aspects of FAIR lies with repositories and higher level services [2, 3], while the interoperability and re-usability according to the FAIR principles is inherent in the raw data produced by the instrument. To implement a FAIR data life cycle, the instrument's raw data must be improved as a starting point for subsequent automatic processing. Unfortunately, there is only a vague understanding about the FAIRness of raw data and concrete examples are rare. However, there are frameworks to assess the FAIRness of data, such as the FAIR Data Maturity Model, which can serve as a guideline [4, 5].

In this paper, we report on our findings and measures to improve the FAIRness of raw data produced by two particle physics instruments at the beginning and the end of the instrument life cycle. The A4 instrument at the MAMI is already dismounted and is particularly interesting since the output has been designed before the FAIR principles were published. As the A4 data still generate results of scientific interest [6], we create a post-processing workflow to convert and enrich the existing raw data sets improving their FAIRness. As a second use case, the Any-Light-Particle-Search experiment ALPS II at DESY [7] represents an ideal test bed to implement workflows applied at the instrument level since it is currently in the commissioning phase. We concentrate

* gerrit.guenther@helmholtz-berlin.de.

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE3BC009

on the interoperability and reusability of data from these two instruments by focusing on the physical and knowledge representation of data.

Throughout the paper, the convention of [8] is adopted to distinguish between data and metadata. Here, data refer to the primary output of detectors or other objects of outstanding scientific interest while metadata belong to information that help to analyze the primary data such as the physical quantities and their units. If both types of data are addressed the term (meta)data is used. We consider the interoperability and reusability of all metadata available at the file level including physical parameters, preliminary results, or log book entries.

PHYSICAL REPRESENTATION

When starting a measurement, the instrument control software orchestrates the interaction of various devices and processes which generate streams of (meta)data. Often these streams are stored in separate files arranged in a nested folder structure. The folder hierarchy and names create a loosely knitted meaning that helps humans to assign the files of different folders to different sources. For example, at A4 the (meta)data of a measurement are placed in a separate folder whose name contains the run number that allows to identify the chronological order of measurements. This structure enables navigation through the whole data set facilitating later data reuse. Our work concentrates on making this structure visible and navigable to those without prior knowledge of the data set.

In case of the A4 experiment, hierarchy and naming convention of the generated folder structure are of rather vague and intuitive nature but represent a helpful classification to orientate within the data set and, thus, is preserved when publishing the A4 (meta)data. Moreover, it helps to identify original and published versions of the same data sets. The raw data of the A4 instrument are in ASCII text file format, or in the community-specific ROOT standard [9]. For data publication, file formats stay unchanged although the content of the ASCII text files is rephrased in accordance to XML. To help human agents reading the ASCII and ROOT files, we add documentations in HTML to the data set which detail the file format and provide guidance on extracting data from the files. As XML can typically be handled by operating systems by default, the documentation is limited to a short Python script to parse the XML format.

The manual to read ROOT files is more complex since (i) a complete software installation is required to handle this file format and (ii) the data of the A4 instrument is stored using a self-written, customized class which is not part of the ROOT file but is required to read data from file. Although the procedure is common practice and meets the requirements of the community, it represents a serious hurdle for human and machine agents to access the data. As such, the HTML documentation, which we added for ROOT, includes a complete description how to read data from the customized class. This requires the installation of the ROOT software, furnishing of C++ files which define the class, and a Make file that compiles and implements the class in the ROOT software. Through HTML documentation we can achieve human interoperability and partial machine interoperability. The ROOT file contains metadata annotation and, thus, is to some extent self-describing which means that humans can access notes on the structure and meaning of (meta)data.

The situation of the next-generation instrument ALPS II is quite different as (meta)data of a measurement are continuously generated and stored in a local database enabling advanced options to browse and process (meta)data. In certain intervals chunks of the data are read from the local database and converted to a HDF5 file [10] before data becomes globally available on a long-term basis through a repository. The output of the ALPS II instrument is not less complex but transfers the complexity into the file structure where HDF5 groups mimic the role of folders or files helping humans to orientate. Compared to the A4 instrument output, a HDF5 file has the advantage that (meta)data are stored within a single file and, thus, are more portable in the sense that you can't loose parts by accident. Further, there are various open software programs to open and process HDF5 files which supports the interoperability of the de facto standard in neutron and photon science [11–13].

KNOWLEDGE REPRESENTATION

Beside the physical file format, the interoperability part of FAIR covers the representation of the file's content that a machine must be able to process and interpret unambiguously. Here, structure and semantics are key elements to represent (meta)data. Structure creates context between information by spatial grouping and hierarchical arrangement of (meta)data, while semantics concerns the language used to express the meaning of the content, ranging from human- and machine-understandable wording to a precise relation of information. Both are particularly human- and machine-readable when following a standard.

Regarding the A4 data set, several instrument devices and automatic data processes produce tables in arbitrary formats to log experimental parameters or create preliminary results for a measurement. The tables encoded as plain text are converted according to the XML data format to improve machine-readability by separating and organizing the values while missing metadata are added. By convention, tables imply hierarchy and logic to some degree. For example, column values of a table are expected to share the same units while the first column contains an independent value that the latter column values of the same row depend on. To add deeper semantics for human and machine agents to metadata (such as column titles, quantities, and units), we employ resource description framework (RDF) annotations (see Fig. 1) [15] in agreement with the RDF Data Cube Vocabulary [16] which is a suitable basis for complex, domain-specific frameworks such as the FAIR core semantic model described by Trojahn *et al.* [17]. The RDF grants flexibility in creating logic without loss in precision by combining well-defined terms and

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE3BC009

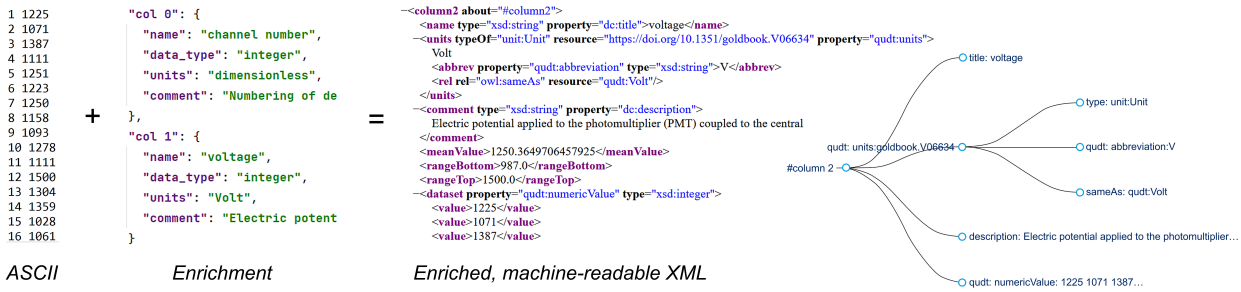

Figure 1: Schematic post-processing workflow to convert and enrich ASCII text file to XML data format. The inset on the left shows a representation of the RDFa table annotation structure seen by a machine [14], containing e.g. column title, units, and comments.

relations to specific ontologies such as QUDT [18] or the BIMERR Weather Ontology [19], as suggested before [20]. The annotation creates semantic interoperability with web standards, and improves reusability by making the data set directly machine-readable.

In case of ALPS II, HDF5 allows for declaration of various data structures within the file, ranging from integers over high-dimensional arrays to ASCII text, that renders the (meta)data machine-readable. The detector data is stored in two streams (with separate calibration and monitor channels) [21] similar to the data produced by LIGO [22] to make contact with related experiments of the community. By adopting the NeXus Definition Language (NXDL) [23], (meta)data such as table dimensions, values, and units are annotated in a standardized, machine-readable representation. The NXDL provides a detailed schema to arrange (meta)data in a HDF5 file relying on naming conventions and pre-defined meaning of terms, although it lacks the precision of ontologies whose terms are defined by an unambiguous Internationalized Resource Identifier (IRI), such as when defining units as free text (e.g. as 'micrometer' or 'mum'). The NeXus data format is a standard for large-scale instrumentation in photon and neutron science and makes (meta)data understandable to a wide audience. Some NXDL concepts such as the sample doesn't have an exact counterpart in the research field of particle physics but the general framework could be adopted to describe ALPS II raw (meta)data in accordance with photon and neutron data.

At both instruments we strive for the use of persistent identifiers (PID) to unambiguously annotate metadata such as ORCID to identify humans [24], instrument PID to define the source that generates the raw data [25], or ROR for facilities involved in the research [26]. These PID help humans and machines to identify basic elements of the project generating the data and contribute to their findability.

CREATING CONTEXT

While it is obvious where to start reading a journal article, there are no clear rules to assess an unknown data set. This poses a problem for reusability since a human or machine agent requires a minimum set of information to assign the data to a scientific field. Concerning experimental raw data,

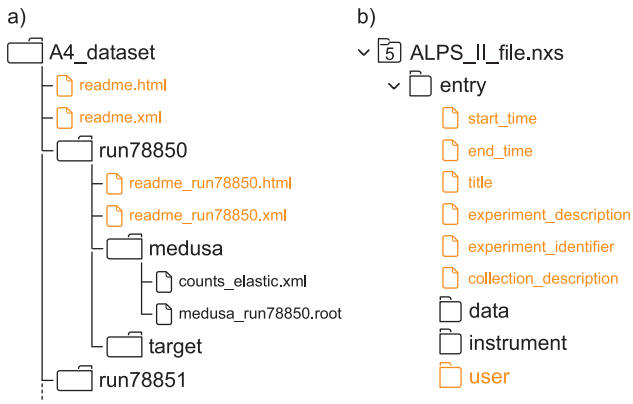

Figure 2: **a)** Schematic structure of the A4 raw data set after being improved for publication. The XML and HTML readme files at the top level and run level (orange) are added to provide context information and to span a network of links allowing human agents to browse the content. **b)** Proposed structure of the ALPS II NeXus-like HDF5 file with context related data fields and the user group, containing contacts to the experimental team, at the top-level (orange).

this set of information could comprise the name of the instrument, the time the data was taken and other metadata depending on the research field and instrument. Fortunately, this kind of metadata either (i) doesn't change during the lifetime of an instrument such as instrument name, hosting facility, experimental technique, (ii) changes rather rarely like the instrument staff, or (iii) is automatically collected such as measurement time or guest users from a proposal.

The A4 data set misses most contextual metadata and a place to put it. For this reason, an additional readme file was created, as suggested elsewhere [27–29], at the top level of the folder hierarchy in the HTML format which is human-friendly and a global standard as part of the world wide web. It contains generic information to create context, such as instrument name, contacts to the experimental team, and external references to publications about the instrument design or scientific results. Moreover, the top-level readme file contains internal links to run-level readme HTML files placed one level below in the folder hierarchy containing a single measurement which provide information about key-

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE3BC009

parameters and links to the files of the measurements. By using the links, a human agent is able to browse through the content of the data set from the top-level readme file to the XML files containing the (meta)data of a measurement (see Fig. 2a).

To make this information intelligible also to machine agents, the top-level as well as the run-level HTML readme files are accompanied by XML versions which provide the same metadata according to the DataCite standard [30]. With this approach we create a machine-readable file network in parallel interlacing the data set.

The experimental team of the A4 instrument maintained an electronic lab notebook which contains valuable information submitted automatically by the instrument control software or manually entered by members of the experimental team. Former entries are useful to check the status of instrument devices but the latter are especially interesting as they provide useful insights into decisions affecting the course of the experiment, observations, or conclusions which are of more semantic nature and may provide context to the (meta)data. As a result, the electronic lab notebook was imbedded to the human- as well as to the machine-readable network in HTML and XML file format.

In case of the ALPS II instrument the NXDL provides predefined fields with this kind of rather high-level information at the top level of the file structure (see Fig. 2b). When publishing (meta)data, the information is added automatically during the creation of files from the database.

For both instruments, we create the context required to make the data reusable. This same context also renders the data findable.

IMPLEMENTATION

For the decommissioned A4 instrument, raw experimental data is published in a semi-automatic process in which a set of measurements is arranged to supplement a scientific result or journal publication. This creates an additional high-level context as a result of the retrospect refering to an analysis of the raw instrument (meta)data. To create publishable (meta)data, a Python script collects, converts and annotates the raw measurement data. The range of measurements, the handling of files and metadata annotation is controlled through configuration files in the JSON format while the creation of readme files and links between the files are automatically created. A report produced by the Python script is added to the published data set as provenance information allowing to track the changes made during conversion.

The ALPS II instrument will generate publishable data sets automatically at regular intervals, e.g. on a daily basis. A Python script employs DESY's DOOCS library to access the database-like storage system and converts the (meta)data to the HDF5 file format by using the h5py module [21]. Details of the automatic process as well as the connection to a repository are work in progress and will be finalized when the commissioning phase is completed. The scheme devloped in the context of ALPS II shall also serve as a blueprint for data handling of future on-site particle physics experiments at DESY like BabyIAXO and MADMAX.

CONCLUSION

The work presented in this paper shows that almost all measures to increase the interoperability and re-usability according to the FAIR principles are applicable at the instrument level to create publishable raw (meta)data. Here, the instrument control software plays a key-role since it orchestrates the devices when starting a measurement and is able to call synchronous processes to add and convert (meta)data. The produced raw data are stored in file formats that are machine-readable, standardized, and open (i. e. under a public license). The (meta)data within the files are (i) rich, in that sense that the information is sufficient to analyze the data, and (ii) represented in a standardized way to help human and machine agents to interpret them unambiguously. The metadata contain a minimal set of context information allowing humans and machines to assess the given data at first glimpse which would be an important feature in a world of big data.

The identified measures of this work to improve the FAIRness of raw instrument (meta)data can be implemented at the instrument level of ALPS II and A4's next-generation instrument. This will allow an automatic processing and establish the basis for a FAIR data life cycle.

ACKNOWLEDGEMENTS

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, I. J. Aalbersberg, *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.*, "FAIR Meets EMIL: Principles in Practice", in *Proc. ICALEPCS'21*, Shanghai, China, 2022, pp. 574–580. doi:10.18429/JACoW-ICALEPCS2021-WEBL05

[3] Assessing the FAIRness of a prototypical PaN instrument at BESSY II, https://zenodo.org/record/6059994

[4] F. D. M. M. W. Group, "FAIR Data Maturity Model. Specification and Guidelines", *Zenodo*, vol. 9, pp. 2651–2659, 2020. doi:10.15497/rda00050

[5] A. Hasnain and D. Rebholz-Schuhmann, "Assessing FAIR data principles against the 5-star open data principles", in *Semant. Web: ESWC 2018 Satell. Events*, 2018, pp. 469–477.

[6] B. Gou *et al.*, "Study of two-photon exchange via the beam transverse single spin asymmetry in electron-proton elastic scattering at forward angles over a wide energy range", *Phys. Rev. Lett.*, vol. 124, no. 12, p. 122 003, 2020. doi:10.1103/PhysRevLett.124.122003

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE3BC009

[7] A. D. Spector, "Approaching the first any light particle search II science run", *SciPost Phys. Proc.*, vol. 12, p. 039, 2023. doi:10.21468/SciPostPhysProc.12.039

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

[9] R. Brun and F. Rademakers, "ROOT — an object oriented data analysis framework", *Nucl. Instrum. Methods Phys. Res. A*, vol. 389, no. 1, pp. 139–152, 1997. doi:10.1016/S0168-9002(97)00048-X

[10] 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:10.1107/S160057751401604X

[11] The HDF group. HDF Viewer, version 3.1.3, 2006, https://www.hdfgroup.org/downloads/hdfview/

[12] Manipulation and analysis toolkit for instrument data. mantid project, doi:10.5286/SOFTWARE/MANTID

[13] J. Filik *et al.*, "Processing two-dimensional X-ray diffraction and small-angle scattering data in DAWN 2", *J. Appl. Crystallogr.*, vol. 50, no. 3, pp. 959–966, 2017. doi:10.1107/S1600576717004708

[14] RDFa Play, https://rdfa.info/play/

[15] Resource description framework RDF, https://www.w3.org/RDF/

[16] The RDF Data Cube Vocabulary, https://www.w3.org/TR/vocab-data-cube/

[17] C. Trojahn, M. Kamel, A. Annane, N. Aussenac-Gilles, and ao Long Nguyen, "A FAIR core semantic metadata model for FAIR multidimensional tabular datasets", *J. Chem. Phys.*, pp. 174–181, 2022. doi:10.1007/978-3-031-17105-5_13

[18] Quantities, units, dimensions and types (QUDT), doi:10.25504/FAIRsharing.d3pqw7

[19] BIMERR weather ontology, https://bimerr.iot.linkeddata.es/def/weather/

[20] D. Jacob, R. David, S. Aubin, and Y. Gibon, "Making experimental data tables in the life sciences more FAIR: a pragmatic approach", *GigaScience*, vol. 9, no. 12, p. giaa144, 2020. doi:10.1093/gigascience/giaa144

[21] S. Karstensen *et al.*, "Creating of HDF5 files as data source for analyses using the example of ALPS II and the DOOCS control system", presented at ICALEPCS'23, Cape Town, South Africa, 2023, paper THPDP101, this conference.

[22] Open data from the third observing run of LIGO, Virgo, KAGRA and GEO, doi:10.48550/arXiv.2302.03676

[23] M. Könnecke, F. A. Akeroyd, H. J. Bernstein, A. S. Brewster, S. I. Campbell, *et al.*, "The NeXus data format", *J. Appl. Crystallogr.*, vol. 48, no. 1, pp. 301–305, 2015. doi:10.1107/S1600576714027575

[24] D. Butler, "Scientists: Your number is up", *Nature*, vol. 485, p. 564, 2012. doi:10.1038/485564a

[25] M. Stocker *et al.*, "Persistent identification of instruments", *Data Sci. J.*, vol. 19, no. 1, p. 18, 2020. doi:10.5334/dsj-2020-018

[26] Research organization registry (ROR), https://ror.org/about/

[27] Guide to writing "readme" style metadata, https://data.research.cornell.edu/data-management/sharing/readme/

[28] Research data management for Purdue, https://purr.purdue.edu/kb/metadata/readme-research

[29] Checklist for creating a readme file, https://osf.io/7hcuv

[30] Datacite metadata schema version 4.4, doi:10.14454/3w3z-sa82