PAPER • **OPEN ACCESS**

RefXAS: an open access database of X-ray absorption spectra – improvements and outlook

To cite this article: S Paripsa *et al* 2025 *J. Phys.: Conf. Ser.* **3010** 012124

RefXAS: an open access database of X-ray absorption spectra – improvements and outlook

S Paripsa[1*†], A Gaur[2†], F Förste[3†], D E Doronkin[2,4], W Malzer[3], C Schlesiger[3], B Kanngießer[3], E Welter[5], J-D Grunwaldt[2,4] and D Lützenkirchen-Hecht[1]

[1] Condensed matter - X-ray physics, University of Wuppertal, Wuppertal, Germany
[2] Institute for Chemical Technology and Polymer Chemistry, Karlsruhe Institute of Technology, Karlsruhe, Germany
[3] Technische Universität Berlin, Berlin, Germany
[4] Institute of Catalysis Research and Technology, Karlsruhe Institute of Technology, Karlsruhe, Germany
[5] Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany

[*]E-mail: paripsa@uni-wuppertal.de [†]These authors contributed equally.

Abstract. Under the DAPHNE4NFDI consortium, RefXAS has evolved as a comprehensive open-access database for X-ray absorption spectroscopy (XAS). We have implemented a platform that allows users to submit raw datasets and associated metadata via a web interface. The database supports automated metadata handling and quality control, ensuring that uploaded data adheres to predefined standards. Recent developments include the integration of a standardised download package, improved filtering systems, and the transition to institutional data storage at DESY, alongside future plans to incorporate the NeXus format, enhancing machine learning applications. This paper provides an overview of these advancements and their implications for the XAS community.

1. Introduction

The management, storage and analysis of X-ray absorption spectroscopy (XAS) data are critical challenges within the photon science community such as the increasing volume and complexity of XAS datasets, along with the lack of unified data formats and accessible reference databases. To address these gaps and enable efficient re-use, we, as part of the DAPHNE4NFDI consortium [1], have developed the RefXAS database [2] – an open-access XAS reference database. The goal is to improve the storage, organisation, and analysis of curated high-quality XAS spectra, with a strong focus on documentation of important metadata and standardisation. This platform provides pre-processing tools that allow users to visualise and compare XAS data from different facilities, enhancing data re-usability and reproducibility (FAIR data principle). To date, RefXAS has successfully integrated crucial functionalities such as automated quality control and automated metadata handling, ensuring that the submitted spectra meet pre-defined quality standards. Looking forward, there are plans to expand the database to support a wider range of sample types, broadening its usefulness across various research fields.

One of the great challenges is to automatically handle data of different formats and datatypes, including the XAS Data Interchange (XDI) format [3], which facilitates the exchange of single XAS spectra, and the Hierarchical File System version 5 (HDF5) format [4], which is suitable for multispectral X-ray experiments. More recently, efforts within the XAS community have focused on defining an application for processed XAS data based on the NeXus format [5]. Furthermore, as the XAS community advances towards the adoption of a NeXus standard, there is growing potential for applying this structured data format to improve machine learning (ML) applications in XAS research. RefXAS is well-positioned to benefit from these developments. Overall, we refer the interested reader to our recent publication, which discusses aforementioned points in detail [2].

This paper presents an overview of the current state of RefXAS, detailing improvements in data handling, and future directions, including the role of NeXus-based standardisation and its potential impact on ML integration.

2. Improvements

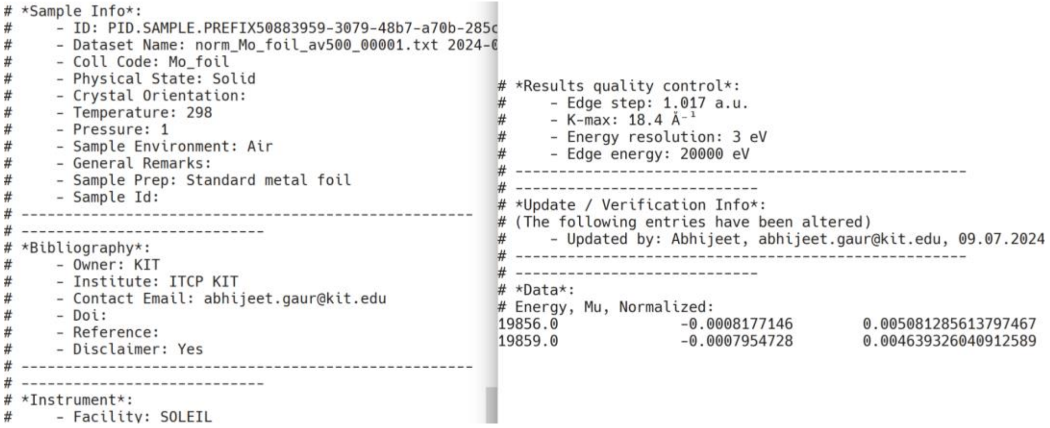
2.1 Standardised upload/download

At the interface of RefXAS, when users upload a dataset, they are required to provide detailed metadata about the experiment, sample, and instrument settings etc. This metadata is organised in a well-thought-out format and is fully visible to anyone who searches or downloads the dataset. This amount of transparency ensures that all users have access to the same metadata as originally provided by the uploading person, promoting reproducibility and enabling a deeper understanding of the data. We also want this formatted metadata to be available upon download of the data file. Hence, one of the significant advancements in our database has been the creation of a standardised, defined download package that includes an array of data and metadata, ensuring accessibility and reproducibility. Upon clicking the “Download this work (ZIP)” – button, users are provided with a ZIP [6] file that contains key elements including plots, metadata files, and human-readable data.

The human-readable metadata file provides a structured and detailed summary of each dataset, see Figure 1. It includes general information, such as the date and time when the file was created, along with the metadata pillars **Sample Info**, **Bibliography**, and **Instrument**, see Figure 2. Additionally, it records the results of the quality control, covering key metrics like the edge step, k-max, energy resolution, and edge energy. The file also includes update and verification details, documenting any changes made to the dataset by curator/database editor. Regardless of the format of the initial upload (current tested formats are .txt, .spec, .xdi, .dat, other formats are implemented gradually), the output provided is always standardised. The file is consistently structured with three essential columns—**Energy (energy of photon beam)**, **Mu (raw XAS data)**, **Normalized (normalized XAS data)**—allowing immediate compatibility with software like Athena [7]. This uniformity is an essential step toward establishing a standardised reference data framework for XAS, which makes data sharing easier and facilitates interoperability across platforms and researchers. To address potential copyright concerns and encourage data sharing, RefXAS make use of a CC BY 4.0 license, ensuring proper citation for shared datasets.

2.2 Transition to sustainable storage

A crucial development in our infrastructure is the ongoing transition from AWS-S3 [9] to either a sustainable dCache file storage solution provided by Helmholtz Federated IT Services (HIFIS) at DESY [10] or an alternative Helmholtz-managed storage solution offered by the Scientific Computing Center (SCC) of KIT [11].



```
# *Sample Info*:
#      - ID: PID.SAMPLE.PREFIX50883959-3079-48b7-a70b-285c
#      - Dataset Name: norm_Mo_foil_av500_00001.txt 2024-0
#      - Coll Code: Mo_foil
#      - Physical State: Solid
#      - Crystal Orientation:
#      - Temperature: 298
#      - Pressure: 1
#      - Sample Environment: Air
#      - General Remarks:
#      - Sample Prep: Standard metal foil
#      - Sample Id:
# -----------------------------------------------------
# ----------------------------
# *Bibliography*:
#      - Owner: KIT
#      - Institute: ITCP KIT
#      - Contact Email: abhijeet.gaur@kit.edu
#      - Doi:
#      - Reference:
#      - Disclaimer: Yes
# -----------------------------------------------------
# ----------------------------
# *Instrument*:
#      - Facility: SOLEIL
```

```
# *Results quality control*:
#      - Edge step: 1.017 a.u.
#      - K-max: 18.4 Å⁻¹
#      - Energy resolution: 3 eV
#      - Edge energy: 20000 eV
# -----------------------------------------------------
# ----------------------------
# *Update / Verification Info*:
# (The following entries have been altered)
#      - Updated by: Abhijeet, abhijeet.gaur@kit.edu, 09.07.2024
# -----------------------------------------------------
# ----------------------------
# *Data*:
# Energy, Mu, Normalized:
19856.0              -0.0008177146        0.005081285613797467
19859.0              -0.0007954728        0.004639326040912589
```

Figure 1. Format of the .txt-output file downloaded from RefXAS. The file contains important metadata under defined fields along with the quality control results, update / verification info for traceability, and a data section (right) with three standardised columns.

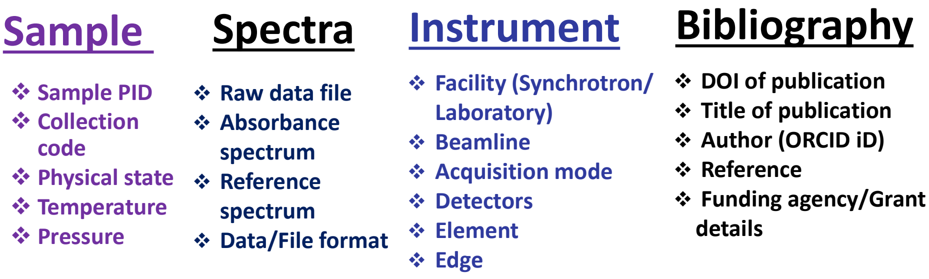


Figure 2. Categorized metadata fields and further sub-fields, formulated to provide information about the uploaded XAS spectra at RefXAS [8].

This transition enhances functionality and security for long-term data storage while ensuring institutional authentication via OpenID Connect (OIDC) [12], ensuring streamlined access control via single sign-on through the Helmholtz AAI (Authorisation and Authentication Infrastructure) [13]. This institutional-level authentication, tailored for scientific research, facilitates FAIR data sharing between institutions, reducing reliance on commercial cloud services and lowering associated costs. Integration of dCache with the RefXAS web-server API is underway, enabling automated data uploads and ensuring secure, reliable long-term access.

Future plans include establishing an institutional web address and promoting data publications through institutional repositories [14, 15].

3. Standardisation through NeXus format and Improved Quality Control

Data standardisation in XAS is an important step towards improving data usability, interoperability, and reproducibility. Our development has prioritised both the adoption of emerging data standards such as the NeXus format and the implementation of automated quality control procedures. These two aspects - data standardisation and quality control - are connected, as the integration of standardised formats directly impacts the efficiency and accuracy of automated data assessments. As the XAS community moves towards standardisation, the NeXus format slowly emerges as a robust solution for storing complex datasets in a structured, hierarchical and self-describing framework that utilises the HDF5 structure to integrate both data and metadata through standardised dictionaries in a tree-like form. One of the key advantages of the NeXus format is its emphasis on metadata inclusion. Machine learning (ML) algorithms rely heavily on well-organised and comprehensive (meta)data to make use of the data and drive accurate predictions [16]. NeXus ensures that this metadata is consistently structured and always included, unlike flat formats like ASCII-based (e.g. .dat or .txt) files, which may separate or omit crucial metadata, necessitating manual intervention and increasing the risk of errors or missing information.

Our current data output follows a human-readable format and uses a more flexible structure. Transitioning from our format to NeXus/HDF5 requires restructuring the metadata using a *# Family.Key: Value* format, that is, storing both metadata and data hierarchically, which enhances interoperability and ML integration. Efforts are already underway to automate this process, ensuring that RefXAS adopts this standardised formats. This effort will help bridge gaps between legacy data formats and modern analysis tools. Currently, the NeXus application definitions for XAS have been actively discussed within international XAS society [17] and once the proposal is finalized and approved by NeXus committee, we plan to incorporate the approved NeXus format into our database.

The international XAS community actively develops a NeXus standard specific to XAS through initiatives like pynxxas [18]. This tool provides a library for reading and writing XAS data in the NeXus format and represents an ongoing effort to unify data structures across the field. A possible integration of this tool would align with broader community efforts, thereby contributing to the overall standardisation of XAS data management.

Our RefXAS database implements automated quality control to ensure data reliability and meet reproducibility standards through defined quality criteria. The current automated data processing and quality assessment system in RefXAS determines certain quality criteria. These include the edge step which is directly proportional to the elemental concentration and thickness of the sample (transmission) and the energy resolution which directly influences the ability to resolve fine spectral features for data interpretation. Further the usable k-range and the amplitude reduction factor are relevant for better interpretation of the data.

The automated quality control is developed under Python3 [19] and is based on functions from the package *Larch* [20]. The utilized function parameters are optimized for metal foils and are hard coded and fixed. This generally guarantees a stable processing and quality assessment of metal foil measurements but may constraint the utilization on differing samples. The fixed parameters may also not be optimal for the analysis of absorption spectra obtained with laboratory set-ups. To overcome this restriction a fitting routine for the parameters has to be introduced.

An implementation of a fitting procedure introduces new challenges. First the evaluation of the data will take more time. Depending on the complexity of the parameter space and the goodness of the provided starting parameter the fit can take a significant amount of time. Second, the fit is not guaranteed to succeed and if it did, the found optimum is not guaranteed to be the global optimum but rather a local optimum. Careful consideration of the parameter space is essential to optimise the fit routine. Selecting which parameters to vary and which to keep fixed minimises the search space, enhancing both speed and convergence. Setting appropriate starting parameters based on sample types further improves the efficiency and success of the fitting process. In order to find a global optimum different optimisation routines are possible, from slow brute force approaches to advanced routines like DIRECT [21] or SHGO [22]. These approaches are generally slower than local optimisation routines. An analysis of the kind of found optimum with local optimisation routines like least square, should therefore be applied in order to check if the found local optimum is the global or close to the global optimum.

The integration of NeXus with quality control procedures offers a solution to enhance both data storage and automated assessment in RefXAS. NeXus facilitates the storage and retrieval of quality control parameters in a structured and machine-readable format, making it easier to implement techniques like ML for detecting patterns and anomalies in large datasets. By adopting NeXus, RefXAS will enable standardised data storage that is directly compatible with quality control routines. Overall, the integration of NeXus-based standardisation and a general fitting routine for an improved data processing will benefit the RefXAS database, and thereby the community, by providing reliable evaluations and a source of high quality data for research. Yet, its implementation has to be handled with care.

4. Machine Learning

The application of machine learning approaches to evaluate XAS data is an ever-growing field in the XAS community. To allow users to use and retrieve stored raw and processed data for their machine learning projects we are developing a dedicated application programming interface (API). In addition, integrating a large language model (LLM) within the database could significantly enhance user interaction by e.g. providing dynamic guidance during the data upload process. An LLM could help users by suggesting optimal metadata fields, ensuring completeness and accuracy in submissions, thereby reducing the likelihood of errors.

To yield the highest accuracy the utilized machine learning models should preferably be trained on real datasets. These datasets should be large, scaled, curated, commented and of high quality. All these requirements are in the scope of the RefXAS research database. We therefore aim to create a standard database for the XAS community filled with real measured referenced data. As of today, no such database is published. Current machine learning approaches in the XAS community rely either on simulations or on local datasets.

Such a standardized curated reference dataset is already a standard for different research fields, e.g. the MNIST dataset for number classification [23] or the CIFAR-100 dataset for image classification [24]. They help to boost the development of more and more advanced machine learning models in their designed fields. A similar database would be highly beneficial for the XAS community, since it could be used to evaluate and validate the performance of developed machine learning models. The comparability could initiate a competitive development boost which ultimately will also lead to more and more accurate machine learning models.

Another advantage of machine learning based on the RefXAS research database could be the improvement of the data processing discussed in section 4. With an advanced machine learning model, the time and stability of the data processing could be optimized. The model could either be utilized to process the data directly or alternatively return optimal parameters for the established processing routine.

5. Conclusions and outlook

This paper presents the ongoing development of the RefXAS platform, which aims to provide a structured and accessible database for X-ray absorption spectroscopy (XAS) data. While quality control measures are still being refined, we highlight current efforts to establish appropriate fitting techniques. We show a new standardised download format that includes all metadata and defined data output that can be used for further data analysis. Additionally, we discuss integrating institutional storage such as at DESY, ensuring sustainable long-term data management. This is feasible within the first half of 2025. As next steps the inclusion Sample IDs, the implementation of the NeXus format and machine learning-compatible data structures will further enhance the platform's functionality, especially in terms of data standardisation and analytical efficiency. A key distinguishing feature of RefXAS compared to other existing XAS databases is its focus on providing well-curated reference spectra along with essential metadata and online processing tools, making the data ready for reuse in a FAIR-compliant manner. Our platform simplifies data visualisation and analysis through automated tools. We are actively engaging with the maintainers of the International XAFS DB portal (https://ixdb.jxafs.org/), which acts as a central integration point for multiple databases. This collaboration aims to position RefXAS as both a standalone reference resource and a contributor to broader data-sharing initiatives within the XAS community.

Acknowledgements

This publication was written in the context of the work of the consortium DAPHNE4NFDI in association with the German National Research Data Infrastructure (NFDI) e.V. NFDI is financed by the Federal Republic of Germany and the 16 federal states, and the consortium is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 460248799 and own funding at the institutions. The authors would like to thank for funding, support and acknowledge all institutions and actors who are committed to the community and its goals.

References

[1] Barty A, et al. (DAPHNE4NFDI) 2023 (doi:10.5281/zenodo.8040606).
[2] Paripsa S, Gaur A, Förste F, Doronkin D E, Malzer W, Schlesiger C, Kanngießer B, Welter E, Grunwaldt J-D and Lützenkirchen-Hecht D 2024 *J. Synchrotron Rad.* **31** 1105-1117 doi:10.1107/S1600577524006751.
[3] Ravel B and Newville M 2016 *J. Phys. Conf. Ser.* **712** 012148 doi:10.1088/1742-6596/712/1/012148.
[4] Koranne S 2011 *Handbook of Open Source Tools* (Boston, MA: Springer) pp 191-200.
[5] Könnecke M, et al. 2015 *J. Appl. Crys.* **48** 301-305 doi:10.1107/S1600576714027575.
[6] Katz P 1989 *ZIP file format specification (*https://pkware.com).
[7] Ravel B and Newville M 2005 *J. Synchrotron Rad.* **12** 537-541 doi:10.1107/S0909049505012719.
[8] Gaur A, Paripsa S, Förste F, Doronkin D E, Malzer W, Schlesiger C, Kanngießer B, Welter E, Grunwaldt J-D and Lützenkirchen-Hecht D 2023 *Conf. on Res. Data Infras. (CoRDI)* **1** doi:/10.52825/CoRDI.v1i.258.
[9] Amazon S3 storage 2023 (https://aws.amazon.com/de/s3/).
[10] dCache file system at HIFIS 2024 (https://hifis-storage.desy.de/).
[11] KIT Scientific Computing Center (SCC) 2024 (https://www.scc.kit.edu/).
[12] OpenID Foundation 2014 (https://openid.net/specs/openid-connect-core-1_0.html).
[13] Helmholtz AAI 2024 (https://hifis.net/aai).
[14] Gashnikova D, et al. 2024 *ACS Catal.* **14** 14871-14886 doi:10.1021/acscatal.4c02077.
[15] Hövelmann S C and Murphy B M 2024 (https://doi.org/10.57892/100-49).
[16] Gorai M, Nene M. J. 2019, pp. 369-374, doi:10.1109/ICCCIS48478.2019.8974498.
[17] Q2XAFS 2023 (https://github.com/XraySpectroscopy/nexus_definitions).
[18] De Nolf W, Newville M, Fonda E 2024 (https://github.com/XraySpectroscopy/pynxxas/tree/main).
[19] Van Rossum G and Drake F L 2009 *Python 3 Reference Manual* (Scotts Valley, CA: CreateSpace).
[20] Newville M 2013 *J. Phys.: Conf. Ser.* **430** 012007 doi:10.1088/1742-6596/430/1/012007.
[21] Gablonsky J, Kelley C 2001 *J. Glob. Optim.* **21** 27-37 doi:10.1023/A:1017930332101.
[22] Endres S C, Sandrock C and Focke W W 2018 *J. Glob. Optim.* **72** 181-217 doi:10.1007/s10898-018-0645-y.
[23] Deng L 2012 *IEEE Signal Processing Magazine* ***29*** 141-142 doi:10.1109/MSP.2012.2211477.
[24] Krizhevsky A 2009 (Learning Multiple Layers of Features from Tiny Images).