Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials: Methods

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstm20

Integration of X-ray absorption fine structure databases for data-driven materials science

Masashi Ishii, Kosuke Tanabe, Asahiko Matsuda, Hironori Ofuchi, Takahiro Matsumoto, Toyonari Yaji, Yasuhiro Inada, Hiroaki Nitani, Masao Kimura & Kiyotaka Asakura

To cite this article: Masashi Ishii, Kosuke Tanabe, Asahiko Matsuda, Hironori Ofuchi, Takahiro Matsumoto, Toyonari Yaji, Yasuhiro Inada, Hiroaki Nitani, Masao Kimura & Kiyotaka Asakura (2023) Integration of X-ray absorption fine structure databases for data-driven materials science, Science and Technology of Advanced Materials: Methods, 3:1, 2197518, DOI: 10.1080/27660400.2023.2197518

To link to this article: https://doi.org/10.1080/27660400.2023.2197518

Published online: 17 Apr 2023.

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2023, VOL. 3, NO. 1, 2197518
https://doi.org/10.1080/27660400.2023.2197518

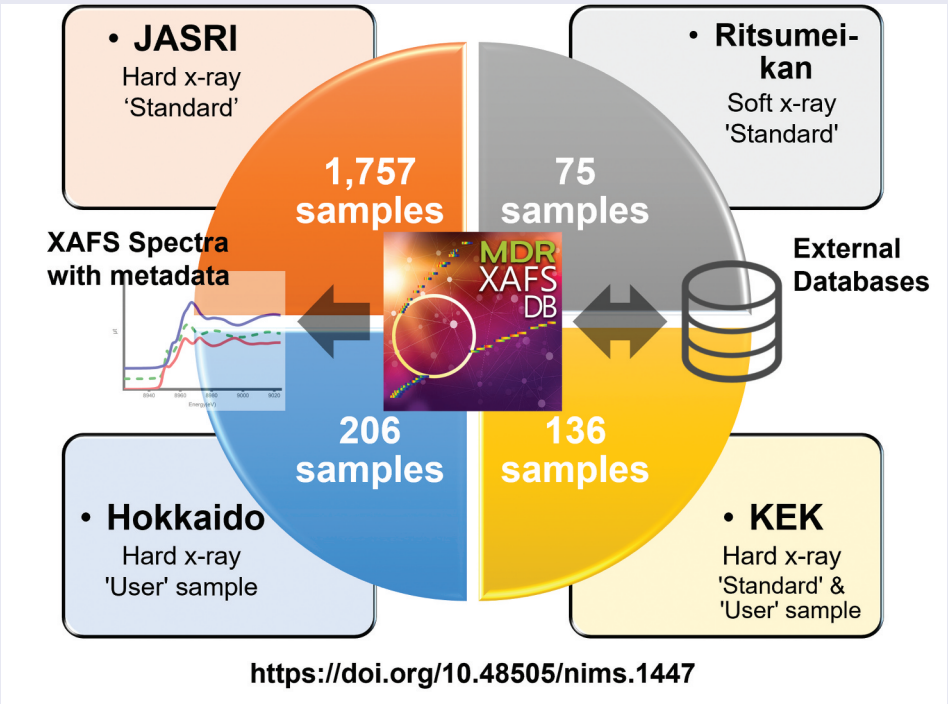
Integration of X-ray absorption fine structure databases for data-driven materials science

Masashi Ishii[a], Kosuke Tanabe[a], Asahiko Matsuda[a], Hironori Ofuchi[b], Takahiro Matsumoto[c], Toyonari Yaji[d], Yasuhiro Inada[d], Hiroaki Nitani[e], Masao Kimura[e] and Kiyotaka Asakura[f]

[a]Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), Tsukuba, Japan; [b]Center for Synchrotron Radiation Research, Japan Synchrotron Radiation Research Institute (JASRI), Sayo, Japan; [c]Information-Technology Promotion Division, Japan Synchrotron Radiation Research Institute (JASRI), Sayo, Japan; [d]SR Center, Ritsumeikan University, Kusatsu, Japan; [e]Institute of Materials Structure Science, High Energy Accelerator Research Organization (KEK), Tsukuba, Japan; [f]Institute for Catalysis, Hokkaido University, Sapporo, Japan

ABSTRACT

With the aim of introducing data-driven science and establishing an infrastructure for making X-ray absorption fine structure (XAFS) spectra findable and reusable, we have integrated XAFS databases in Japan. This integrated database (MDR XAFS DB) enables cross searching of spectra from more than 2000 samples and more than 700 unique materials with machine-readable metadata. The introduction of a materials dictionary with approximately 6000 synonyms has improved the search performance and links with large external databases have been established. In order to compare spectra in the database, the energy calibration policies of each institution were compiled, and the energy calibration methods across institutions were shown. This clarified how to utilize the MDR XAFS DB as a knowledge base. The database created through this cross-institution initiative is a model case for the further development of databases for other methods and material informatics using them.



ARTICLE HISTORY
Received 14 November 2022
Revised 11 February 2023
Accepted 25 March 2023

KEYWORDS
X-ray absorption fine structure; data integration; metadata; materials data repository; DOI; RDF

1. Introduction

While new data-driven scientific discoveries are progressing in various fields [1], ensuring sources of data has become a serious challenge. In particular, data collection in experimental science requires innovations due to the time-consuming tasks involved in data acquisition. There have been trials in many studies, for example, in the development of high-throughput experiments

CONTACT Masashi Ishii ISHII.Masashi@nims.go.jp Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), Tsukuba, Ibaraki 305-0044, Japan

using robotics and combinatorial techniques [2–4]. However, measurements that require a variety of experimental environments, such as operando [5] and low-temperature measurements, are not always suitable for such high-throughput experiments. For the accumulation of data from experiments that require diverse environments, one possible solution is the integration of data through the cooperation of related researchers [6]. Given the diverse range of users involved, the requirements for this data integration are as follows:

- The benefits of data integration should be not only in data-driven science but also in everyday research.
- The data and metadata should be in as few formats as possible (ideally one format).
- The publication infrastructure should be prepared as a repository with policies for data utilization, such as the FAIR Principles [7].
- The database infrastructure should have search functionality and not just storage online.

FAIR is an acronym for Findable, Accessible, Interoperable, Reusable, and is a basic guideline for the utilization of data.

The X-ray absorption fine structure (XAFS) [8,9] discussed in this paper is a typical synchrotron radiation experimental technique that provides the atomic-level local structure (bond length, coordination number, etc.) and electronic states of a specific element by exciting its inner-shell electrons. Atomic-scale observation areas have a high commonality even if the samples are intended for various applications or are processed in multiple ways. In other words, many researchers across different fields can discuss a single spectrum and feedback the knowledge they obtained from their samples. The establishment of a basis, by which various XAFS spectra can be superimposed and compared, activates research. We have established an infrastructure for sharing XAFS spectra by integrating XAFS databases in Japan. In this paper, we clarify the problems with integrating data and discuss the solutions attempted in this initiative.

2. Activities of XAFS database

In order to understand international trends in XAFS databases, we have summarized below well-known data provision services:

(1) Farrel Lytle Database (http://ixs.iit.edu/database/data/Farrel_Lytle_data/)

This is a collection of data measured by F. W. Lytle and is probably the world's oldest and largest XAFS database operated by the International X-ray Absorption Society (IXAS). There are over 7000 RAW data items, and PROCESSED data compressed into a standard format are also available.

(2) IXAS X-ray Absorption Data Library (https://xaslib.xrayabsorption.org/elem/)

This is operated by IXAS and publishes 20 absorption edges, with a total of 276 spectra, measured primarily at the Advanced Photon Source (APS) and the Stanford Synchrotron Radiation Lightsource (SSRL). The unique sample type is 105. Data is stored in the XAFS Data Interchange (XDI) Format [10], with metadata beginning with # + Key + Value in the header. It provides superior reuse of data.

(3) ID21 SULFUR XANES SPECTRA DATABASE (https://www.esrf.fr/home/UsersAndScience/Experiments/XNP/ID21/php.html)

This is a collection of data provided by the ID21 beamline users at the European Synchrotron Radiation Facility (ESRF). The database is particularly rich in chemical information on samples, which makes it easy to reuse data. Graphical and text data are provided. The database contains 43 inorganic and 29 organic material spectra.

In response to such XAFS database activity outside Japan, the database constructed in this initiative has successfully integrated the major XAFS databases currently available in Japan. The features of these databases are summarized below:

(4) BL14B2 XAFS Standard Sample Database (https://support.spring8.or.jp/BL/bl14b2/xafs/standardDB/)

The largest XAFS database in Japan, owned by SPring-8 and operated by Japan Synchrotron Radiation Research Institute (JASRI), contains spectral data on 1913 chemical substances. All of the measured samples are defined as 'Standard'. For example, for commercial products, information such as the supplier and model number are included in the metadata, making them traceable. The data can also be obtained in bulk by installing the downloader software provided.

(5) Hokkaido University XAFS DB (https://www.cat.hokudai.ac.jp/catdb/index.php?action=xafs_login_form&opnid=2)

Hokkaido University XAFS DB is the oldest XAFS database in Japan. It was developed in collaboration with the Japan XAFS Society (JXS) and is operated by the Institute for Catalysis (ICAT). Its history and operational policy are described in reference [6]. This reference pointed out the necessity of data integration for the XAFS community, and this was one of the triggers for this project. Currently, approximately 300 spectral data are included in the database.

(6) Ritsumeikan University Soft X-ray XAFS Database (http://www.ritsumei.ac.jp/acd/re/src/sx_xafs_db/)

This is open to the public at Ritsumeikan University, which has a soft X-ray synchrotron radiation facility. The database is operated by the Ritsumeikan SR Center. While most of the data are

hard X-ray XAFS spectra, this database is a valuable data source that complements the spectra of light elements. Currently, 194 spectra from 98 samples are available using the following detection techniques: Total Electron Yield (TEY), Partial Electron Yield (PEY), Partial Fluorescence Yield (PFY), Inverse Partial Fluorescence Yield (IPFY), and Total Fluorescence Yield (TFY).

(7) Photon Factory XAFS database (https://pfxafs.kek.jp/xafsdata/)

This database is published by the Institute of Materials Structure Science (IMSS), which operates the Photon Factory (PF). Data are registered by facility personnel and PF users, and currently 148 spectral data are publicly available. The metadata must be parsed from the header of the data file.

3. Integration of XAFS databases: issues and trials

We have integrated the databases (4)–(7) above in this initiative and created a new public infrastructure, the MDR XAFS DB [11]. The most important function of an integrated database is cross searching, and the two main issues in realizing this are summarized below:

- Designing and collecting metadata describing spectra and sample details
- Unifying the vocabulary used in the metadata, including not only metadata items (keys) but also descriptions (values)

Since XAFS experiments are usually performed at large synchrotron radiation facilities, the conditions of the storage ring for X-ray generation and the optical system for extraction of monochromatic X-rays can almost all be automatically obtained as metadata. The problem is how to collect user-dependent metadata, such as experimental conditions, in a defined format, that is, keys and values expressing sample composition, shape, customized measurement parameters, etc., since these can be written in a variety of ways. Therefore, the format of user-dependent metadata needs to be defined and structured. Another problem is that each synchrotron radiation facility has its own metadata descriptions. In the following, such individual metadata is referred to as 'local metadata'. Local metadata must eventually be integrated with data that is shared with other facilities. Even if the above issue is resolved, if the vocabulary used for keys and values is not unified, the search performance of the integrated database will deteriorate. In this study, we focused on the project goals of integrating XAFS spectral data and cross searches, and we found the following practical solutions to the above issues.

3.1. Design and collection of metadata

Although the data format of XAFS spectra is based on simple columns of incidence and absorption X-ray intensities in a certain photon energy range, various formats are available. In Japan, there are 9809 (PF and SPring-8 Standard), REX [12], and Athena [13] formats, etc., that are compatible with post-experimental data analysis software. Metadata is placed in the header, providing the metadata necessary for analysis and some additional information. However, considering data reuse, these few pieces of metadata are not sufficient, and a wide variety of metadata needs to be organized, as described below. In such cases, it is not desirable to include a few lines of metadata as a header, and it is necessary to prepare a structured metadata file separate to the data file. In other words, it is necessary to maintain the existing data file, add a structured metadata file, and consider how to use it as a new information source to achieve the desired functionality.

Here we describe the general concept of metadata and the methods we adopted to achieve this goal. Figure 1(a) conceptually shows a general metadata hierarchy (stacked metadata model). Figure 1(b) shows schematically the scale of the users of each hierarchy level. The first (top) level is metadata that is always present in any study, such as names, institutions, etc. Its users are broad, and its content is shallow and requires no specialized knowledge. The second level is large category metadata, such as specific measurements (e.g. synchrotron radiation experiments) and samples, which require a certain level of specialized knowledge and have fewer users. The third (bottom) level is metadata specific to XAFS that is highly specialized and has in-depth content with little commonality. Its users are limited to a small number of researchers in the materials field. In general, as shown in Figure 1(a), the number of metadata keys increases as the hierarchy becomes deeper, and it is necessary to handle a variety of contents. The relationship between (a) and (b) is that of a pyramid and an inverted pyramid. We believe that there is more than one way to use metadata, but the appropriate key should be used according to the purpose. It is desirable that all the keys are used for wide and shallow and narrow and deep use, as shown in Figure 1. Since the purpose of the MDR XAFS DB is a cross search, we extracted the keys in the first and second levels with a careful review, according to the purpose of the search.

We organized local metadata as shown in Table 1. The keys are classified according to the following purposes:

(1) Keys for general information
(2) Keys related to the reproducibility and reliability of XAFS experiments
(3) Keys necessary for the integration of XAFS spectrum data

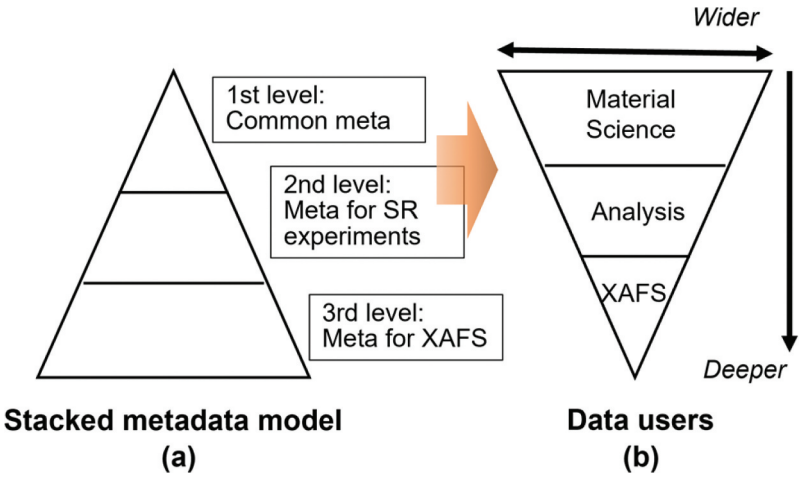


Figure 1. (a) Stacked metadata model with a hierarchy of keys that increase in number as they become more specialized, and (b) the scale of users at each level of the hierarchy.

Table 1. Categorization of keys contained in local metadata.

Purpose	Typical keys	Use case
Genearl information	Date, Experimenter, Facility, Beamline, Method, Sample	Comparison with other experimental data Discovering relevant data
Reproducibility and reliability of XAFS experiments	Monochromator, Mirror, Slit, Energy calibration, Number of measurement points, Step width, Ion chamber gas, Amplifier gain	Accuracy evaluation, Detection limits, Reproduction of experiments, Precise analyses
Integration of XAFS spectrum data	Column name, Unit, Data format	Big data creation, Statistical analysis, Machine learning

In the case of (2), it is highly specialized and not necessary for all researchers of materials, but it is essential for XAFS researchers. Therefore, (2) corresponds to the third level in Figure 1(a). And (3) is information necessary for recent data-driven research. That is, in order to perform big data creation, statistical analysis, and machine learning, information about the definition of the content in each column and its data format is necessary at the data merging stage. In addition, since multiple data formats are mixed in the MDR XAFS DB, as mentioned above, this information is necessary for XAFS spectrum analysts.

Consequently, most of the metadata in (2) and (3) are necessary for data use but not for cross searches. It is clear that general information in (1), e.g. beamline name, measurement technique, and sample name, is suitable for cross searches. And the number of metadata commonly handled here is likely to be less than 10. We will discuss in Section 4.2 what keys to assign and uses for these general metadata, including the constraints of the actual data infrastructure.

3.2. Unification of vocabulary

Examples of successful lexicon creation can be seen in Wikidata projects (https://www.wikidata.org/wiki/Wikidata:Main_Page). There, each vocabulary is uniquely managed by assigning IDs to each vocabulary in turn, and synonyms are registered to prevent vocabulary fluctuations. National Institute for Materials Science (NIMS) has adopted a similar system to manage research vocabulary and has established the materials vocabulary platform (MatVoc), which manages material names and other information using IDs called QIDs. This platform is already in use in the search system and was released to the public in January 2023 (https://matvoc.nims.go.jp/explore/ja/dictionary/Q713). We have used this dictionary to streamline the process of checking whether the material is the same as previously registered data. Currently, this work is performed manually by the database editor, but in the future, it may be used by users to identify names when registering data, and furthermore, it may be automated by machines. Lexicographic control is extremely important for material names, which are extremely diverse in the way they are described. However, as the registration of spectra by individuals begins in the future, it is quite possible that common names and abbreviations will be included in the metadata for beamlines and facilities as well, and the importance of vocabulary management is expected to increase. In fact, as discussed later, facility and beamline policies are incorporated into the energy calibration and metadata contents, thus they can be parameters for data screening.

Furthermore, these IDs are also used as Uniform Resource Identifier (URI), which forms a space of material-related lexicons, a namespace, and is publicly

available (http://dice.nims.go.jp/ontology/mdr-xafs-ont/Item#).

In this space, one can find the standardized name of materials and their QIDs and chemical formulas (if present). For example, the QID for tin(II) chloride dihydrate is Q2307, and the following URI has content in machine-readable format (http://dice.nims.go.jp/ontology/mdr-xafs-ont/Item#Q2307).

There are currently 713 entities registered as XAFS-related material names, and the number of synonyms is about 6000. Within MatVoc, many materials are assigned Chemical Abstracts Service (CAS) registry numbers to manage the vocabulary in a favor of linkage with large external databases. The mapping to external URIs and the resulting validation of data linkage are discussed in Section 5.4. The details of the concept of data and vocabulary management in the project are not limited to the MDR XAFS DB but are general in nature and will be presented at another time.

4. Construction of MDR XAFS DB

4.1. Database policy

As described in Section 2, earlier efforts to build XAFS databases were done individually. Taking a broad view, it can be concluded that we are in a transitional period from the past, where spectral data only need to be understood by the person who measured them, and the recent policy that aims for a cyber society where understandable metadata are added to the data and shared with many people. In fact, some databases still follow the tradition of leaving information in the file name or sample name, which should be recorded separately as metadata, to serve as a reminder to the person who recorded it. On the other hand, databases that seek to collect data systematically have machine-readable metadata, even though they cannot follow pioneering standard data formats such as NeXus [14]. Therefore, deep data linkage is possible through an interface that allows correspondence to be established. Although these differences in policies among the participating institutions were a challenge in integrating the databases, a construction policy was formulated and the integrated database MDR XAFS DB was constructed based on this policy. Here, Material Data Repository (MDR) [15], as the database infrastructure, is operated as part of a data platform project that has been underway at NIMS since 2017.

MDR has functions and operational policies suitable for open data in accordance with the FAIR Principles, which is becoming a fundamental concept for data utilization. Notably, data registered in the MDR is assigned a Digital Object Identifier (DOI) to enhance the visibility of the data. It also has an Application Programming Interface (API) function, which enables not only a Graphical User Interface (GUI) but also large data unit operations that are suitable for data-driven science. The repository in this project is divided into three main areas: publications, datasets, and collections that systematically archive data. At the time of writing this paper, approximately 1272 publications and 2370 datasets have been registered. Each data set in the XAFS DB is stored in the datasets area, and all data are also registered in a collection for systematic browsing. Currently, there are 15 similar systematically organized datasets, that is, collections. The MDR is an open data repository and can be used according to the license granted to each piece of data.

Considering the background so far, i.e. the requirements from the XAFS community, including the cross searches described in Section 3, and MDR's engineering abilities, we decided on the following construction policy for the MDR XAFS DB:

- Each spectral data provided by each institution must be accompanied by a structured local metadata file in Yet Another Markup Language (YAML) format.
- Keys in the local metadata should be standardized so that the data can be searched seamlessly without being aware of the differences between data-providing institutions.
- The keys to be standardized are the names of materials, chemical formulas, absorption edges, beamline names, and monochromator crystals.
- The set of metadata and the spectral data of the sample and reference sample should be defined as '1 Work', and each Work should be assigned a DOI.
- Each data providing institution is responsible for the quality and rights of the data, and data that have already been published should be used.
- The data to be released in the MDR XAFS DB should be open-access spectra and their supplementary data only, and the license should be Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) [16].

4.2. Metadata implementation for cross searches

The policy of Section 4.1 had to be consistent with the cross-search requirements discussed in Section 3. That is, the names of materials, chemical formulas, absorption edges, and spectrometer crystals had to be extracted from the local metadata provided in YAML format by each participating institution and then embedded in the MDR metadata. Since MDR is not a specialized repository for a specific area of materials science, it is not suitable for creating an advanced database customized for a single purpose, i.e. XAFS. On the other hand, it is advantageous for linking with

other data in MDR because it integrates data from a wide range of areas that are not limited to XAFS. In any case, based on this data provision concept, the MDR has its own data structure and rules for input (schema) [17], so it was not possible to fit all the key values for these cross searches into the MDR metadata. For example, with beamline names there is no commonality except for synchrotron radiation experiments, and there are no applicable keys in the MDR metadata schema. Therefore, the following keys for cross searches were extracted from the local metadata of each organization and implemented as values for 'Keyword', which is one of the keys in the MDR metadata schema. The following is an example of keywords extracted in YAML format:

subjects:
- subject: Nickel # Material name
- subject: Ni # Chemical formula
- subject: Ni K-edge # Absorption edge
- subject: Pure metal # Material superordinate
- subject: Si(111) # Monochromator crystals
- subject: BL-12C # Beamline name
- subject: Photon Factory # Data provider
- subject: XAFS # Measurement method (fixed)
- subject: collection – MDR XAFS DB # Identification of collection (fixed)

The comment text after the # is for ease of understanding for the reader and the definition of the value. Although metadata keys should be precisely defined, the polymorphic key 'Subject' is utilized here. This is because it follows DataCite's schema for obtaining DOIs (https://datacite.org/), but it should be noted that this key is used only for the index for cross-search in MDR. As described below, we have demonstrated that these simplified keys are sufficient for screening data. When cross-searching many fields, the use of a univocal key may inadvertently limit the search target. The advantage of the MDR keyword function is that users can filter the data by sequentially selecting these keys. For example, selecting 'Absorption edge' filters out relevant excitation elements, followed by 'Material superordinate' to obtain to the desired material system. Here, the vocabulary used in the keywords should be the nomenclature as described in Section 3.2 so that users can search the data seamlessly regardless of the institutions registered. Furthermore, it is also possible to select an institution by choosing 'Data provider' in the keywords.

4.3. Database management

These cross-institutional initiatives require systematic database management. This section describes how data are registered, assigned DOIs, and maintained. As shown in Figure 2, data registration begins with the submission of spectral data and local metadata including necessary information, such as data provider information and rights statements. Registration is completed when it is confirmed the registration data are displayed correctly on the test server. Within MDR, after the DOI is issued via electronic submission, the data is added to the MDR XAFS DB in the MDR Collection and eventually released to the public. The cross-search keywords described in Section 4.2 are also used to obtain DOIs and are the target of searches by DataCite, an organization that grants DOIs for research data. Automating and simplifying the registration procedure make it easier for users to register data directly in the future. Data registration is a joint initiative of materials scientists, engineers in charge of MDR, and service team members to handle data from the data-providing institutions that have contracts with NIMS. The contract procedure guarantees the legality of data use, and the names of these responsible institutions also appear in the keywords mentioned above. The granting of a DOI makes spectral data not just stored data but also carries with it the responsibility of publication. For example, due to the persistence of DOIs, if a serious error is found, a tombstone page is created indicating the reason for the error. Indeed, tombstone pages have been created for seven spectral data so far. This situation is undesirable, and further consideration should be given to how

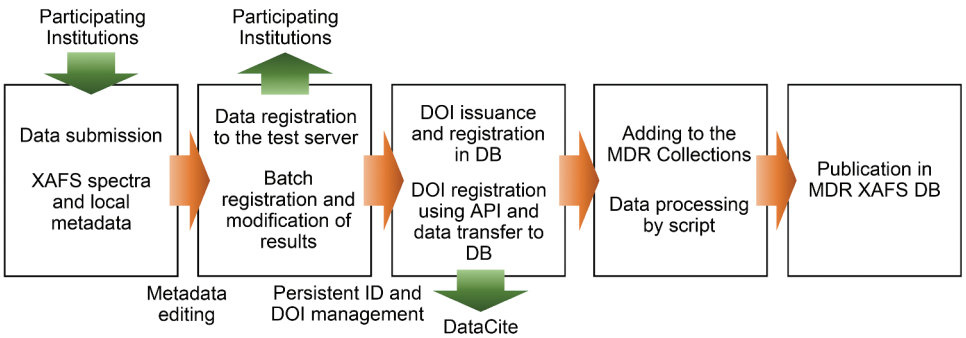


Figure 2. Spectra registration flow for publication in MDR XAFS.

much effort needs to be devoted to the peer review of registration data.

5. Contents of MDR XAFS DB

5.1. Statistics

As of September 2022, the statistical information of the MDR XAFS DB, which was created by integrating the databases of the four institutions described above, is as follows:

Total number of data: 2174 (contains 7 invalidated data with DOIs)

Total number of absorption edges: K-edge 1310 and L-edge 864

Unique absorption edges: K-edge 47 and L-edge 23

Unique materials: 713

Figures 3(a,b) summarize the number of K-edge and L-edge data, respectively, in histograms. As shown in these figures, the number of absorption edges is more than 100 spectra at the NiK-edge and W L-edge to the unregistered edge. In these figures, the number of highly monochromatic incident X-ray measurements using Si(311) as the monochromator crystal are also shown in the line graph. Approximately 45% of the K-edge and 30% of the L-edge are high-resolution spectral measurements, and the MDR XAFSDB can easily filter these high-resolution spectra using the keyword, 'Si(311)'.

Figure 4 shows the number of registered absorption edges sorted in descending order of number. The inset shows the top 10 absorption edges marked in yellow in the figure and their spectral numbers for both K-edge and L-edge. More detailed registration numbers are listed on the MDR XAFS DB readme page (https://mdr.nims.go.jp/concern/datasets/vh53wz94c). The accumulation is also shown. The results show that 90% of spectra are covered by 24 elements in K-edge and 13 elements in L-edge, which roughly correspond to 50% of the major absorption edges, indicating that there are many absorption edges with low registration numbers. Ideally, these curves should increase linearly or follow a curve according to a strategic spectrum collection plan. We are considering extending the K-edge spectrum to the Zn-Zr region, where a gap is seen in Figure 3(a), and the L-edge spectrum to lighter elements. Establishing a cooperative system in the community, such as by supplying samples to participating institutions, is also desirable.

5.2. Metadata analysis

In this project, we have conducted a sample nomenclature with an emphasis on linking with other material data. However, in practical terms it is not sufficient to only use nomenclatures. Instead, it is necessary to map with more general, external information, for example, linking with the ID of a well-known large external database or providing detailed product information. Therefore, we investigated the keys related to samples in the local metadata of each data-providing institution. The metadata keys related to the samples and their numbers for the four institutions are summarized in Table 2. Since the names of the keys in

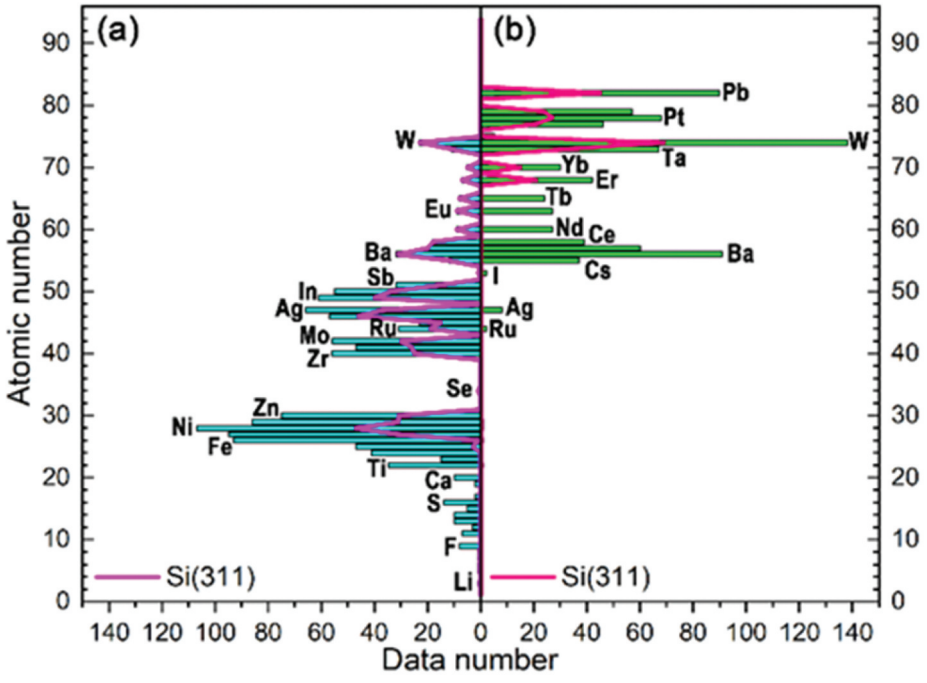
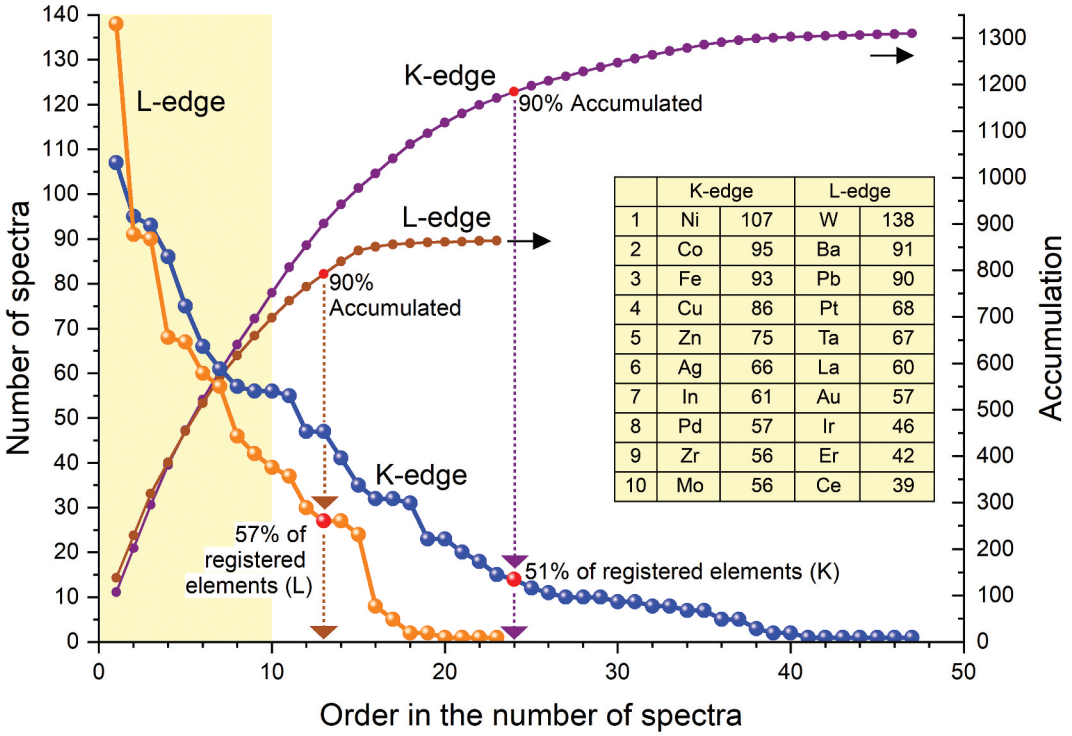


Figure 3. Number of data for (a) K-absorption edge and (b) L-absorption edge shown in histograms.



	K-edge		L-edge	
1	Ni	107	W	138
2	Co	95	Ba	91
3	Fe	93	Pb	90
4	Cu	86	Pt	68
5	Zn	75	Ta	67
6	Ag	66	La	60
7	In	61	Au	57
8	Pd	57	Ir	46
9	Zr	56	Er	42
10	Mo	56	Ce	39

Figure 4. Number of absorption edge spectra registered in the MDR XAFS DB sorted in descending order and their accumulation.

the local metadata of each institution are not unified at this time, keys with the same meaning are placed on the same line.

As summarized in Table 2 and the following paragraphs and beyond, it is clear that each facility has its own characteristics. Local metadata about the sample is entered using a user interface provided by the facility and merged with facility-specific metadata (e.g. storage ring current) and beamline metadata (e.g. optical element settings). In other words, metadata is not designed by individual users. Considering that once metadata is established, it will be used by many users, it is important to recognize that the characteristics will have a significant impact on the MDR XAFS DB.

In SPring-8, the metadata keys are designed to focus on identifying individual samples rather than linking with external databases. Therefore, information, such as supplier, model number, and lot number, is attached to almost all samples. In this way, each sample should have a well-defined individual ID along with a nomenclature ID. This process leads to complete data management, such that each study sample is traceable and retains its provenance and related properties. The average number of metadata on samples per Work (hereafter referred to as the average number of metadata) is the highest at 4.92 per Work. Here, it is necessary to explain why there are fewer chemical formulas than the number registered. In this database, there are registered samples, such as alloys and composites, that have names but no identification chemical formula. To the best of our knowledge, there are no data where the registrant forgot to include the chemical formula, so we conclude that the lack of a chemical formula does not prevent the use of the data.

Unlike SPring-8, Ritsumeikan University has set up metadata keys that can be linked to external databases. In fact, more than 90% of the registered data have CAS registry numbers. In addition, all samples are provided with additional data that is needed to understand the

Table 2. Metadata keys related to samples and the number of keys.

JASRI		Ritsumeikan University		Hokkaido University		KEK	
key	Number of value	key	Number of value	key	Number of value	key	Number of value
name	1757	name	75	name	206	name	136
chemical_formula	1684	chemical_formula	75	chemical_formula	206	chemical_formula	121
		CAS_number	68	CAS_number	169		
supplier	1753	manufacturer	31			manufacturer	2
model_number	1737	Product_number	24			product_number	1
lot_number	1715	sample_lot_number	16				
		additional_data	75	additional_metadata	121	additional_data	62
Total	8646	Total	364	Total	702	Total	322
Average	4.920/work	Average	4.853/work	Average	3.408/work	Average	2.368/work

experiments. Reflecting the fact that the measurements are made with soft X-rays and not transmissions, the sample shape information, such as 'powder on carbon tape', is provided. The average number of metadata is 4.85 per Work, which is comparable to that of SPring-8. All samples from Ritsumeikan University have chemical formulas.

Metadata for Hokkaido University and KEK were extracted from sample names freely written by users. In many cases, sample names incorporate experimental conditions in addition to the substance names and are written in original, non-standardized notations. For example, 'SUS316L Ni K-edge 18.2 K' is a typical example of an original sample name. Although experts, or those who did the experiment, can generally guess the meaning, the metadata creation method needs to be improved for future usage by third persons and computers that perform machine-learning analysis. The Japanese Society for Synchrotron Radiation Research (JSSRR) and JXS are currently working on a unified metadata format, and it is expected that users will provide the values (sample names) in the standardized metadata keys by themselves at the time of experiments in the future. The sharing of these issues with the XAFS community in the framework of the MDR XAFS DB project is expected to have a positive effect on data registration and cross-disciplinary data integration going forward. The average number of metadata for the Hokkaido University and KEK are 3.41 and 2.37 per Work, respectively. For the data from Hokkaido University, 80% of the extracted substance names were manually assigned CAS registry numbers in this project.

5.3. Energy calibration

The most important issue in XAFS measurements is the lack of a clearly defined absolute photon energy. When discussing fine structural details, such as peak attribution in X-ray absorption near edge structure (XANES) spectra, a comparison of various compounds is necessary. At the minimum, the relative energy relationship must be explicitly defined. In the MDR XAFS DB, where there are many independent registrants and measurers, it is inherently desirable to have a common energy standard. While an absolute energy calibration method using 'glitches' in the spectra caused by multiple-beam diffraction [18], highly accurate energy identification attempts [19], and well-organized historical tables [20] have been proposed, MDR XAFS DB adopts the relative energy calibration method using standard samples. In fact, this is because the absolute energy of any absorption edge has not been determined at this time. On the other hand, as shown below, there are no standardized guidelines for relative energy calibration, and data suppliers provide their own energy calibration methods.

All soft X-ray spectra provided by Ritsumeikan University adopt a method of calibrating a characteristic peak to a defined energy. An example of the definition of that energy calibration in local metadata in YAML format is shown below:

```
measurement:
  energy_calibration:
  – standard_sample: alpha-Al2O3
      calibration_position: white line peak maximum
      energy: 1567.71
      energy_unit: eV
```

This machine-readable metadata states that the energy of the white line peak of alpha-Al_2O_3 was set to 1567.71 eV for this measurement.

In all hard X-ray spectra provided by JASRI, metallic foils stable in air are used as reference samples. In cases where no suitable metallic foil is available, metallic powders, oxides, or metallic foils with adjacent absorption edge energies are used. This procedure is well established, so that all spectra provided by JASRI for the same absorption edge and the same monochromator crystal are uniquely calibrated. The spectra are not simply measured relative to a standard sample but are calibrated in a similar way to Ritsumeikan University as follows:

- For the Cu K-edge, the pre-edge peak is set to E = 8980.23 eV.
- When measuring absorption edges other than the Cu K-edge, energy calibration at the Cu K-edge should be performed first.
- If the energy of the absorption edge to be measured differs significantly from the value in the literature, then energy calibration is performed again using the value.

Many of the spectral data provided by Hokkaido University are attached to reference spectra, and although there is no prescribed calibration procedure, it is possible to compare spectra using a single energy axis at many absorption edges.

Therefore, as shown in the actual example of the Cu K-edge in Figure 5, (a) if we consider only the JASRI data, spectra of various materials can be shown in the same figure as is, and (b) with the Hokkaido University data, multiple spectra can be superimposed by appropriate calibration. However, as can be seen from the energy axis, there is no common reference point for both institutions. And when merging data, it would be ideal to use a common reference sample and calibrate the data before registration in the database. Figure 5 plots the data for each institution, but in this example, the Cu foil could be the common reference sample. Strictly speaking, the reference samples need

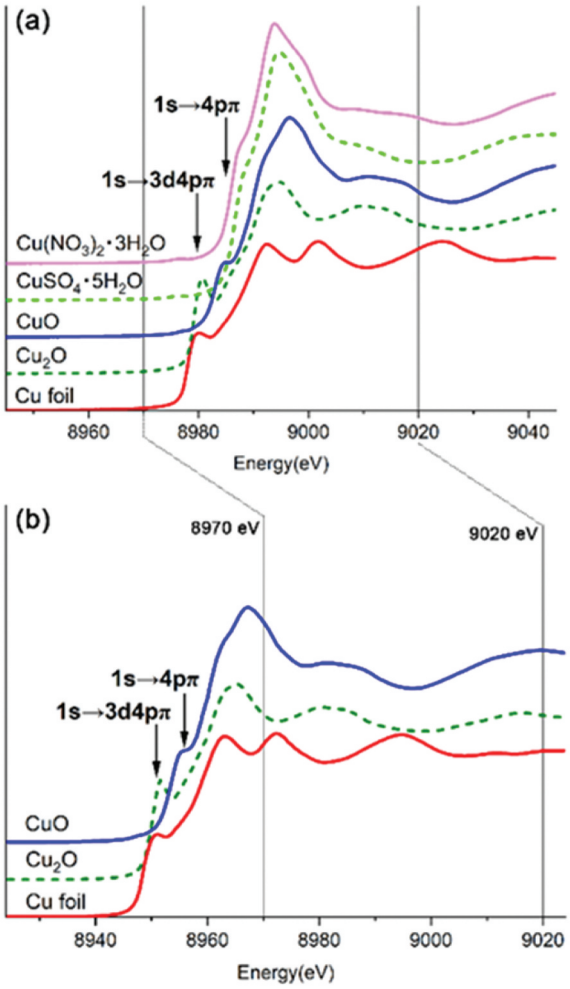


Figure 5. Examples of Cu K-edge spectra provided by (a) JASRI and (b) Hokkaido University.

to be identical and not just have the same material. But the limitations of such a method should be understood, due to the characteristics of each facility, beamline, and instant of X-rays.

Figure 6 shows the results of verifying this limitation using the actual spectra of Cu foils in the MDR XAFS DB, where first derivative dμt/dE spectra of Cu K-edge data provided by JASRI, Hokkaido University, and KEK are superimposed by applying two different methods of energy offsets. Figure 6(a) shows where the pre-edge peaks are aligned, and Figure 6(b) shows where the pre-edge leading edges (the first peak of dμt/dE spectra) are aligned. Since this figure is a differential spectrum, the energy at zero on the vertical axis E(dμt/dE = 0) indicates the peak or dip in the original XAFS spectrum. Here, Hokkaido University, KEK, and JASRI data are labeled with Hok, KEK, and SP8, respectively.

The inset summarizes the energy difference at E (dμt/dE = 0) for each of Hok and KEK from that of SP8. The energy difference with respect to SP8 is denoted as ΔE. The inset also shows the Cu K-edge XAFS spectrum as a dashed line, which shows which peak (dip) corresponds to which E(dμt/dE = 0). From these figures, the following can be understood:

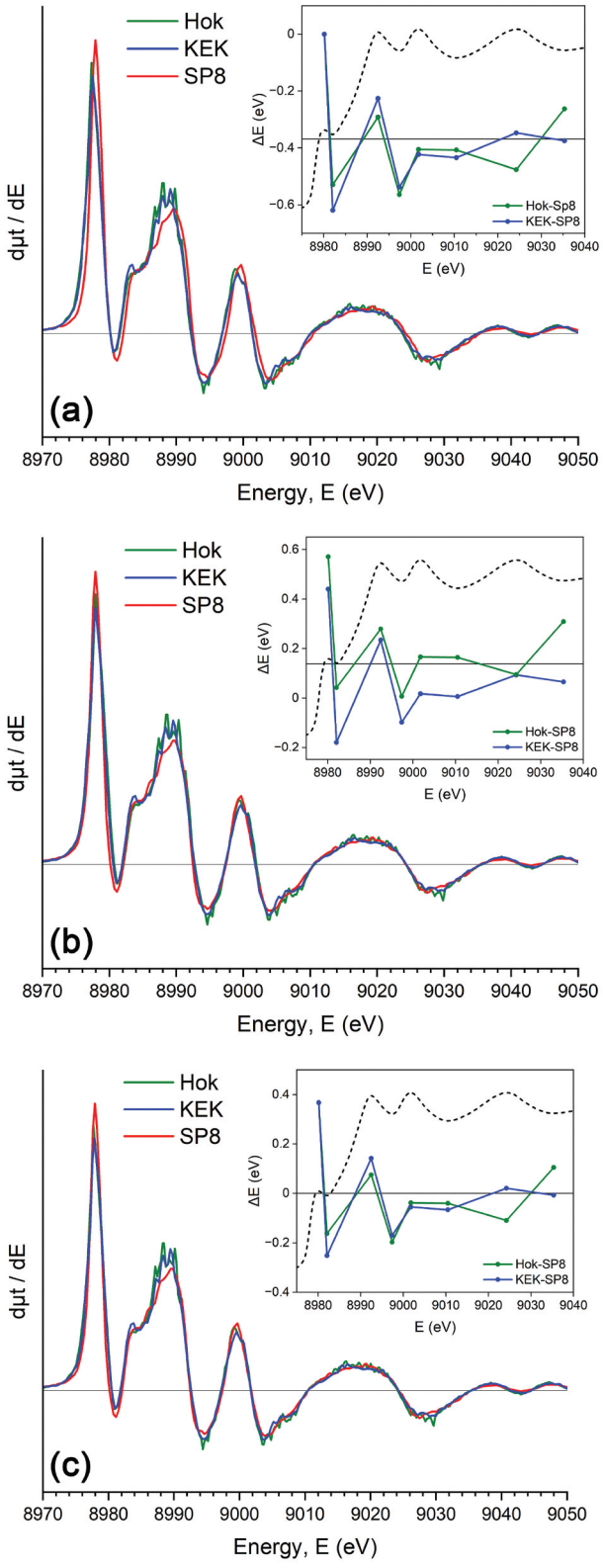


Figure 6. Comparison of two energy offset methods, (a) pre-edge peak and (b) leading edge alignments. (The inset shows the difference in photon energy at the peaks and dips.) (c) Result of the proposed method, i.e. energy correction that makes the sum of ΔEs zero.

The ΔE averaged over Hok and KEK together for Figures 6(a,b) were 0.37 eV and 0.14 eV, respectively, as indicated by the auxiliary lines in the inset. The absolute values are larger in Figure 6(a), indicating that the energy calibration of the three spectra is not

as well done as for Figure 6(b). This means that the commonly used method of aligning pre-edge peaks is not always optimal.

The pre-edge peaks have a large influence as a factor that makes ΔE large. In fact, the width of ΔE, i.e. the difference between its maximum and minimum values, is 0.62 eV and 0.75 eV in Figures 6(a,b), respectively. But it is 0.34 eV and 0.40 eV if the pre-edge peaks and dips are not included. This fact suggests that the electronic state of the pre-edge is sensitive to variations in individual samples, as well as to the intrinsic properties of Cu.

An example of an optimal method other than the offset using pre-edge peaks is shown in Figure 6(c). When a differential spectrum, as shown in Figure 6, is obtained, several ΔEs with the spectrum to be compared are obtained in the energy range to be analyzed, as shown in the inset. The offset energy, which gives the sum of these ΔEs zero, is considered to be plausible as a calibration. In fact, in Figure 6(c), the offset energies of Hok and KEK are 29.40 eV and 0.030 eV, respectively, to reduce the difference from SP8. In order to increase the reliability of the integrated XAFS database, it may be necessary to standardize the preparation and management of reference samples and X-ray beam monitoring methods.

5.4. Data federation

The Resource Description Framework (RDF) is an international model for data federation [21]. This method of representing information as a 'triple', the subject, predicate, and object, has been adopted in biotechnology for more than a decade. To facilitate data reuse in materials science, we have implemented RDF-based Semantic Web data linking the MDR XAFS DB. The federated RDF for connecting with huge external databases that is published in RDF format is available at 'MDR XAFS DB Readme' page (https://mdr.nims.go.jp/concern/datasets/vh53wz94c).

Here, data are described in triples using the SKOS (Simple Knowledge Organization System), an internationally standardized predicate for knowledge organizations [22].

This federated RDF describes connecting the QIDs of the aforementioned materials dictionary to the Compound IDs of PubChem, a huge and well-known database (https://pubchem.ncbi.nlm.nih.gov/), with the predicate SKOS:closeMatch. Here, the strictness of RDF can be understood from the fact that the definition of this predicate, SKOS:closeMatch, is given in the linkage with SKOS and is replaced by the namespace shown in Appendix B, http://www.w3.org/2004/02/skos/core#closeMatch. Using this RDF and the definition of MDR's Work (https://dice.nims.go.jp/en/ontology/about.html#mdr), we can combine XAFS spectra with DOI in MDR and SMILES (Simplified Molecular Input Line Entry System) and the molecular weight in PubChem into one table using the following SPARQL (SPARQL Protocol and RDF Query Language) [23] query, where integbio (https://integbio.jp/rdf/pubchem/sparql) is used as the endpoint for PubChem.

```
SELECT distinct ?label ?url ?smiles ?mw
WHERE {
  ?qid skos:closeMatch ?cid;
    rdfs:label ?label.
  ?mdr obo:RO_0000057 ?qid;
    rdfs:seeAlso ?url.
  SERVICE <https://integbio.jp/rdf/pubchem/
  sparql>;{
    ?cid sio:has-attribute ?attribute.
    ?attribute a sio:CHEMINF_000376;
      sio:has-value ?smiles.
    ?cid sio:has-attribute ?attribute2.
    ?attribute2 a sio:CHEMINF_000338;
      sio:has-value ?mw.
    }
} order by ?mw
```

The URIs of each namespace represented by prefixes such as 'rdfs': in this SPARQL and the variables used are summarized in Appendix B.

For example, the XAFS spectra of 49 organic compounds were linked to PubChem using skos:closeMatch, and SMILES and molecular weight information were added to these XAFS spectra. Since these organic compounds are organometallics covering almost all the major absorption edges shown in the inset of Figure 4, 1185 spectra can be used to discuss electronic states and structures with the PubChem reference data. Most of them are inorganic materials, but the comparison of electronic states using spectra provides a connection between organic and inorganic materials. One of the advantages of XAFS is that it can make links between these large material differences, and the MDR XAFS DB extends this advantage with Semantic Web technology.

6. Issues to be resolved

Below is a summary by the JXS of the remaining issues:

- While standard sample data collected systematically by participating institutions are easy to release, several barriers remain for the release of a wide variety of data provided by users, for example, how to deal with rights, such as data possession or how to describe metadata for special samples.
- How to maintain the quality of the data and whether to set criteria for data publication are

two other issues. At the minimum, it is necessary to follow the database policy described in Section 4.1, but it does not include quality assurance. Ideally, it is better to register only data that can be used reliably by anyone for any purpose, but it is difficult to determine the criteria for judging the reliability of data. Therefore, we have to decide how to create an equitable review process.

- How to design a unified metadata format across institutions and fill it in efficiently is another issue. It is not easy to create a unified metadata format that covers all the various XAFS methods, and there is no guarantee that everyone will follow that format. Although a minimal mapping and naming of metadata, as in the MDR XAFS DB, is useful for cross searches, we have not found a way to write machine-readable metadata, as discussed in Table 1, that fully guarantee the reproducibility and reliability of the experiments.
- How should the metadata of multi-dimensional data, such as time-resolved and micro-XAFS imaging data, be described and stored? MDR XAFS DB allows a variety of data formats. In fact, many of the registered metadata contain definitions of the formats used. However, when data formats for multi-dimensional methods are implemented, the definitions cannot be fully described in the metadata, and the guarantee that all data can be reused is rapidly lost. A common data format needs to be created to ensure database usability.

These issues will continue to be discussed, but the most important thing is to develop a culture of open data and show the specific benefits in return. We expect that these issues will be resolved sequentially as the MDR XAFS DB initiative moves forward.

7. Conclusion

Four Japanese institutions have collaborated to integrate X-ray absorption fine structure (XAFS) spectral databases. More than 2000 spectral data have been integrated in the photon energy range from soft to hard X-rays. The database MDR XAFS DB has achieved seamless cross searchability with the use of sample nomenclature so that database users do not have to be aware of the differences in the local metadata of the facilities that provide the data. The introduction of Semantic Web technologies also demonstrated the potential for collaborative use with external data. However, there are still issues to be resolved, such as the acceptance of multidimensional data by time- and space-resolved measurements and unification of metadata, which is necessary for more domain-specific use. The culture of open data has not yet been established in materials science, but we hope that this initiative will be a trigger to promote the utilization of materials data.

Acknowledgements

This research was supported by the TIA collaborative research program 'Kakehashi,' JSPS KAKENHI Grant Number 21K18024, and Grant-in-Aid for Transformative Research Areas (A) 22H05109 by JSPS, Japan. We also thank H. Nagao, H. Yoshikawa, M. Kanzaki, M. Shimizu, and K. Inaishi for their technical support.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Japan Society for the Promotion of Science [21K18024,22H05109]; Tsukuba Innovation Arena (TIA) [2022 Kakehashi_#32].

ORCID

Masashi Ishii http://orcid.org/0000-0003-0357-2832
Kosuke Tanabe http://orcid.org/0000-0002-9986-7223
Asahiko Matsuda http://orcid.org/0000-0001-5989-027X
Hironori Ofuchi http://orcid.org/0000-0003-1718-4427
Takahiro Matsumoto http://orcid.org/0000-0001-6949-5492
Yasuhiro Inada http://orcid.org/0000-0001-5772-4788
Hiroaki Nitani http://orcid.org/0000-0002-7155-0304
Masao Kimura http://orcid.org/0000-0001-8645-2224
Kiyotaka Asakura http://orcid.org/0000-0003-1077-5996

References

[1] Hey T, Tansley S, Tolle K, et al. The fourth paradigm: data-intensive scientific discovery. Redmond (WA): Microsoft Research; 2009.
[2] Pyzer-Knapp EO, Pitera JW, Staar PWJ, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput Mater. 2022;8(84):1–9.
[3] Vaucher AC, Zipoli F, Geluykens J, et al. Automated extraction of chemical synthesis actions from experimental procedures. Nat Commun. 2020;11(1):1–11.
[4] Jandeleit B, Schaefer DJ, Powers TS, et al. Combinatorial materials science and catalysis. Angew Chem Int Ed. 1999;38:2494–2532.
[5] Nurk G, Huthwelker T, Braun A, et al. Redox dynamics of sulphur with Ni/GDC anode during SOFC operation at mid- and low-range temperatures: an operando S K-edge XANES study. J Power Sources. 2013;240:448–457.
[6] Asakura K, Abe H, Kimura M. The challenge of constructing an international XAFS database. J Synchrotron Rad. 2018;25:967–971.
[7] FAIR principles – GO FAIR. [cited 2022 Oct 14]. Available from: https://www.go-fair.org/fair-principles/
[8] Kincaid BM, Eisenberger P. Synchrotron radiation studies of the K-edge photoabsorption spectra of Kr, Br2, and GeCl4: a comparison of theory and experiment. Phys Rev Lett. 1975;34(22):1361–1364.

[9] Rehr JJ, Albers RC. Theoretical approaches to x-ray absorption fine structure. Rev Mod Phys. 2000;72(3):621–654.
[10] XAS data interchange format draft specification, version 1.0. [cited 2022 Oct 14]. Available from: https://github.com/XraySpectroscopy/XAS-Data-Interchange/blob/master/specification/spec.md
[11] Ishii M, Nagao H, Tanabe K, et al. MDR XAFS DB. [cited 2022 Oct 14]. Available from: https://doi.org/10.48505/nims.1447
[12] Taguchi T, Ozawa T, Yashiro H. REX2000: yet another XAFS analysis package. Phys Scr. 2005; T115:205–206.
[13] Demeter: XAS data processing and analysis. [cited 2022 Oct 14]. Available from: https://bruceravel.github.io/demeter/
[14] Flannery D, Cottrell SP, King PJC. The application of the NeXus data format to ISIS muon data. Phys B Condens Matter. 2003;326(1–4):238–243.
[15] MDR: NIMS materials data repository. [cited 2022 Oct 14]. Available from: https://mdr.nims.go.jp; Ranganathan A, Matsuda A, Tanifuji M, et al. The 14th International Conference on Open Repositories (OR2019); 2019 Nov 26; Hamburg, Germany. The Development of an Integrated Next Generation Data Repository for Materials Science. [cited 2022 Oct 14]. Available from: https://doi.org/10.5281/zenodo.3553963
[16] Creative Commons BY-NC-SA Attribution-NonCommercial-ShareAlike 4.0 International [cited 2022 Oct 14]. Available from: https://creativecommons.org/licenses/by-nc-sa/4.0/
[17] Materials Data Platform Center, National Institute for Materials Science. MDR schema (version 2.0.0. [Data set]; 2022 [cited 2022 Oct 14]. Available from: https://doi.org/10.48505/nims.3239
[18] Arthur J. Use of simultaneous reflections for precise absolute energy calibration of x-rays. Rev Sci Instrum. 1989;60(7):2062–2063.
[19] Kraft S, Stümpel J, Becker P, et al. High resolution x-ray absorption spectroscopy with absolute energy calibration for the determination of absorption edge energies. Rev Sci Instrum. 1996;67(3):681–689.
[20] Bearden J. XR wavelengths. Rev Mod Phys. 1967;39(1):78–124.
[21] World Wide Web Consortium (W3C). RDF Resource Description Framework (RDF). [cited 2022 Oct 14]. Available from: https://www.w3.org/RDF/
[22] World Wide Web Consortium (W3C). SKOS simple knowledge organization system – home page. [cited 2022 Oct 14]. Available from: https://www.w3.org/2004/02/skos/
[23] World Wide Web Consortium (W3C). SPARQL 1.1 query language. [cited 2022 Oct 14]. Available from: https://www.w3.org/TR/sparql11-query/

Appendices

Appendix A

A list of DOIs for the spectra used in this paper is summarized below (Table A1).

Table A1. List of DOI for the XAFS spectra used in Figures 5 and 6.

Figure	Material	DOI
5(a)	Cu(NO3)2_3H2O	https://doi.org/10.48505/nims.2028
	CuSO4_5H2O	https://doi.org/10.48505/nims.1786
	CuO	https://doi.org/10.48505/nims.1767
	Cu2O	https://doi.org/10.48505/nims.2026
	Cu foil	https://doi.org/10.48505/nims.1759
5(b)	CuO	https://doi.org/10.48505/nims.3543
	Cu2O	https://doi.org/10.48505/nims.3544
	Cu	https://doi.org/10.48505/nims.3543
		https://doi.org/10.48505/nims.3544
6	Cu	https://doi.org/10.48505/nims.3544
		https://doi.org/10.48505/nims.3672
		https://doi.org/10.48505/nims.1759

Appendix B

URIs of each namespace represented by prefixes in SPARQL discussed in Section 5.4 are summarized as follows (Table B1).

This SPARQL query is based on the schema defined in http://rdf.ncbi.nlm.nih.gov/pubchem/compound/ and the schema published in PubChem http://rdf.ncbi.nlm.nih.gov/pubchem/compound/. The variable definitions used in the query are as follows (Table B2).

The contents of the query, which are translated into human-readable format, are as follows:

Get the PubChem Compound ID and the MDR XAFS DB URI of the corresponding material (including reference samples) while obtaining the SMILES and molecular weight of the material from PubChem.

Table B1. URI list for SPARQL used for data federation.

prefix	URI
compound	http://rdf.ncbi.nlm.nih.gov/pubchem/compound/
obo	http://rdf.ncbi.nlm.nih.gov/pubchem/compound/
rdfs	http://rdf.ncbi.nlm.nih.gov/pubchem/compound/
sio	http://rdf.ncbi.nlm.nih.gov/pubchem/compound/
skos	http://rdf.ncbi.nlm.nih.gov/pubchem/compound/

Table B2. Variable list for SPARQL used for the data federation URI list for SPARQL.

Parameter	Definition
?label	Name of the material as defined in our material dictionary
?qid	QID of the material
?cid	Compound ID of the material (PubChem data)
?url	URL of the spectral data of the material
?attribute	The name of a property attributed to the CID. Here, it means SMILES.
?smiles	SMILES of the material (PubChem data)
?attribute2	The name of aproperty attributed to the CID. Here, it means molecular weight.
?mw	Molecular weight of the material (PubChem data)