OSCARS
Open Science Clusters' Action for Research & Society

CDIF-4-XAS: Mappings from Community Standards to CDIF

D2: Semantic description of at least two XAS community standards using a CDIF profile (XAS-CDIF)

Steven Richard (Committee on Data of the International Science Council [CODATA]), Arofan Gregory (CODATA), Patrick Austin (Scientific Computing Department, Science and Technologies Facilities Council, UKRI), Heike Görzig (Helmholtz-Zentrum Berlin für Materialien und Energie), Simon Hodson (CODATA), Rolf Krahl (Helmholtz-Zentrum Berlin für Materialien und Energie), Markus Kubin (Helmholtz-Zentrum Berlin für Materialien und Energie, Helmholtz Metadata Collaboration), Leandro Liborio (Scientific Computing Department, Science and Technologies Facilities Council, UKRI), Abraham Nieva de la Hidalga (Cardiff University), Deirdre Lungley (UK Data Archive); Flavio Rizzolo (Statistics Canada); Vyacheslav Tykhonov (CODATA)

Funded by the European Union

CDIF-4-XAS project report metadata

Project number	01-314
Project acronym	CDIF-4-XAS
Project name	Describing X-ray spectroscopy data for cross domain use
Call	1st OSCARS Open Cal
Topic	OSCARS-PaNOSC
Type of action	Task 2 Overview of standards, vocabularies (and ontologies), data formats and practices within the XAS area (landscape analysis)
Project starting date	01/10/2024
Project duration	24 months
REPORTING PERIOD	
Period covered	from 01/10/2024 to 28/02/2025
Reporting period number	2
Periodic report date and version	24/10/2025, version 1
Authors	Steven Richard (Committee on Data of the International Science Council [CODATA]), Arofan Gregory (CODATA), Patrick Austin (Scientific Computing Department, Science and Technologies Facilities Council, UKRI), Heike Görzig (Helmholtz-Zentrum Berlin für Materialien und Energie), Simon Hodson (CODATA), Rolf Krahl (Helmholtz-Zentrum Berlin für Materialien und Energie), Markus Kubin (Helmholtz-Zentrum Berlin für Materialien und Energie, Helmholtz Metadata Collaboration), Leandro Liborio (Scientific Computing Department, Science and Technologies Facilities Council, UKRI), Abraham Nieva de la Hidalga (Cardiff University), Deirdre Lungley (UK Data Archive); Flavio Rizzolo (Statistics Canada); Vyacheslav Tykhonov (CODATA)
Contributions	SR and AG were the lead authors. All other authors contributed editing the draft during revision, and discussed and agreed on the coverage, content, and formatting of the document during the reported period.

CHANGE HISTORY

VERSION	PUBLICATION DATE	CHANGE
1.0	24/10/2025	First public version

Funded by the European Union

OSCARS

Contents

Funded by
the European Union

1. Introduction

This document and the associated spreadsheet and other materials present an initial attempt to take the major community standards used for X-Ray Absorption Spectroscopy (XAS) and map them to the standards recommended for cross-domain FAIR sharing of data by the Cross Domain Interoperability Framework (CDIF) guidelines. The current standards landscape within the XAS community has been extensively described in the document "Overview of X-Ray Absorption Spectroscopy standards, vocabularies (and ontologies), data formats and practices" [1]. This document builds on the analysis presented in that document as a concrete exploration of how the data and metadata from the two most common XAS standards can be expressed in a CDIF-described package to facilitate use across domain, institutional, and application boundaries. It is felt that a concrete application of the standards and guidelines involved will clearly indicate the next steps for implementation.

While the focus of this document is technical, there are also implications for how other activities by domain groups and standards bodies can most effectively be conducted. These implications can be provided as feedback to various groups, and these will be mentioned here. The specific recommendations to be made to such groups do not, however, form part of this deliverable, but will be formulated more completely in other project deliverables in the future.

Funded by
the European Union

2. Existing Standards and Scenarios of Use

The document “Overview of X-Ray Absorption Spectroscopy standards, vocabularies (and ontologies), data formats and practices” [1] clearly identifies two major standards within the XAS community which have become the focus for harmonization of data description and formatting. The first of these is the NeXus format [2, 3] using HDF5 as the file format and the NXxas application profile to hold relevant metadata and description. The second standard is the XAS Data Interchange (XDI) standard [2], which describes a text-based format for both metadata and data (for an overview of these see the “Section 5.1 Community defined standards” in [1]).

It should be noted that these two standards are used for different things: NXxas is used to describe the raw data resulting from one or more measurements or scans. While it can have additional application data added to it or linked from it, it builds on the raw data which it describes for these functions. It is used as a means of exchanging the raw data between applications, a task which may involve large amounts of data. Consequently, HDF5 has emerged as the best way to package the data in these scenarios.

XDI is much more focused on the analytical aspects of XAS. It is designed to “encapsulate a single spectrum of XAFS along with relevant metadata” [4]. The intention here is to provide this data for exchange between analysis programs, spreadsheets, and data visualization tools. In practical terms, the volume of data to be expressed as XDI is much smaller than that for NXxas, as both the data and metadata are combined in a single text-based format.

From the perspective of CDIF, the goal is to allow any potential users of the data to find and assess the data for FAIR purposes, which involves understanding what is in the data sets described. Further, it should be possible to programmatically access the data and transform it into the form necessary for reuse. This “FAIR” functionality is independent of the specific processes supported by the different XAS community standards, but relies on the information contained in those standard formats (as well as some additional information for cataloguing purposes.) These FAIR functions, however, are very much aligned with the stated goals of those who produce the community standards, as we see in their stated goals [5]:

- 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 following an agreed data schema).
- The publication infrastructure should be prepared as a repository with policies for

Funded by
the European Union

- data utilization, such as the FAIR Principles.
- The database infrastructure should have a search functionality and not just provide online storage.

For the purposes of the current mapping exercise, two major scenarios emerge: (1) a data set which is encoded in HDF5 according to the NXxas application profile of NeXus will be transformed into a "cross-domain" CDIF metadata description to accompany the HDF5 file; and (2) an XDI text file will have its data contents described using the "cross-domain" CDIF recommendations for needed metadata. In both cases, the consumer of the data may be ignorant of the source formats – that is, their systems may not be designed to work with the standards employed. Regardless, they should be provided with enough information to maximise the machine-accessibility of these data, and to support search and other FAIR functions.

It should be noted that within the XAS community, cataloguing metadata is generally held external to the data files, and this is largely true of both the community standards described here. CDIF combined the cataloguing and data description metadata into a single format for FAIR purposes, and in these cases the XAS data will often be supplemented by cataloguing metadata held in typical "standard" formats such as the DataCite metadata scheme. Additionally, the entire package may need to be supplemented with information regarding access, licensing, and fair use. Such information often comes from institutional policy guidance, and may not be expressed in a standard form.

One requirement of CDIF, which is not fully met by existing community standards, is the need for formal definitions of the concepts used in describing the data. These can be exposed as labels for columns of fields and similar descriptors of data. CDIF demands that these be described in a standard fashion, because FAIR users may be unfamiliar with the use of terms in the institution which has produced the data. While there are some formal definitions in the XAS community specifications, these are not expressed in a standard or comprehensive form. They can, however, form the basis of such an expression.

Note that CDIF metadata is always expressed in a JSON-LD syntax, optimized for FAIR use/reuse over the Web. It leverages both the familiar Javascript features of JSON, as well as the more powerful RDF features of Linked Data approaches.

The standards recommended by CDIF include Schema.org for cataloguing and discovery [6], W3C PROV for provenance information [7], W3C SKOS for controlled vocabularies [8], and the Data Documentation Initiative's Cross-Domain Integration (DDI-CDI) specification [9].

Funded by
the European Union

Each of these specifications are used according to profiles described in the CDIF guidelines, although the provenance guidelines are currently under development.

CDIF currently focuses on describing data sets encoded in text-based formats. The need to describe binary formats (including HDF5) has been noted in the initial release of the guidelines, and work in this area is on-going.

Funded by
the European Union

3. Inputs and Outputs

Given the scenarios described above, we can produce high-level schematics of the transformation processes which will be supported. These should make the spreadsheet mappings easier to follow, as there are several different inputs and outputs involved.

3.1 XDI-to-CDIF

The schematic in Figure 1 shows the mapping from XDI to the recommended CDIF-4-XAS specification. The inputs include an XDI file, which contains both a metadata section and a data section, and a set of cataloguing metadata (such as a DataCite file). The XDI file may be "passed through" the transformation to act as a data file in the resulting package, or the data may be pulled out and reduced to a simple data-only text file. The metadata is transformed in either case, and placed in different areas of the CDIF package (JSON-LD file and data), based on what type of information it is. The cataloguing of metadata is a fairly straightforward mapping into corresponding Schema.org fields.

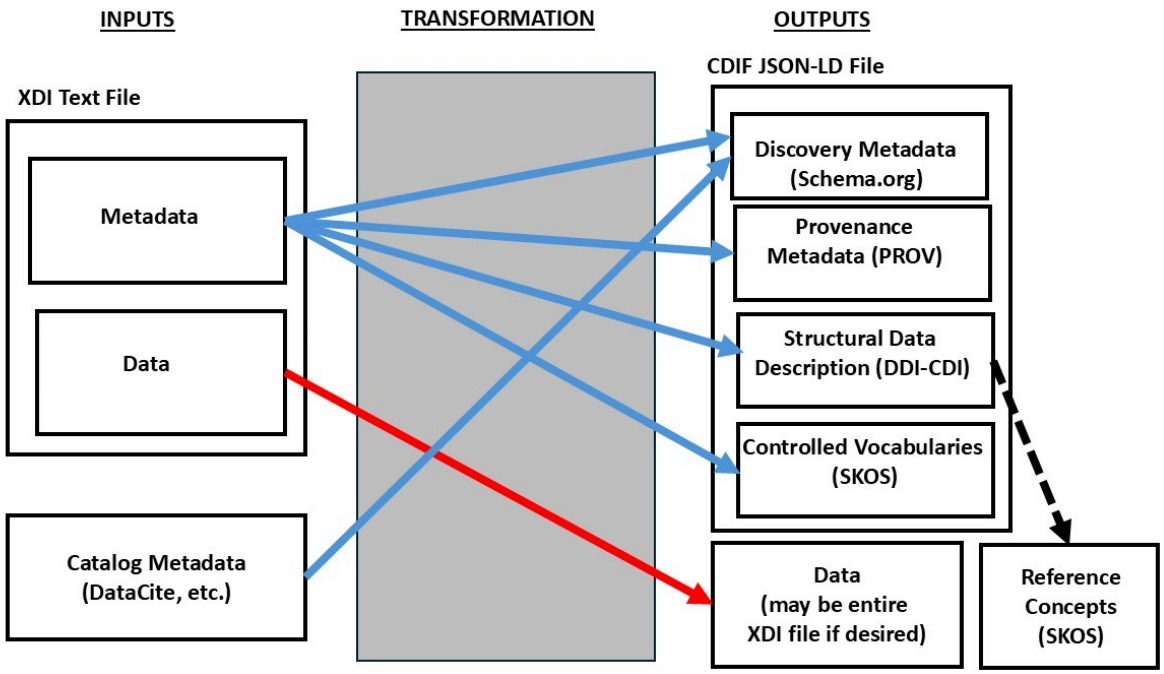

Figure 1 High level schema of XDI metadata extraction and mapping to the proposed CDIF-4-XAS profile

Note that CDIF uses SKOS to describe enumerated codes/values found in the data – this is the set of controlled vocabulary metadata inside the CDIF JSON-LD package. (The dashed

Funded by
the European Union

line indicates the use of these concepts using links.) The Reference Concepts are a static set of terms drawn from authoritative sources within the XAS community (or from authorities recognised by that community). This is also expressed as SKOS concepts in a SKOS concept scheme, but would be published externally to the CDIF package for each data set (it does not vary from data set to data set, so a single set of definitions can be reused).

Note also that the resulting resources are completely “open” in that they are all expressed in open standard formats or as structured text (for the data). No proprietary software is needed to access the files, and all of the standards are Web-based (JSON-LD/RDF, text) so that it is as easy as possible to access and use the resulting metadata. From a FAIR perspective, this is ideal, and very much in line with best practice as endorsed by the CDIF guidelines. The provided metadata is also very rich, giving complete definitions of terms, and indicating the structure and semantics of the data in a (potentially) comprehensive fashion.

3.2 NXxas-to-CDIF

The schematic in Figure 2 shows the mapping of NXxas HDF5 files to the recommended CDIF-4-XAS specification. This mapping is very similar to the XDI mapping. The main difference is that the data are always kept in a binary format that of the HDF5 source file. All data are referred to/from the metadata using path expressions which can be resolved by HDF5 software packages. The metadata is all stored in “open” JSON-LD form, in the same type of structure used for the XDI metadata.

Funded by
the European Union

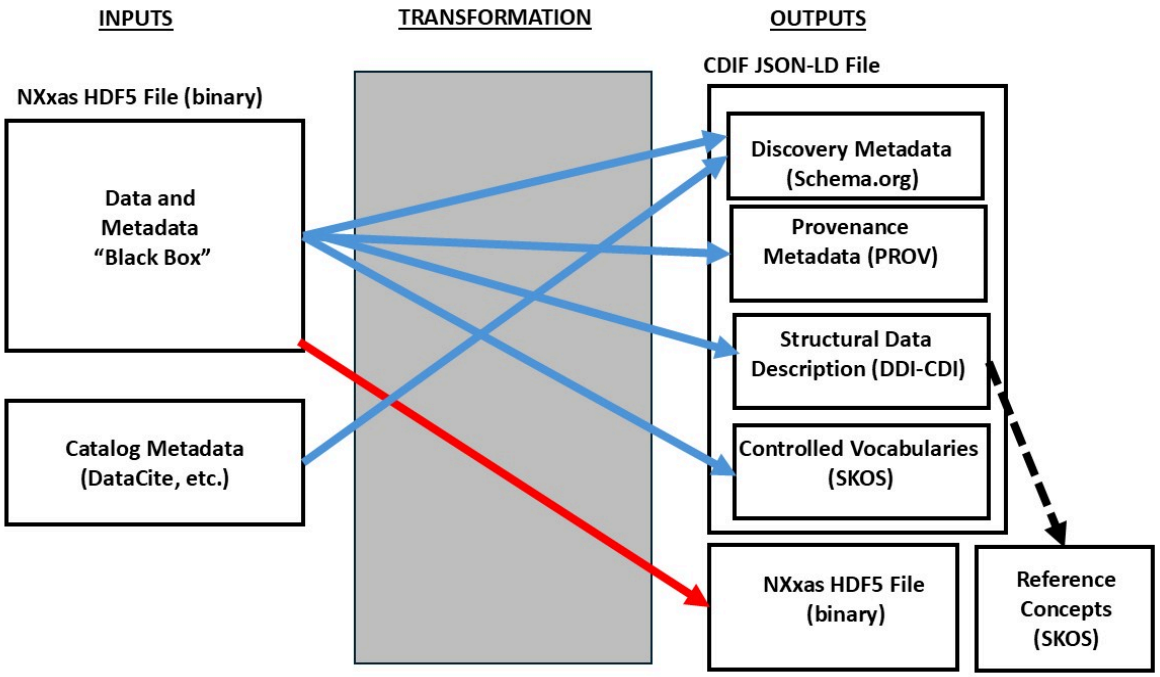

Figure 2 High level schema of NXxas metadata extraction and mapping to the proposed CDIF-4-XAS profile.

As before, the static Reference Concepts are expressed as SKOS, published separately, and the cataloguing metadata is (at least in part) coming from DataCite or some other standard cataloguing metadata format.

It should be noted that the mappings of metadata from NXxas to CDIF are different from those defined for XDI. This is due to the nature of NXxas within the NeXus/HD standardisation, which is designed to store and catalogue large sets of raw data combined with other measurement techniques, while XDI is designed for single, processed spectra data. Further, there is a wealth of information regarding the sample and collection of the measurements, which is not necessarily present in the XDI format. Much of this metadata is expressed in PROV in the CDIF package, which is not as heavily used in the XDI mapping.

The biggest difference here is that, from a FAIR perspective, the need to have dedicated HDF5 software to access the data makes this slightly less open. However, the HDF5 libraries and associated tools are available under a BSD-like license for general use [10]. The practical cost of accessing the data for those who do not already use HDF5 is higher than with the XDI mapped data, because there is a requirement for applications to work with more than just plain text, so they have to process the specific HDF5 binary format, with its attendant

Funded by the European Union

learning curve. Despite this, both CDIF output packages have very rich metadata which is equally open, and from this perspective both are ideally FAIR.

3.3 Static Concept Definitions

The production of the Reference Concepts is performed in an off-line fashion. For this deliverable, we offer a draft set of concept definitions coming from the NXxas specifications [2, 3], XDI [4], IUCr [11], and IUPAC [12]. This is provided both as a human-readable spreadsheet, and in a machine-processible SKOS form (which is equivalent). Ideally, this set of Reference Concepts would be agreed by the community in a process led by the existing standardization groups. Once agreed, a single canonical version – including SKOS – could be published for use not just by CDIF but by all members of the community for any application where it proved useful. Agreed definitions are a foundational aspect of FAIR, and this includes their expression in a standard, machine-actionable form.

It should be noted that while we are using SKOS here, an ontology described in OWL would be another form of publication which would be fundamentally in line with the FAIR principles. Because SKOS is a more accessible standard format in many domains, it is preferred for the purposes of the CDIF guidelines. Both standard forms could be published in an aligned fashion for different applications.

Funded by
the European Union

4. Further Work

Several work items or issues have arisen from the initial mapping work and the discussions held during its development. In this section, these are briefly noted, with the intention that these will lead to specific actions, as appropriate, later in the project.

1. **New Data Description Features of DDI-CDI:** In this mapping, it became clear that there was a concrete requirement to be able to describe the HDF5 binary data file using the DDI-CDI metadata recommended by CDIF. This was a feature already targeted by the CDIF Working group as a priority, and it is hoped that the NXxas use case can serve as one significant input to that development. We believe that the approach taken here is in line with the overall design of CDIF in implementing support for the FAIR Data Principles.

2. **Extensions to the DDI-CDI Profile in CDIF:** Additionally, there may be some minor extensions to the CDIF profile of DDI-CDI which are needed, for the inclusion of classes which are required for XAS data but which were not in the current set of CDIF recommendations.

3. **Contributions to the CDIF Provenance Profile:** The CDIF Working Group is now focusing on data provenance and context, and part of what was used in these mappings comes from drafts of an extension to PROV being developed as part of the CDIF work. The requirements coming from this mapping are in line with the approach being taken by CDIF, and it is intended that the XAS use case be provided to that group, so that it can be directly addressed by the emerging solution in the CDIF recommendations as they evolve.

4. **Engaging the XAS Community in Publishing FAIR Reference Concepts:** This mapping has highlighted the need for agreed formal concept definitions within the domain. Further, it has provided an example of how these might be expressed as a machine-actionable resource, based on a draft taken from existing sources. This project cannot do more than advocate for the development of a real set of agreed reference concepts, but we intend to make the effort to promote this idea to the appropriate domain authorities, and in general to support the development of such a resource in line with FAIR implementation.

Funded by
the European Union

5. References

[1] A. Nieva de la Hidalga *et al.*, 'Overview of X-Ray Absorption Spectroscopy standards, vocabularies (and ontologies), data formats and practices', Zenodo, Mar. 2025. DOI: https://doi.org/10.5281/zenodo.14920226

[2] NeXus International Advisory Committee, P. R. Jemian, R. Berg, T. Richter, and J. Wuttke, 'NXxas documentation', 3.3.2.26. NXxas - nexus v2024.02 documentation. Accessed: Nov. 02, 2024. [Online]. Available: https://manual.nexusformat.org/classes/applications/NXxas.html

[3] NeXus International Advisory Committee, P. R. Jemian, R. Berg, T. Richter, and J. Wuttke, 'NXxasproc documentation', 3.3.2.26. NXxasproc - nexus v2024.02 documentation. Accessed: Nov. 02, 2024. [Online]. Available: https://manual.nexusformat.org/classes/applications/NXxasproc.html

[4] M. Newville, B. Ravel, V. A. Solé, and Wellenreuther, 'XAS Data Interchange Format Draft Specification, version 1.0', XAS Data Interchange Format. Accessed: Oct. 01, 2024. [Online]. Available: https://github.com/XraySpectroscopy/XAS-Data-Interchange/blob/master/specification/spec.md

[5] M. Ishii, 'International XAFS DB Portal', International XAFS DB Portal. Accessed: Jan. 28, 2024. [Online]. Available: https://ixdb.jxafs.org/

[6] 'Schema.org', Schema.org. Accessed: Feb. 03, 2025. [Online]. Available: https://schema.org

[7] T. Lebo, S. Satya, and D. McGuinnes, 'PROV-O: The PROV Ontology', PROV-O: The PROV Ontology. Accessed: Oct. 02, 2025. [Online]. Available: https://www.w3.org/TR/prov-o/

[8] A. Miles and S. Benchhofer, 'SKOS Simple Knowledge Organization System Reference', SKOS Simple Knowledge Organization System Reference. [Online]. Available: https://www.w3.org/TR/2009/REC-skos-reference-20090818/

[9] Data Documentation Initiative Alliance (DDI Alliance), 'DDI CDI Version 1.0', DDI-CDI v1.0. Accessed: Oct. 02, 2025. [Online]. Available: https://ddialliance.org/ddi-cdi_v1.0

Funded by the European Union

[10] The HDF Group, 'HDF Licenses', Licences The HDF group. Accessed: Oct. 02, 2025. [Online]. Available: https://www.hdfgroup.org/licenses/

[11] International Union of Crystallography, 'Online Dictionary of Crystallography', Online Dictionary of Crystallography. Accessed: Aug. 01, 2025. [Online]. Available: https://dictionary.iucr.org/Main_Page

[12] International Union of Pure and Applied Chemistry, J. Kaiser, and S. Chalk, 'The IUPAC Compendium of Chemical Terminology', IUPAC Gold Book. Available: https://goldbook.iupac.org/

Funded by the European Union