Journal of
Synchrotron Radiation

ISSN 0909-0495

Received 28 March 2012
Accepted 26 August 2012

Towards data format standardization for X-ray absorption spectroscopy

B. Ravel,[a]* J. R. Hester,[b] V. A. Solé[c] and M. Newville[d]

[a]National Institute of Standards and Technology, Gaithersburg, MD 20899, USA, [b]Bragg Institute, ANSTO, Locked Bag 2001, Kirrawee DC, NSW 2232, Australia, [c]European Synchrotron Radiation Facility, 6 rue Jules Horowitz, BP 220, 38043 Grenoble Cedex 9, France, and [d]Center for Advanced Radiation Studies, University of Chicago, Building 434A, Argonne National Laboratory, Argonne, IL 60439, USA. E-mail: bravel@bnl.gov

A working group on data format standardization for X-ray absorption spectroscopy (XAS) has recently formed under the auspices of the International X-ray Absorption Society and the XAFS Commission of the International Union of Crystallography. This group of beamline scientists and XAS practitioners has been tasked to propose data format standards to meet the needs of the world-wide XAS community. In this report, concepts for addressing three XAS data storage needs are presented: a single spectrum interchange format, a hierarchical format for multispectral X-ray experiment, and a relational database format for XAS data libraries.

Keywords: XAFS; standardization; data formats.

1. Introduction

In the years since the seminal 1971 paper (Sayers *et al.*, 1971) demonstrating quantitative analysis of the X-ray absorption spectroscopy fine structure (XAFS), X-ray absorption spectroscopy (XAS) has developed into a mature field used by researchers in a broad array of scientific disciplines. During those decades the world-wide XAS user community has developed best practices for all aspects of the XAS measurement, including beamline design, sample preparation, measurement practice, theory, and data processing and analysis. The measurement of XAS data is today one of the core competencies of synchrotron science and beamlines dedicated to XAS are available at almost every synchrotron in the world. In these four decades there has never been a sustained and broadly supported effort within the XAS community towards standardization of data formats despite the substantial benefit that data format standardization has brought to other disciplines, notably X-ray scattering. The authors of this report represent a working group consisting of beamline scientists and XAS practitioners from around the world working under the auspices of the International X-ray Absorption Society and the XAFS Commission of the International Union of Crystallography (IUCr) to develop a proposal for data format standardization for XAS and related measurement techniques. This report summarizes the results of the first meeting of this working group.

Data format standardization addresses a number of problems shared by measurement facilities, the scientists and engineers who develop and maintain the hardware and software at those facilities, the scientists who develop software related to theory and analysis, and the scientists who use all of those things in their research. These problems include (i) representing spectral data, (ii) representing metadata related to the spectral data, (iii) composing relationships between XAS spectra measured at different times or at different facilities, (iv) composing relationships between XAS spectra and other data measured during the same experiment, (v) archiving of data and metadata for storage and future use, (vi) preparation of data for subsequent processing and analysis, (vii) comparison of measured data with applicable theory, and (viii) deposition of data with journals or other repositories. Stated more simply, we seek standards for data formatting to facilitate sharing of data in the broadest possible sense. *We seek to share data across continents, decades and analysis toolkits.*

In the following we outline data format proposals for three different applications. In this report the term 'data format' includes both the syntax specification and the semantic and scientific meanings assigned to the data names appearing in the file. Two data format proposals address data interchange of a minimal unit of currency: a single XAS spectrum. A further proposal suggests using a hierarchical format to represent the result of a complex multispectral experiment in which one or more XAS spectra are measured along with other X-ray or non-X-ray measurement techniques. The final proposal is a database standard for data libraries.

2. An interchange format for XAS spectra

The irreducible unit of XAS data is a single XAS spectrum. The bare minimum of data and metadata required to encode an XAS spectrum is:

(i) A table of numbers representing the energy axis, $\mu(E)$, and the intensity of the incident X-ray beam (which is often referred to as I_0), as well as the uncertainties in the measurement of those numbers. In some situations this table consists instead of the photoelectron wavenumber and $\chi(k)$ and their associated uncertainties.

(ii) Identification of the absorbing species and absorption edge.

(iii) The value of the *d*-spacing of the crystal monochromator or line spacing of the grating monochromator used to provide monochromatic beam for the experiment.

In the language of this report the table of numbers is the data and the other items are examples of metadata, *i.e.* descriptive information about the specific instance of XAS data. This set of data and metadata has been identified as essential and irreducible for a full assessment and analysis of an XAS spectrum. The incident intensity is included in the data table so that systematic errors in a spectrum related to the photon source and optics can be distinguished from those related to the measured sample. The identification of the absorbing element and edge are required to identify unambiguously the elemental origin of the spectrum. The value of the monochromator *d*-spacing or line spacing is required to apply corrections to the energy axis owing to miscalibration of energy, angle or encoder value.

In addition to the required data table entries and metadata values, a particular representation of data might require additional data table entries or metadata values. The data interchange format must be sufficiently flexible to allow the addition of new data table columns and metadata values. For example, the data table might include signals from additional detectors or encoder values related to the energy axis. Additional metadata might include such things as the start and end times of a scan, details about the photon source or the optics used to condition the photon beam, details about the preparation of the sample, details about the sample environment, or other descriptive information.

The spectrum encoded in the data table and described by the metadata might represent a single scan of XAS data. An individual scan, however, is only one example of a spectrum that might be represented by an interchange standard. It is common practice, for example, to measure two or more scans on an individual sample and to merge those spectra by performing a statistical average. Subsequent data processing and analysis is then performed on the merged $\mu(E)$ spectrum. The merged $\mu(E)$ spectrum, not the individual scans, is the unit of currency we wish to capture in the interchange format. Often an XAS spectrum is but one result of a complex experiment. As an example, an XAS spectrum might be measured at a point in an X-ray fluorescence map or an XAS spectrum measured in an energy-loss channel might be extracted from a non-resonant inelastic scattering experiment. The $\mu(E)$ spectrum that might be extracted from this larger data set is in the unit of currency we wish to capture in the interchange format.

With the established contents of the data table, a dictionary of commonly recognized metadata and a specified format for the interchange file, several problems related to data transmission are obviated. XAS data in this interchange format can be readily transferred between people, beamlines, desktop and Internet applications, data archives and journals. Our working group is approaching a general accord on the content of the data table and metadata dictionary and we are currently evaluating two proposals for the syntax of the interchange format.

2.1. The XAS data interchange format proposal

The XAS data interchange (or XDI) proposal is a plain-text format loosely based on the structure of Internet email. Like the syntactic separation of email headers from the body of the email, the XDI header precedes and is syntactically separated from the data table. The header contains a small set of syntax elements used to positively identify metadata content in a manner that is easily understood both by human and computer readers. The representation of the metadata is line-based with obvious separations between metadata names and values. These names can correspond to a dictionary of defined metadata, but the syntax is sufficiently flexible to allow the introduction of additional kinds of metadata. An example might be parameters related to specific software for data processing or analysis.

The data table follows the header as white-space-separated columns of numbers. Fig. 1 shows an example of data in the XDI format. The syntax elements distinguishing the metadata header from the data table are chosen so that data in the XDI format can be imported without filters into many existing XAS analysis packages, although most existing programs are not equipped to use the metadata. Additionally, many general-purpose data processing and visualization packages can readily use files in this format. Thus, the XDI format is

```
# XDI/1.0
# Beamline.name: APS 10ID
# Beamline.d-spacing: 3.1356
# Scan.element: Fe
# Scan.edge: K
# Column.1: energy eV
# Column.2: mu
# Column.3: i0
#///
# Fe K-edge, Lepidocrocite powder
# on 4 layers of tape
#-------------------------------
# energy mu i0
6899.9609 -1.3070486 149013.70
6900.1421 -1.3006104 144864.70
6900.5449 -1.3033816 132978.70
6900.9678 -1.3059724 125444.70
...
```

Figure 1
An example of data in the XDI format. Only the first few data points of the data table are shown. The XDI version number is identified in line 1. Subsequent lines show a few examples of metadata with their syntactic elements, the hash symbol (#) identifying a header line and the colon symbol (:) separating a metadata name from its value. The line of slashes separates structured metadata from free-format user-supplied comments. The line of dashes separates the header from the data table. The columns are labeled immediately before the beginning of the data table.

immediately useful even in existing software that does not explicitly recognize and use the format and is able to represent XAFS data uniformly and consistently.

2.2. The xasCIF proposal

The Crystallographic Information Framework (CIF) (Hall *et al.*, 1991) was adopted by the IUCr as its preferred format for data exchange in 1991, and is now the *de facto* standard for communicating crystal structures in the scientific community. Despite its name, there is nothing particularly 'crystallographic' in the CIF specifications. CIF consists of a tightly specified text-based syntax coupled with a comprehensive dictionary mechanism for defining the meanings of tags appearing in data files. A data file written according to the CIF syntax consists of one or more data blocks, with each data block notionally corresponding to a single structure. Each data block contains a set of key-value pairs interspersed with tabular information. There are no restrictions on the order in which key-value pairs or tables appear in data blocks. Thus, with appropriate choice of data names to act as relational keys, a CIF data block may be mapped directly into a relational database.

A simple example of what a CIF data file containing XAS data might look like is shown in Fig. 2. Note that multiple tables in a single data block allow CIF data files to organize metadata in tabular form where appropriate. The data names `_xafs_detector.label` and `_xafs_ionisation_detector.label` in this example act as relational keys linking the two metadata tables together. Further details of the CIF syntax can be found in *International Tables for Crystallography* (Hall & Westbrook, 2005), and the advantages of relational databases are discussed further in §4.

```
data_v2o5_nanotube
  _xafs_absorber.atom    V
  _xafs_absorber.edge    K
  _xafs_source.identification   'KEK-PF BL20B'
  _xafs_source.location         'Tsukuba, Japan'
loop_
    _xafs_detectors.label
    _xafs_detectors.position
    _xafs_detectors.type
monitor      monitor    ionisation
io-detector  detector   ionisation
foil         foil       ionisation

loop_
    _xafs_ionisation_detector.label
    _xafs_ionisation_detector.gas_pressure
    _xafs_ionisation_detector.length
    _xafs_ionisation_detector.amplifier_type
    _xafs_ionisation_detector.amplifier_gain
monitor         1      10    'Keithley'    10
io-detector     1      20    'Keithley'    10
foil            1      5     'Keithley'    11

loop_
    _xafs_reduced.energy
    _xafs_reduced.absorbance
 5248.52108 0.813707373
 5258.29435 0.798733337
 5268.26606 0.781069442
 5278.27878 0.764530778
 5288.28697 0.748170706
 5298.19834 0.731395959
...
```

Figure 2
An example of how XAS data might be expressed in CIF format. Only the first few data points are shown. Comments follow hash (#) characters. Note that the actual content and data names chosen by this working group would differ from those shown here.

3. Hierarchical formatting of complex data sets

While text-based file formats are adequate for single spectra, they are not well suited for large or complex data sets. While a single spectrum is our single unit of currency, its information relevance as a single spectrum is quite different when it can be compared with other spectra acquired at other sample points in the context of a mapping experiment. The relationship among those spectra is well described by a hierarchical approach. This need becomes more obvious when we combine an arbitrary number of XAS measurements with other X-ray and non-X-ray measurements in a single experiment. For instance, one might save complete fluorescence spectra while performing a fluorescence XAS mapping experiment. While all this can be achieved with text files, binary file formats are far better suited to that task. A clear example in the XAS field is full-field transmission X-ray microscopy where storage of image data is seldom made in plain text.

XAS scientists are often uncomfortable with binary file formats. The main concern is transferability and readability of a binary file among individuals and across computing platforms. On the other hand, those same XAS scientists never object to storing experimental data on a physical format, such as a hard disk, a compact disk or a memory stick. It is taken as granted that the data can be recovered from the physical format and data are no longer stored in printed form. Clearly they are confident about the availability of tool sets to access the file system of the physical medium and to make the stored data accessible.

The Hierarchical File System version 5 (HDF5) format is an extant binary format which offers a close analogy to a file system. This file format is developed and maintained by the HDF group (HDF GROUP, 2012). An overview of the HDF5 capabilities from a scientific perspective can be found in the article by Dougherty *et al.* (2009). To summarize, any set of data can be stored in an HDF5 file in a self-descriptive way: numbers are retrieved as numbers retaining their original precision, text is retrieved as text, multi-dimensional data sets are retrieved in their original type and dimensions, machine endian-ness is automatically handled, automatic compression and decompression is supported, and so on. Data analysis programmers face fewer problems correctly reading an HDF5 file than a text-based file for which there is not a good library.

HDF5 itself provides a support and a set of tools that make sure the data are readable in their original form but does not add anything concerning their meaning. The versatility of the HDF5 format assures that the data conventions and the dictionary of XAS metadata developed for the data interchange format can be readily mapped onto HDF5. This, along

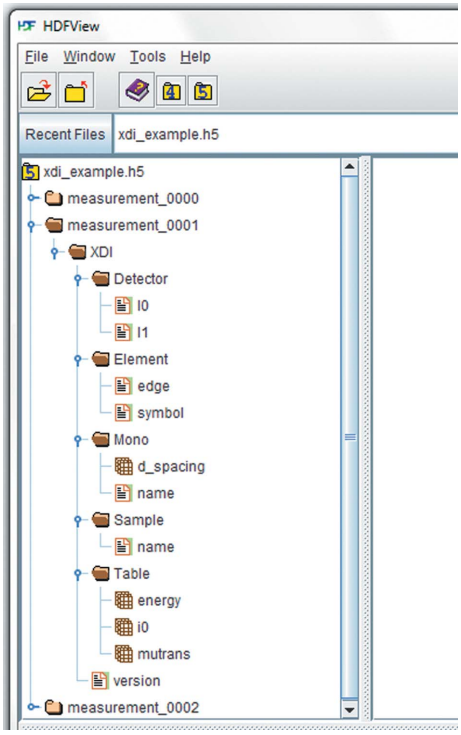

Figure 3
Think of an HDF5 file as a portable hard disk. The data and metadata content of an XAS measurement are accessed in a hierarchy. The HDF5 file can store any number of XAS measurements along with data and metadata from other X-ray and non-X-ray measurements.

with the fact that most European synchrotrons are already using HDF5 or planning to do so, implies that an HDF5-based implementation will be available for use in XAS analysis programs.

4. XAS data libraries

While the XDI format presented in §2.1 represents a single XAFS spectrum and is the fundamental quantum of exchangeable XAFS data, it is unable to convey much context in relation to other measurements. This includes not only other measurements made at the same time, as discussed in §3, but also measurements made on the sample under different conditions, such as temperature, or on different samples measured during the same experimental session. And yet, one of the motivations for adopting a standardized format is to compare and exchange XAFS spectra, and especially spectra on model or reference compounds measured at different places and times that may be used in comparative analysis of XANES and EXAFS spectra. For this, libraries of XAFS spectra are needed. While the XDI format itself is not capable of holding a library of spectra, its well described header format and clear meanings of column data make it easily inserted into and extracted from XAFS spectra libraries.

While some attempts (Lytle *et al.*, 1995; Newville *et al.*, 1999) have been made at building searchable on-line XAFS spectral libraries, these have suffered from difficulties in uniformity of format, vetting the quality of data, control and management of a centralized library, and the mechanism for adding and extracting spectra in a usable format. Parts of these difficulties are technical, while other parts are clearly more social and political. In this section we will focus on the technical aspects, and note that these may mitigate some of the more complicated non-technical issues.

The organization of digital data is well matched to using a relational model (Codd, 1970), and relational database management systems using this model and especially the structured query language (SQL) are both ubiquitous and well supported. The relational model differs from the hierarchical model described in §3 in that data are organized in tables without a strict hierarchy, but with relations defined between table entries. This allows more complex and interconnected relationship of data properties. For example, a database of XAFS spectra using a relational model could readily and efficiently support searching through a collection of spectra by absorbing element, absorption edge, beamline of data collection or date of collection.

Several high-quality and freely available SQL database systems exist and are proven solutions for many database applications. In particular for data XAFS spectral libraries, which we would expect to be relatively small by today's standards, relational databases that map to a single file are a reasonable solution. For reference, the Lytle collection contains roughly 16000 files of raw unfiltered data, and uses 500 Mb, which is probably a good estimate of the size needed for a database encompassing model compound spectra. The Public Domain SQLite (Hipp, 2012) library provides a relational database that can be accessed with SQL in a single portable file, and is used in a wide range of desktop applications, such as web browsers and media players, and is widely deployed for embedded data management in commercial mobile devices. The file format is well described, if binary, and many tools and programming languages can access the data contained in these files.

XAFS spectral libraries could be built using SQLite for data storage and the same metadata used in XDI files. This would allow, for example, a desktop application that worked like a modern media player to import, sort and organize spectra by absorbing element, beamline, facility, sample name and so on as described in the XDI metadata. Some of the advantages of such an approach, in which a standard SQLite database file contains a large suite of data, including the relations between spectra, include (i) a library can be shared in its entirety, (ii) libraries do not have to be centralized and public, but can hold data to be shared between trusted colleagues or publicly, and (iii) centralized repositories can hold 'libraries of libraries' as well as a full suite of data. We propose to use SQLite for spectral libraries, and are in the process of building a portable desktop application based on SQLite to hold and manage libraries of XAFS spectra. While this application will try to read in many data formats, it will use the same metadata fields as described for XDI and output plain text files only in XDI format. An example screenshot of the still-in-development application is shown in Fig. 4, using metadata imported from the XDI format. The spectral libraries used by this application

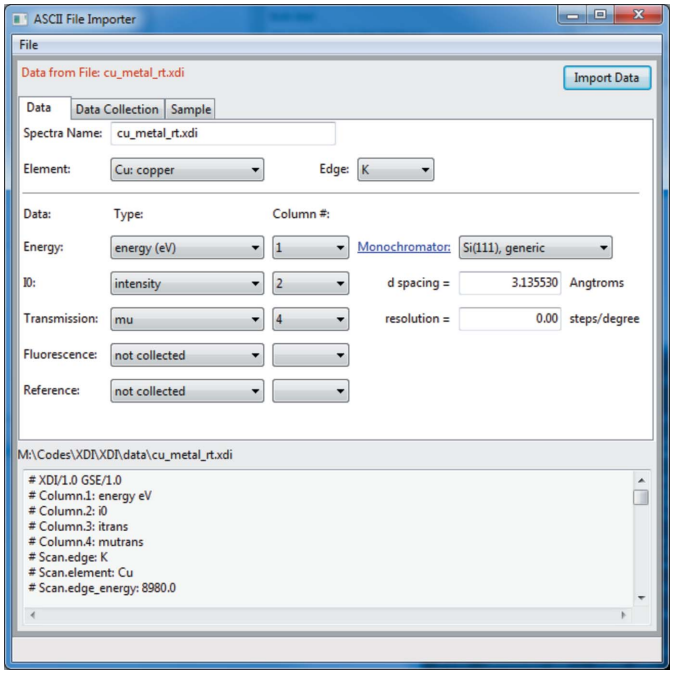

Figure 4
Example screenshot of an application using SQLite to manage libraries of XAFS spectra, using metadata imported from the XDI format. The spectral libraries used by this application are portable and XDI-format XAFS spectra can be extracted from the library for use in existing data analysis applications.

are portable, and XDI-format XAFS spectra can be extracted from the library for use in existing data.

5. Conclusion

The current draft of the metadata dictionary is under discussion using Internet social media resources, including a mailing list (XAS Data Format Working Group, 2012*a*) and a wiki (XAS Data Format Working Group, 2012*b*). The wiki is the forum at which the content of the metadata dictionary and the format proposals are presented. Further development continues on both the mailing list and the wiki. All community members are invited to contribute to our deliberations.

One of the challenges of defining the metadata dictionary is to distill the very large number of details that comprise an XAS measurement into a set of agreed-upon items of sufficient importance to merit a dictionary entry. Another challenge is to capture the salient aspects of all photon sources used for XAS measurement. For example, the polychromator for a dispersive XAS beamline is not adequately described only by its *d*-spacing. A final challenge is that metadata might not be well represented by the model of name/value pairs presented in the XDI proposal. xasCIF inherits from CIF a syntax for handling non-shallow data structures, as demonstrated by the first two `loop_` structures in Fig. 2. A modification to the XDI proposal to handle non-shallow metadata in that fashion would be straightforward.

We note that the data format needs of the growing field of X-ray emission spectroscopy (XES) are quite similar to those of the XAS community. The solutions suggested for XAS here map naturally onto the field of XES. The XAS data interchange and library standards can be adopted with little modification for XES spectra, while the hierarchical format can be adopted for various kinds of dispersive measurement geometries or resonant inelastic scattering experiments.

The proposals outlined above cover storage of a single spectrum (XDI and xasCIF), multi-dimensional data sets with potentially complex interrelationships (HDF5) and data libraries. Our working group will provide a common dictionary of metadata for use across all applications. For each of these applications our working group plans to provide both a vetted standard and appropriate application programming interfaces (APIs) so that these data format standards can be easily incorporated into new and existing software, including software for data acquisition, data archiving and data analysis. The metadata dictionary, the file format standards and the details of the APIs will be topics of future reports from this working group.

Finally, we acknowledge that two important topics have not been discussed in this report. We have addressed neither the representation of raw data as collected at the beamline nor the representation of the chain of analysis of an XAS spectrum. We in the working group consider both of these topics to be beyond the scope of our work. The nature of raw data is often idiosyncratic and characteristic of the particular beamline at which it was measured. Furthermore, existing beamlines have existing control and acquisition systems with working solutions for the representation of raw data. Our concern for data interchange is, therefore, focused on the representation of data that have been processed into the form of $\mu(E)$ [or perhaps $\chi(k)$]. We note that the syntax of either XDI or xasCIF is adequate for conventional XAS measurements consisting of signals from a small number of scalars. Either format could also be used by theory to encode $\mu(E)$ or other functions. The HDF5-based format from §3 is an attractive solution for XAS experiments involving more complex arrangements of detectors. That hierarchical format could also be applied to the capture of a complete analysis chain, including algorithm parametrization, user interaction and application of theory.

The authors thank Hiroyuki Oyanagi for organizing our working group and for offering continual encouragement and many suggestions. We are grateful to all the organizers of the Q2XAFS workshop. Finally, we thank Ken McIvor for his initial work on the formal grammar of the XDI format and for many helpful discussions.

References

Codd, E. F. (1970). *Commun. ACM*, **13**, 377–387.
Dougherty, M. T., Folk, M. J., Zadok, E., Bernstein, H. J., Bernstein, F. C., Eliceiri, K. W., Benger, W. & Best, C. (2009). *Commun. ACM*, **52**, 42–47.
Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.

Hall, S. R. & Westbrook, J. D. (2005). *International Tables for Crystallography*, Vol. G, ch. 2.2. Dordrecht: Kluwer.
HDF GROUP (2012). *HDF5*, http://www.hdfgroup.org/HDF5/.
Hipp, D. R. (2012). *Sqlite*, http://www.sqlite.org/.
Lytle, F., Boyanov, B. & Segre, S. (1995). *IXAS XAFS database*, http://ixs.iit.edu/database.
Newville, M., Carroll, S. A., O'Day, P. A., Waychunas, G. & Ebert, M. (1999). *J. Synchrotron Rad.* **6**, 276–277.
Sayers, D. E., Stern, E. A. & Lytle, F. W. (1971). *Phys. Rev. Lett.* **27**, 1204–1207.
XAS Data Format Working Group (2012*a*). *Xasformat mailing list*, http://millenia.cars.aps.anl.gov/mailman/listinfo/xasformat.
XAS Data Format Working Group (2012*b*). *Data format working group wiki*, https://github.com/XraySpectroscopy/Data-Format-Working-Group/wiki.