Formats for Sharing XAS Data, Metadata, and Results

Matthew Newville

Center for Advanced Radiation Sources,
The University of Chicago

Q2XAFS 2023: 2023-Aug-17

Thanks to: Bruce Ravel, V. Armando Sole, James Hester, Wout De Nolf, Mauro Rovezzi, Benjamin Watts, John Rehr, Benedikt Eggert . . .

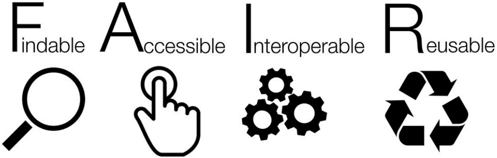
More info (examples, codes): https://tinyurl.com/nxxas2023

Motivation: Sharing XAS Data

The XAS community wants to be able to share XAS data and results with each other and with the wider scientific community, such as in online databases.

Many journals expect or require published data to be available as supplemental material in a downloadable, machine-readable format.

Many facilities and funding agencies are (or may soon) require data from X-ray beamlines be readily available to the public under *FAIR Data* principles.



We have seen this coming, we know we have to face it.

Can we share XAS spectra, and maybe analysis results in a way that works for us, the wider scientific community, the facilities, and general public?

Previous Work: XAFS Data Interchange (XDI) Format

Data formats were discussed at Q2XAFS2011. A Working Group (B. Ravel, J. Hester, V. A. Sole, G. Wellenruether, M. Newville) was formed to discuss and recommend formats for XAFS: See B. Ravel, et al, *J. Sync Rad* **19**, p869–874 (2012)

This work has two basic recommendations:

1. use plain-text (ASCII) files with clear and well-defined keyword tags for an individual XAFS spectrum: XDI or xasCIF.

   > *. . . 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. . .*

2. Use HDF5-based formats for more complex datasets:

   > *The HDF5-based format [above] 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.*

next: Review XDI, then: HDF5

The XDI Format

The initial design for XDI presented at Q2XAFS2011 and in the 2012 paper were refined, implemented, and presented at Q2XAFS2015 and the 2015 XAFS conference.
B. Ravel and M. Newville, *J. Physics: Conf Series* **712**, p12148 (2016).

Example XDI Data File

```
# XDI/1.0
# Column.1: energy eV
# Column.2: i0
# Column.3: itrans
# Element.edge: K
# Element.symbol: Zn
# Scan.edge_energy: 9659.0
# Mono.name: Si 111
# Mono.d_spacing: 3.13550
# Beamline.name: 13-BM-D
# Beamline.harmonic_rejection: Rh-coated mirror
# Facility.name: APS
# Facility.energy: 7.00 GeV
# Facility.xray_source: APS bending magnet
# Scan.start_time: 2008-04-10T17:00:26
# Detector.I0: 10cm  N2
# Detector.I1: 10cm  N2
# Sample.name: ZnSe
# Sample.prep: powder on tape, 6 layers
# ///
# room temperature
#----------------------------
#   energy          i0          itrans
    9509.000       103316.7       169556.2
    9514.000       100838.7       165838.2
    9519.000       100983.7       166450.2
```

- All lines in the header begin with `#`.
- The first line must have `# XDI`, with version number.
- Metadata must be formatted with syntax `# Family.Field:  Value`
- After `#///` freely formatted comments can be given.
- The header ends with `#----` followed by an optional line with column labels.
- There is 1 data table with consistent number of rows and column. Each row being a different energy.
- names of columns and some metadata values are strictly specified, with a dictionary of Family, Field names provided.

https://github.com/XraySpectroscopy/XAS-Data-Interchange/

Array Data in XDI Files

XDI specifies names for data arrays and for metadata. There is a limited and clearly defined list of names (case insensitive) for arrays.

Label	Meaning	Units (default)
energy	mono energy	eV, keV, pixel
angle	mono angle	degrees, radians
i0	monitor intensity	arbitrary
itrans	transmission intensity	arbitrary
ifluor	fluorescence intensity	arbitrary
irefer	reference intensity	arbitrary
mutrans	mu transmission	-log(itrans/i0)
mufluor	mu fluorescence	ifluor/i0
murefer	mu reference	unspecified

Some array labels for processed data are also defined:

k	wavenumber	Å^{-1}
chi	EXAFS	unitless
normtrans	normalized mu transmission	unitless
normfluor	normalized mu fluorescence	unitless
normrefer	normalized mu reference	unitless
r	radial distance	Å
chir_mag	magnitude of FT[chi(k)]	unspecified
chir_re	real part of FT[chi(k)]	unspecified
chir_im	imaginary part of FT[chi(k)]	unspecified

Labels are not exhaustive, but are the expected words to use for those meanings: ifluor, not if, not ifluo.

For $\mu(E)$ data, energy or angle should be in the first column. Units and mono d-spacing must be given in the metadata.

Please do not use angle. We are communicating XAS. It is a function of energy.

I am not aware of anyone using XDI for processed data (norm, $\chi(k)$, . . .).

More details: https://github.com/XraySpectroscopy/XAS-Data-Interchange/

MetaData in XDI Files

Metadata is formatted as `# Family.Field: Value` with these Family names:

Family	Contents
Column	data column labels and units
Element	absorbing atom
Mono	monochromator
Detector	detector details and settings
Beamline	beamline and its optics
Facility	synchrotron or facility used.
Sample	sample prep and conditions
Scan	Parameters of the XAS scan

Columns of array data are specified with `# Column.N: Label [Units]` with column number `N`, starting with 1. It is common (but not required) to also put array labels on a line between the line `#----` and the data table. For example:

Column Labels for Arrays

```
# XDI/1.0
# Column.1: energy eV
# Column.2: i0
# Column.3: itrans
... (more header lines)
#--------------
# energy     i0     itrans
```

There is a small set of **required metadata**:

Family.Field	Meaning
Element.symbol	Atomic symbol
Element.edge	IUPAC Level name (K, L3, . . .)
Mono.d_spacing	mono *d* in Å.

and a handful of **recommended metadata**

There are many optiona `Family.Field` pairs, and these can be expanded for some spectra types (XMCD, HERFD, . . .), or beamline-, sample-, or processing-specific metadata..

Several beamlines (including mine) are writing data with an "XDI-like" format, though maybe not with exact array and metadata names.

Multi-spectra files

Recap: XDI defines a single XAS spectrum in plain text, with clearly defined syntax, and has support code. These files will be useful for 50+ years. For databases, supplemental material for journals, and FAIR data sharing, we also want to share:

- many spectra, perhaps many hundreds of spectra.
- "more raw" data like indvidual arrays from multi-element detectors and dead-time-correcting arrays.
- non-XAS data as metadata: XES emission scan, XRD pattern, ...
- theoretical inputs, data processing parameters, intermediate results.

XDI is a good start, but we need something more. Getting something that will be "useful for 50+ years" is challenging.

Multi-spectra files possibilities

Some existing possibilities, with ranking of important criteria for use:

Format	**Readability**	**Longevity**	**Array Support**	**Programmability**
Zip File of XDI	Good	Excellent	Poor	Very Good
XML, JSON, etc	Good	Excellent	Poor	Excellent
CIF	Good	Very Good	Poor	Poor
Sqlite3	Poor (binary)	Excellent	Poor	Very Good
HDF5	Poor (binary)	Very Good	Excellent	Very Good

These are all general-purpose and require specific *structure* and *dictionary of terms* (*ontology* or *schema*) for XAS data. **We need to map XDI to these formats**.

Zip file of XDI is the default – if we do nothing, this will be common.

The Problem with this:

Zip file of raw text files from beamline is too easy to mistake for **Zip file of XDI**.

The formality of XDI and its arrays labeled **i0**, **itrans**, **ifluor** enforces a *minimal data processing* (identifying the actual XAS data!) needed to communicate XAS well.

We should not expect users of public XAS data to know how to do "sum of dead-time-corrected channels from beamline XXX in 2023"

HDF5 and NeXuS

The 2012 paper on XAS Data Formats (B. Ravel, et al, 2012) recommends HDF5 for more complex datasets. HDF5 (Hierarchical Data Format version 5):

- widely used at synchrotrons and in other scientific fields for large datasets.
- efficient at storing large numerical datasets (compressed).
- well-supported for many programming languages.
- uses a simple and familiar hierarchy (filesystem-like), with Groups (directories) storing Datasets (files) with array or other data or other Groups.

Many facilities are favoring on HDF5, including for FAIR data portals.

HDF5 does not specify a Schema to assign meaning to Group and Datasets.

NeXuS is a "community-led" effort to define, support, and validate, schema for HDF5 for scientific data (synchrotron, neutron facilties). Been around 20+ years.

Schemas should build on existing NeXuS conventions, but can be proposed and/or modified and then "accepted": there is a an advisory committee, but they need input from "domain scientists" too.

We have a proposed schema (GitHub Pull Request), and need XAS Community comments. https://github.com/nexusformat/definitions/pull/1293

NXxas: XAS in NeXuS

A layout for XAS in NeXuS format closely mimics the XDI fields. Each HDF5 Group for an XAS Spectrum in a NeXuS file would look like (slightly truncated for space):

Address	**Meaning**
definition	`nxXAS`
data/edge, element	strings for element symbol, edge
data/energy	→ instrument/mono/energy
data/i0, itrans, ifluor, irefer	→ instrument/*/data
data/mutrans, mufluor, murefer	$\mu(E)$ for datatype
data/rawdata	→ scan/data
data/column_labels	→ scan/column_labels
data/mode	measurement mode ('Transmission')
sample/name	string name of sample
sample/prep	string description of sample prep
scan/start_time	date and time of scan
scan/edge_energy	nominal edge energy
scan/data	2D (nCol x nP) raw scan data table
scan/column_labels	array of column labels for scan/data
scan/scan_mode	string describing scan mode
instrument/mono/energy, angle	Array of energy values
instrument/mono/chemical_formula	string for mono crystal (eg, 'Si')
instrument/mono/d_spacing	d-spacing (in Ang) for reflection
instrument/mono/reflection	string crystal reflection (eg, '1,1,1')
instrument/source/beamline_name	string name of beamline
instrument/source/facility_name	string name of facility
instrument/source/probe	string for source probe ('X-ray')
instrument/i0, itrans, ifluor, irefer	Groups for detectors, with data

Follows XDI where possible.

The full raw data table is included, to give access to all collected data.

Metadata:
Each dataset and group can have keyword/value Attributes, or other datasets can be added.

more info, example files at https://tinyurl.com/nxxas2023

NXxas: discussion points

Using NeXuS/HDF5 format :

- easy translation from/to XDI plain text files.
- can have other data (XRD, XES emission spectra, . . .) *in the same data file*.
- uses format and conventions of other synchrotron methods
- Any strict format simplifies downstream use by novice users.

But there are real challenges to adopting XDI

- Any strictly enforced format requires deliberate use of that format, to avoid "Zip file of raw beamline datafiles".
- XDI has not been adopted universally (https://mdr.nims.go.jp/)

The work done to date on data formatting:

- Oyanagi formed a Working Group (WG) on data formatting.
- WG discusses formatting and members write papers.
- WG creates dictionaries of terms, support libraries, and tested examples.
- NeXuS format definition is proposed and support code, examples created.
- Translation tools from raw beamline to XDI or NXxas need to be maintained.

does not ensure that these formats are actually used.

Conclusion: NeXuS/HDF5 for XAS XDI

A NeXuS definition for storing XAS data in HDF5 has been proposed.

- This layout is based on XDI and supports multiple XAS (and other HDF5 data sets) in a single file.
- A large data table for "as collected" data is included along with the usual XAS arrays for standard processing.
- This format can support many variations of XAS data (including HERFD) and could also accomodate XMCD and X-ray Raman with small additions.
- It could be expanded to 2D data for full-field XAS and RIXS plane measurements.
- May be expandable to include analysis parameters and results.
- A strictly formatted file for an XAS spectrum will simplify use, but means that data must be translated to (and maybe out of) this format.

A Working Group (2 to 5 people) can develop, support, and maintain definitions like XDI or NXxas. But they cannot make you use it.

Adoption for use in supplemental material, data libraries and facility data portals requires engagement and common priorities and goals from a larger community.

Comments, discussion, suggestions welcome.