Progress Report: Data Format Working Group

Matthew Newville

Center for Advanced Radiation Sources,
The University of Chicago

Q2XAFS 2024: 2024-July-26

Active Working Group: Wout De Nolf, Marius Retegan, Mauro Rovezzi, Hitoshi Abe, Abhijeet Guar, Shelly Kelly, Gerry Seidler, Edmund Welter.

Thanks to: Bruce Ravel, Armando Sole, James Hester, Pieter Glatzel, Jan-Dierk Grunwaldt, Benjamin Watts, Sebastian Paripsa, Emiliano Fonda, Diego Gianolio, Giannantonio Cibin, Mark Wolfman, Sonal Patel, Masao Kimura, Takahiro Matsumoto, Masashi Ishii, many others.

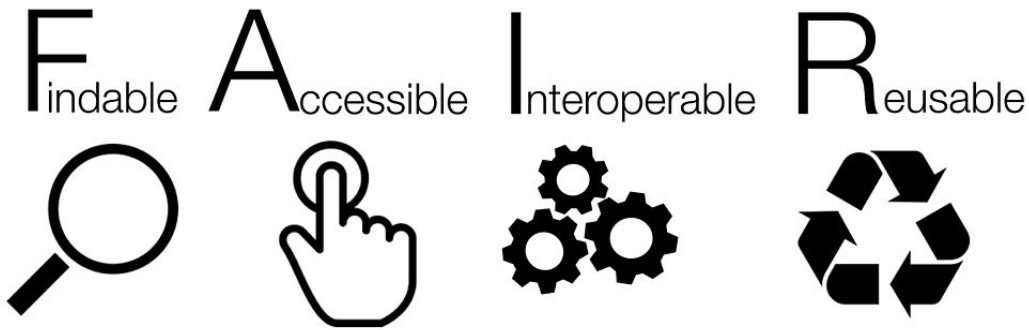
These slides, example data, more links and information:
https://tinyurl.com/nxxas2024

Reminder of 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.



How can we best 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 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.

These recommendations are still the two preferred options.

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

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**

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 are many optional `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.

XDI: Strengths and Weaknesses

XDI represents 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.
- make data more easily digestable (more non-experts, machine-learning) detectors and dead-time-correcting arrays.
- include "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.

HDF5 and NeXuS for multi-spectra files

There are many possibilities for data containers that could handle multiple spectra. Even in 2012 (B. Ravel, et al, 2012), HDF5 was recommended for more complex datasets.

HDF5 (Hierarchical Data Format version 5):

- widely used at synchrotrons and in other scientific fields for large (10 to ::100 GB) 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.

HDF5 is not without some criticisms:

- binary format that can be read only with its own library.
- not great at multiple processors reading/writing.
- files can become corrupted and unrecoverable.

But, at this time, HDF5 is the only one is really worth considering for experimental data with large arrays

HDF5 gives structure. But not meaning.

NeXuS: Adding meaning to HDF5 structure

NeXuS tries to adds meaning by creating layouts or schema for different categories of data. These are used at many synchrotrons (+ neutron, muon sources), especially for scattering/diffraction.

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

Many facilities are mandating (or thinking about it) NeXuS for FAIR data portals.

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

NeXus

Keys goal of our working group:

- Identify **key data and metadata** for communicating XAS.
- Refine the NeXuS XAS (NXxas) definition.
- Map NXxas to/from XDI.
- Develop examples and translation tools into NXxas and XDI.

What are the key **data** and metadata for XAS?

XDI specifies many optional types of data: `itrans`, `mufluor`, ... and can support multiple of these, and a reference channel. It is OK, but not great a multi-element fluorescence data.

NeXuS really wants a single main dataset: "intensity" ("what to plot?"), but also allows lots of auxiliary data.

What is "The XAS Data" we want to share with people a decade from now, with non-experts reading Supplemental Information, and with machine-learning algorithms?

Answer: **pre-edge subtracted, edge-step normalized $\mu(E)$. (yes, E)**

- Requires some "light processing" of raw' data. *Do I use* (I_f/I_0) *or* $-\log(I_t/I_0)$*?*
- Easily described: subtract this line, divde by this value
- Easily un-done or re-done.
- encouraged: some fluorescence channels, do deadtime corrections.

Note: For data with multiple modes (say, HERFD and transmission) or data that includes a reference spectra, each would be separate datasets, but NeXuS supports that easily, and allows for "links" instead of copies.

What are the key data and **metadata** for XAS?

XDI specifies many optional pieces of "metadata" and a simple way to organize it. NeXuS allows very rich, hierarchical metadata.

((. . . imagine lots of discussions and rehashing of ideas of what **could** be done.))

What metadata is really needed to be formal and machine-readable?

Required and Strongly Encouraged Metadata:

- Element Symbol and Element Edge
- name of sample. Something more than "sample 3" would be nice..
- name or abbreviation of laboratory, facility, and/or beamline used.
- date of data collection.
- name of person uploading or collecting this data.
- measurement mode (transmission, fluorescence, HERFD, . . . , Theory, . . .)
- *d*-spacing used (preferred) or crystal cut (at least) for monochromator used to select X-ray energy, so that energies can be recalibrated with high accuracy.

Probably only Element Symbol and Element Edge are absolutely required.

It's great to tell us the coating on the harmonic rejection mirror. Some of us care about that. This can be in human-readable text. Structured is good.

What are the key data and metadata for XAS?

So, we want NeXuS and XDI to easily share:

Data

- 2 arrays: Energy, Intensity for pre-edge-subtracted, normalized $\mu(E)$.
- Allow for multi-dimensional raw data (Nenergy x Ncolumns) for those who might want to reprocess the original data.
- include a description of how pre-edge subtraction and normalization were done.

MetaData

- Element Symbol
- Element Edge
- name of sample. We can hope for more.
- name or abbreviation of laboratory, facility, beamline used.
- date of data collection.
- name (ORCID?) of person uploading or collecting this data, or Experiment ID.
- measurement mode
- mono *d*-spacing or crystal cut.
- a big dictionary or mapping of other useful things to know.

Note:t this can support synchrotron Data, laboratory data, and theoretical XAS spec-

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`
element	string for element symbol [Fe, Pt]
absorption_edge	string for absorption edge [K, L3, M5]
mode	measurement mode ('Transmission')
energy	energy array
intensity	normalized μ array
reference	name of / link to other NXxas group
title	user-supplied title
rawdata	N-dimensional data
rawdata_labels	labels for columns of rawdata
process	text of processing steps
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
instrument/mono/energy, angle	Array of energy values
instrument/mono/chemical_formula	string for mono crystal (eg, 'Si')
instrument/mono/crystal/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

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.

pynxxas: NXxas support code

Many beamlines (including mine) save data into something "XDI-like". Some people already save to HDF5. Great!

No beamline or facility is saving this "lightly processed to normalized $\mu(E)$" data. We all need tools to help do this.

We (Wout de Nouf) we have a start of code to help:

- convert between NXxas and "new XDI" (matching NXxas).
- convert existing data files into either format.

`pynxxas`, https://github.com/XraySpectroscopy/pynxxas

We are willing to help write converters for your beamline, and we would love help making these tools.

Demo of NeXuS / XDI examples

Databases: current status, hope for improvements

The Working Group has not done much on On-line Dataases.

There are a few public on-line XAS databases. Most of these are limited to a spectra from one or a few facilities.

https://mdr.nims.go.jp/catalog : >2000 Spectra from Japanese beamlines. DOI for each spectrum. Not easy to navigate. Data is not easy to use.

https://xaslib.xrayabsorption.edu only 250+spectra, could be improved.

https://xasdb.lightsource.ca/ Pretty nice!

https://www.sshade.eu/db/fame/ A lot of good data!

Most of these aim to provide *curated* XAS data on well-known Standards.

There are also web portals like DATA.ESRF.FR for *un-curated* experimental data.

NXxas: discussion points

We have a workable definition for NXxas and XDI for sharing XAS data widely.

- share pre-edge subtracted, normalized $\mu(E)$. This does place a burden of "light data processing" on data producers or uploaders. Raw data can be included but is not sufficient.
- a truly minimal required set strongly recommended metadata, allowing for lots more.
- these formats support many variations of XAS data like HERFD and calculations, and could be extended to XMCD and X-ray Raman with small additions.
- we all need translation tools, and this work has begun.

But there are real challenges:

How to encourage adoption?

Most on-line databases use XDI or something like it and have $\mu(E)$. But https://mdr.nims.go.jp/ does not.

Does the community want or need a centralized on-line database?