

Data Formats for XAS Data: progress toward better, larger, XAS Databases

Matthew Newville

Center for Advanced Radiation Sources,
The University of Chicago

DAPHNE4NDFI
KIT / DAPHNE4NDFI Meeting 2024-Feb-14

A report on an ad-hoc Working Group established at Q2XAFS 2023, Melbourne

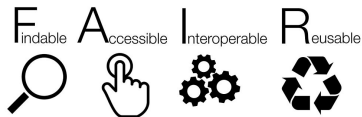
Motivation: Sharing XAS Data

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.

Machine learning methods need good data for training and validation.

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 share XAS spectra with other XAS practitioners, the wider scientific community, the facilities, and general public?

What is the XAS data do we want to share?

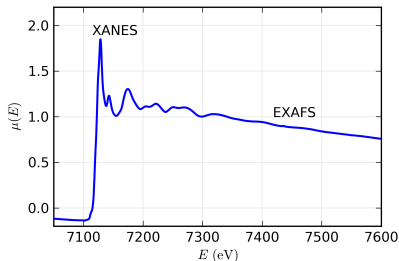
What needs to be included with the data to make it usable?

X-ray Absorption Spectroscopy: XAS, XAFS, EXAFS and XANES.

X-ray Absorption Spectroscopy (**XAS**) is the modulation of the X-ray absorption coefficient at energies at and above an X-ray absorption edge.

XAFS	X-ray Absorption Fine-Structure Spectroscopy (= XAS)
XANES	X-ray Absorption Near-Edge Spectroscopy
EXAFS	Extended X-ray Absorption Fine-Structure

These contain information about an element's chemical state (XANES) and local atomic environment (EXAFS).



Fe K-edge XAFS for FeO.

Main XAS Characteristics:

- local atomic coordination
- valence, oxidation state
- applies to any element ($Z > 2$) .
- works at low concentrations (ppm, μM)
- minimal sample requirements.
- many sample environments possible.
- independent of crystal structure, isotope.

Why XAS data is unique and valuable

XAS gives unique information about *Oxidation State* and *local atomic environment* around the selected element, even at low concentrations, and in complex environments.

Why XAS data is unique and valuable

XAS gives unique information about *Oxidation State* and *local atomic environment* around the selected element, even at low concentrations, and in complex environments.

XAS does require an *Energy tunable X-ray source*. This usually means using a **synchrotron beamline**.

- limited number of beamlines
- all custom instruments, with similar but not identical components.
- custom data acquisition systems and raw data streams (maybe shared per facility).

There is not a standard format for collected “raw” data from these instruments.

Why XAS data is unique and valuable

XAS gives unique information about *Oxidation State* and *local atomic environment* around the selected element, even at low concentrations, and in complex environments.

XAS does require an *Energy tunable X-ray source*. This usually means using a **synchrotron beamline**.

- limited number of beamlines
- all custom instruments, with similar but not identical components.
- custom data acquisition systems and raw data streams (maybe shared per facility).

There is not a standard format for collected “raw” data from these instruments.

There are sufficient data processing and analysis tools for

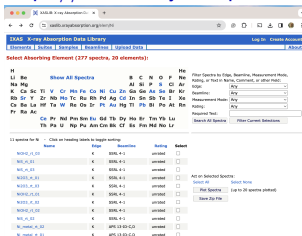
- processing “raw data” from a beamline to a *Standard XAS Spectrum*.
- visualization and analysis of a *Standard XAS Spectrum*.

But we do need that *Standard XAS Spectrum* - preferably in a standard format - for everyone to be able to use.

Motivation: Sharing XAS Data – current status

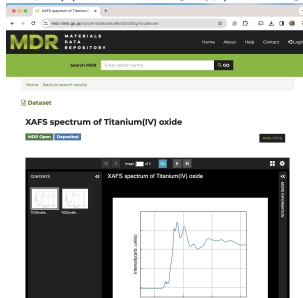
There are a few public on-line XAS databases, and several beamline-specific ones. Most of these aim to provide *curated* XAS data on well-known Standards.

<https://xaslib.xrayabsorption.edu>



not very many spectra (250+).
Could be improved.

<https://mdr.nims.go.jp/catalog>

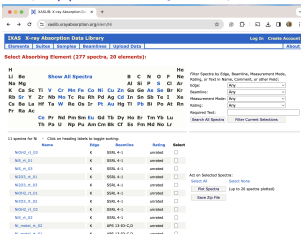


>2000 Spectra from Japanese beamlines
More than XAS.
DOI for each spectrum.
Not easy to navigate.

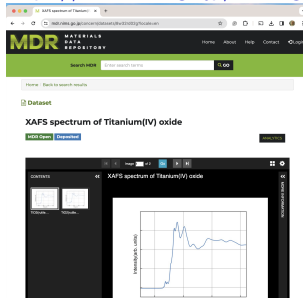
Motivation: Sharing XAS Data – current status

There are a few public on-line XAS databases, and several beamline-specific ones. Most of these aim to provide *curated* XAS data on well-known Standards.

<https://xaslib.xrayabsorption.edu>



<https://mdr.nims.go.jp/catalog>



not very many spectra (250+).
Could be improved.

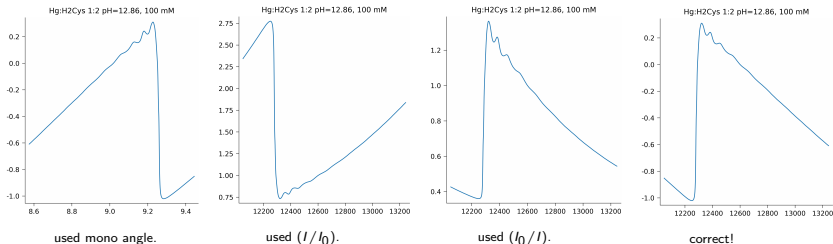
>2000 Spectra from Japanese beamlines
More than XAS.
DOI for each spectrum.
Not easy to navigate.

Also: web portals like DATA.ESRF.FR that share *un-curated* experimental data.

Also: calculated XANES data from Materials Project (... that is not uniformly great).

Motivation: Sharing XAS Data

Are the data in these databases meaningful? Can people (or their machines) read and use these data?

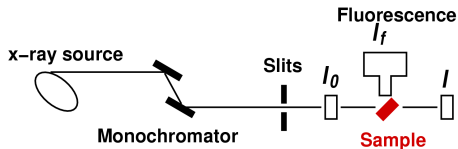


How the data are formatted and described matters, especially as the consumer of the data becomes less experienced, which could include:

- machine-learning algorithms.
- automated data analysis workflows now applied to a data set from a new beamline.
- the interested, capable scientist new to XAS.

We need a common data format of a *Standard XAS Spectrum*

X-ray Absorption Measurements



$\mu(E)$ can be measured two ways:

Transmission measure what is transmitted through the sample:

$$I = I_0 e^{-\mu(E)t}$$

Fluorescence measure fluorescent x-rays from the re-filling the core hole:

$$\mu(E) \propto I_f / I_0$$

Energy usually comes from a monochromator that works by Bragg diffraction

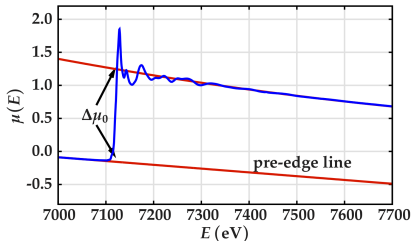
$$\frac{hc}{E} = \lambda = 2d \sin(\theta)$$

where d is monochromator lattice spacing, and θ is angle.

Data Reduction: Pre-Edge Subtraction, Normalization

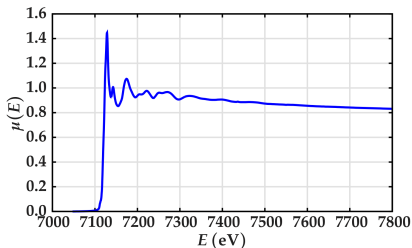
We measure $\mu(E)$ that often has a strong, smooth background.

We care about how big the *edge jump* is, but usually want to normalize spectra to compare them.



Pre-Edge Subtraction

We subtract away the background that fits the *pre-edge* region. This gets rid of the absorption due to other edges (say, the Fe L_{III} edge).



Normalization

We estimate the *edge step*, $\Delta\mu_0(E_0)$ by extrapolating a simple fit to $\mu(E)$ to the edge. We divide by this value to get the absorption from 1 x-ray.

Raw Data to a Standard XAS Spectrum

The collected, “raw” data includes signals from a few (maybe 100) channels from several different detectors.

The resulting *XAS spectrum* may be a slightly complex combination of these, perhaps:

converting collected data columns to XAS spectra:

```
energy = Col.02
intensity = -log(Col.04 / Col.03)
# or
intensity = (Col.05 * Col.09 + Col.06 * Col.10 + Col.07 * Col.11 + Col.08 * Col.12)/Col.03
```

Some beamlines even want that conversion from monochromator angle to Energy in these data reduction steps.

We may also want to do that pre-edge subtraction and normalization (especially for machine-learning applications).

- 1 to subtract a line
- 2 divide by an edge jump

These are simple, easily reproduced, easily documented steps, but we often want to share data after these steps.

XAS spectra need to be *lightly processed* to be shared.

XAS Metadata Working Groups: past and present

Data formats for XAS data were discussed at Q2XAFS2011, Tsukuba, Japan.

A small Working Group (B. Ravel, J. Hester, V. A. Sole, G. Wellenruether, M. Newville) was formed to discuss and recommend formats for XAFS:

B. Ravel, et al, *J. Sync Rad* **19**, p869–874 (2012)

This discusses a few options and proposes and defines a plaintext (ASCII) format:

XDI = XAFS Data Interchange for a single XAS spectrum.

XAS Metadata Working Groups: past and present

Data formats for XAS data were discussed at Q2XAFS2011, Tsukuba, Japan.

A small Working Group (B. Ravel, J. Hester, V. A. Sole, G. Wellenruether, M. Newville) was formed to discuss and recommend formats for XAFS:

B. Ravel, et al, *J. Sync Rad* **19**, p869–874 (2012)

This discusses a few options and proposes and defines a plaintext (ASCII) format:

XDI = XAFS Data Interchange for a single XAS spectrum.

At Q2XAFS 2023, Melbourne, in August 2023, a Working Group was established and has been meeting ~monthly. Participants include:

Europe Mauro Rovezzi, Wout De Nolf, Marius Retegan, Benjamin Watts, Benedikt Eggert, Abhijeet Gaur, Edmund Welter, Diego Gianoli, Giannantonio Cibin, Sofia Diaz Moreno, Sebastian Paripsa, Emiliano Fonda, Gautier Landrot, Nicolas Trcera

Asia/Australia Kiyotaka Asakure, Masako Kimura, Takahiro Matsumoto, Masashi Ishii, Hitoshi Abe, James Hester, Chris Chantler, Chanh Tran.

Americas Bruce Ravel, Shelly Kelly, Mark Wolfman, Jerry Seidler, John Rehr, Morgan Desmau, M Newville.

Goals: a) revisit XDI and what metadata is needed to communicate XAS+ data.
b) guidance on multi-spectra formats, especially HDF5/NeXuS.

What is it we want to share anyway?

There are a few distinct desires for databases and data formats

- provide XAS data on standard reference compounds for analysis, quality assurance, testing theoretical and analysis methods.
- provide XAS data on wide ranges of materials, for machine learning methods.
- provide XAS to make it more generally useful to someone else.

And, there are many variations of XAS we may want to share

- fluorescence XAS needing corrections (deadtime, over-absorption)
- HERFD XAS, needing resolution information
- XMCD, needing polarization description
- time-resolved XAS
- dispersive XAS
- laboratory XAS

What is it we want to share anyway?

There are a few distinct desires for databases and data formats

- provide XAS data on standard reference compounds for analysis, quality assurance, testing theoretical and analysis methods.
- provide XAS data on wide ranges of materials, for machine learning methods.
- provide XAS to make it more generally useful to someone else.

And, there are many variations of XAS we may want to share

- fluorescence XAS needing corrections (deadtime, over-absorption)
- HERFD XAS, needing resolution information
- XMCD, needing polarization description
- time-resolved XAS
- dispersive XAS
- laboratory XAS

What is it we want to share anyway?

There are a few distinct desires for databases and data formats

- provide XAS data on standard reference compounds for analysis, quality assurance, testing theoretical and analysis methods.
- provide XAS data on wide ranges of materials, for machine learning methods.
- provide XAS to make it more generally useful to someone else.

And, there are many variations of XAS we may want to share

- fluorescence XAS needing corrections (deadtime, over-absorption)
- HERFD XAS, needing resolution information
- XMCD, needing polarization description
- time-resolved XAS
- dispersive XAS
- laboratory XAS

Again, the general conclusion from these working groups has been

Share $\mu(E)$ that is “lightly processed and corrected”

The existing XAFS Data Interchange (XDI) Format

The 2012 Working group published 2 basic recommendations in

B. Ravel, et al, *J. Synch Rad* **19**, p869–874 (2012)

- 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...

Using CIF for storing XAS data was not really pursued.

The existing XAFS Data Interchange (XDI) Format

The 2012 Working group published 2 basic recommendations in
B. Ravel, et al, *J. Sync Rad* **19**, p869–874 (2012)

- 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...

Using CIF for storing XAS data was not really pursued.

- 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.

Using a “real” relational databases (say, Postgres) is attractive for metadata and data on-line databases, but is not practical for sharing data with individuals.

The idea was to get XDI done for holding a single spectrum, then go from there.

next: Review XDI, then onto HDF5 / NeXuS

The XDI Format

The design (from Bruce Ravel) for XDI presented at Q2XAFS2011 were refined, implemented (C, Python, Perl) 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
mu	mu, general	arbitrary
intensity	mu, general	arbitrary
i0	monitor intensity	arbitrary
itrans	transmission intensity	arbitrary
ifluor	fluorescence intensity	arbitrary
irefer	reference intensity	arbitrary
mutrans	mu transmission	$-\log(\text{itrans}/i0)$
mufluor	mu fluorescence	$\text{ifluor}/i0$
murefer	mu reference	unspecified

Some array labels for processed data are also defined:

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

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
mu	mu, general	arbitrary
intensity	mu, general	arbitrary
i0	monitor intensity	arbitrary
itrans	transmission intensity	arbitrary
ifluor	fluorescence intensity	arbitrary
irefer	reference intensity	arbitrary
mutrans	mu transmission	$-\log(\text{itrans}/i0)$
mufluor	mu fluorescence	$\text{ifluor}/i0$
murefer	mu reference	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.

Some array labels for processed data are also defined:

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

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

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**

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:

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 <i>d</i> in Å.

and a handful of **recommended metadata**

Column Labels for Arrays

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

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:

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 <i>d</i> in Å.

and a handful of **recommended metadata**

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.

Recommendation: Use Minimal XDI

What do we really need for communicating XAS data?

Name	Meaning	Data Type	Importance
Energy	X-ray energy, in eV, maybe keV	Array	required
Mu or Intensity	absorbance, unitless	Array	required
I0	incident flux, unitless	Array	recommended
Element.symbol	symbol of absorbing element	String	required
Element.edge	symbol absorbing edge	String	required
Mono.d_spacing	d-spacing for mono	Number	recommended
Mono.name	name and crystal cut of mono	String	recommended

Recommendation: Use Minimal XDI

What do we really need for communicating XAS data?

Name	Meaning	Data Type	Importance
Energy	X-ray energy, in eV, maybe keV	Array	required
Mu or Intensity	absorbance, unitless	Array	required
I0	incident flux, unitless	Array	recommended
Element.symbol	symbol of absorbing element	String	required
Element.edge	symbol absorbing edge	String	required
Mono.d_spacing	d-spacing for mono	Number	recommended
Mono.name	name and crystal cut of mono	String	recommended

Some notes and requests for creating data to share:

- convert arrays of **angle** to **energy** please.
- include either **Mono.d_spacing** (preferred) or **Mono.name**. Allows precise recalibration.
- neither **mu** nor **i0** imply any units or scale. (dark currents subtracted, please!).
- **mutrans**, **mufluor**, **itrans**, **ifluor**, **irefer** ... are acceptable. (correct deadtimes please!)
- include name and/or description of sample.
- include name of beamline or laboratory.
- include name of person involved in measurement.
- include date of data collection.

Anything else (temperature, ring current, harmonic rejection mirror) is nice to have', but not required. These might be used in downstream analysis as categories, but not as numbers.

Multi-spectra files in On-line Databases

Recap: XDI defines a single XAS spectrum in plain text, with clearly defined syntax, and has support code. Files in this format will be usable in 50 years or more.

Multi-spectra files in On-line Databases

Recap: XDI defines a single XAS spectrum in plain text, with clearly defined syntax, and has support code. Files in this format will be usable in 50 years or more.

For on-line databases of XAS spectra, we recommend (Working Group consensus):

- use any relational or non-relational database for storing the data.
- provide “export data to individual XDI files” for output when possible.
- verifying and curating data is very challenging. It is OK to provide imperfect data. Maybe allow this to be noted or commented.
- Energy calibration with a reference foil is preferred, but this could be a separate file, maybe measured simultaneously, but maybe not – some monochromators are very very stable these days.
- curating sample preparation, characterization, and environment are important, but we do not know how to make this machine-readable expect “tagged metadata using strings”.

Current on-line databases are organized by nation.

Maybe this could be improved?

Multi-spectra files: data portals and repositories

On-line databases of curated data are very different from data for an experiment. Still, have consistent formats would be helpful.

For supplemental materials for journals and FAIR data portals, we may want to share:

- multiple spectra, perhaps many hundreds of spectra.
- “more raw” data like individual arrays from multi-element detectors and dead-time-correction arrays.
- non-XAS data as metadata, such as
 - ▶ XES emission scans.
 - ▶ elastic energy scan for HERFD/RIXS analyzers.
 - ▶ XRD patterns.
 - ▶ XRF spectra or maps.
- theoretical inputs, data processing parameters, intermediate results.

XDI is OK for a single spectrum, but we need something more for all of these options.

Getting something that will also be “useful for 50 years” is challenging.

Multi-spectra files possibilities

Some existing possibilities, with some 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.

Multi-spectra files possibilities

Some existing possibilities, with some 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 **energy**, **mutrans**, **i0**, **itrans**, enforces *minimal data processing* and identifies the actual XAS data. This is necessary to communicate XAS well (across decades, continents, expertise).

We cannot expect users of public XAS data to know how to sum dead-time-corrected channels from beamline XXX in 2017.

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.
- very efficient at storing large numerical datasets (compressed).
- less ideal than other formats at storing strings or keyword-value dictionaries, but acceptable.
- binary, but well-supported for many programming languages.
- uses a simple and familiar hierarchy, like a filesystem), 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.

Public Service Announcement:

HDF5 files can become corrupted and all data in them lost. Filesystems and relational databases work very hard to avoid corruption. HDF5 does not.

This is not hard to do. I speak from direct experience.

“Read-only” is not always enough.

Use of HDF5 files for “big data” is not going away.

If you like your data, make a backup of your HDF5 data.

NeXuS is a “community-led” effort to define, support, and validate, schema for HDF5 for scientific data (synchrotron, neutron facilities). It has 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. This process seems odd to some people:

why is a self-appointed group of neutron scatterers assert authority to approve how X-ray spectroscopists communicate with one another?

But they all mean well, do know what they are doing, and are happy to help and guide,

We are working on a proposed schema (and GitHub Pull Request), and need XAS Community comments.

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	
absorbing_element	element symbol	Follows XDI where possible.
edge	element edge	
mode	measurement mode ('Transmission')	→ means "link to other dataset".
energy	X, → instrument/mono/energy	
intensity	Y, $\mu(E)$, lightly processed	
irefer	intensity (processed) for reference	
i0, itrans, ifluor, irefer	→ instrument/*/data	The full raw data table is included, to give access to all collected data if desired.
sample/name	string name of sample	
sample/...	description of sample prep, etc	
raw_data	Group containing raw data table	
raw_data/data	2D (nCol × nP) raw scan data table	
raw_data/column_labels	array of column labels for scan/data	Highly structured metadata.
process/	Group for text/code of processing steps	
scan/	Group of data collection parameters	
instrument/mono/energy, angle	Array of energy values	Each dataset and group can also have keyword/-value Attributes
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	Detector Groups have a data array, and maybe lots of metadata.

more info, example files at <https://tinyurl.com/nxxas2023>, and work in progress.

NXxas: discussion points and conclusions

Advantages of 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 used by other synchrotron methods
- Any strict format simplifies downstream use by novice users.

But there are real challenges with 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 does not ensure that these formats are actually used.

- Hiroyuki Oyanagi appointed 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 being proposed and support code, examples created.
- Translation tools from raw beamline to XDI or NXxas need to be maintained.

... hopefully we can do this ...

NXxas: discussion points and conclusions

Advantages of 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 used by other synchrotron methods
- Any strict format simplifies downstream use by novice users.

But there are real challenges with 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 does not ensure that these formats are actually used.

- Hiroyuki Oyanagi appointed 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 being proposed and support code, examples created.
- Translation tools from raw beamline to XDI or NXxas need to be maintained.

... hopefully we can do this ...

Conclusion: XDI and NeXuS/HDF5 for XAS

- ① XDI gives a good way to describe a single XAS spectrum.
A strictly formatted file for an XAS spectrum will simplify use, but means that data must be translated to this format.
- ② NeXuS can be used to store multiple datasets (XAS and others) in a single HDF5 file with (very) formal definitions.
- ③ An NXxas definition is being refined to
 - ▶ mimic and translate to/from XDI
 - ▶ include raw data table for FAIR data sharing of data from X-ray beamlines.
 - ▶ include metadata for, or be extended for related X-ray spectroscopies (HERFD, XES, XMCD, X-ray Raman, 2-D XAS, RIXS...)
 - ▶ include processing parameters / scripts as text.

A Working Group has been formed to discuss and refine many of these concepts and details.

I think that we are ready to have a working NeXuS definition for XAS that can be used for Supplemental Information in Journals and for FAIR Data Portals

Adoption and support needs to come from a broad community.

Comments, discussion, suggestions welcome.