ANSTO comments on NeXus Quo Vadis

Good paper Mark. Thanks for putting down your thoughts and giving us a framework for discussion.

Summary of my comments.

1. The NIAC are custodians of standards, rather than those that define the standards. Move focus to reduced data.
2. The NIAC continues to build tools. Adopts a component model.
3. Replace the NAPI with H5NX
4. Use an existing object oriented data model for HDF e.g. UCAR's common data model
5. Validation. Create xml schema to validate HDF files for instrument types. This should alleviate the need for a dictionary

Nick's opinions

Let's differentiate between 3 file types, using Mark's definitions.

1. Full beamline description (FBD) – this is an archival format which is a 'complete' data set produced by the instrument.
   Users: Data Acquisition Developers, Catalogue Developers.
2. Application Definitions – the minimal dataset required to do data reduction. Optional.
   Users: Data Reduction Developers.
3. Reduced – data that has been corrected for instrument idiosyncrasies.
   Users: Scientists

A standard **reduced data format** is of most value to the scientific community. This is the interchange format that allows instrumentally corrected scientific data to be shared.

A facility should provide a reduced data file to a scientist. It is the facility's responsibility to take whatever archival format they have; database, NeXus or binary and convert this to a reduced dataset. The majority of scientists provided with a reduced data file won't look at the Instrument Definition or Archival file.

NIAC as custodians

Great idea.

Many years ago each instrument definition had an editor, whose responsibility it was to edit the application definition in response to technique community requirements. Giving a community a prescription for data formats has not worked. The NIAC being custodians of standards, rather than those that define the standards should be trialled.

canSAS is a good example of positive interaction between a member of the NIAC and the technique community. The community recognised the benefit of a standard reduced data format and create it. Pete Jemian, a member of the NIAC, has supported them in these efforts. It would be great if we can replicate this success.

The NIAC exists to provide support to scientific communities on the software engineering aspects of reduced data format standards.

Remarketing NeXus

I suggest we remarket NeXus. nexusformat.org becomes a site which targets information for scientists. **Only reduced data format standards and tools** with minimal complexity. For SAS for example, it may only provide links to the canSAS site.

A new site, nexusdevelopers.org supports developers (data acquisition, catalogue and data reduction developers). The content would be most of the current nexusformat.org information.

Business as usual

NeXus FBD definitions are of most value to

1. data acquisition developers (producers)
2. catalogue (e.g. iCAT/Tardis) developers (consumers)
3. data reduction developers (e.g. Mantid) (consumers)

who are professional software developers. The value to the scientific community is indirect. The standard helps these developer groups to be more productive through reduced design effort and software component reuse. The NAPI and the Common Data Model are examples of this. Scientists should see improved reliability and reduced costs.

The NIAC exists to provide support to software developers with FBD data format standards and software tools.

Where we haven't seen value is in the reuse of data reduction code.

ANSTO survey

Q: Do ANSTO scientists use the napi, FBD or application definitions?

A: NO (with a few exceptions e.g. mapping NeXus datasets to LAMP datasets – but this process is done by hand and the NeXus definition don't add value)

Q: Do ANSTO data acquisition developers use the napi, FDB or application definitions?

A: YES (not the application definitions).

Q: Has it made data acquisition developers more productive?

A: NO. It's about the same as a do-it-yourself data format. The effort to learn and implement the napi and NeXus definitions does not reduce the time to create the format for a new instrument.

Q: Does the FBD enable scientists to use third party applications to read their data files.

A: NO. (The LAMP case could have been done with any HDF5 internal file structure)

Q: Does introducing a dictionary mapping the ANSTO internal file structure to program variables enable scientists to use other applications to read their data files.

A: YES, but only for applications into which the dictionary can be compiled. No solution at this stage for IGOR, IDL etc.

Other users of FBD

1. Instrument Simulation (e.g. MCSTAS, Vitesse)
2. Self-configuring instrument control applications (e.g. GDA)
3. Data reduction applications (e.g. Mantid)
4. Cataloguing (e.g. iCAT)

Once an instrument is modelled using NeXus base classes, the schema can be used to generate models for simulation, instrument control and data files. There is a lot of effort required to make this work. Once it does work, you can imagine that you only have to change the model when you build or make changes to an instrument. No other coding is required. Regarding object oriented NeXus; this may be good reason to add inheritance to the base class definitions. We already have encapsulation. I don't think we want polymorphism.

NeXus validation service

Either FBD or reduced data files are validated, along with the applications that read them.

If you can validate a NeXus file, then you don't need a dictionary.

The reason that NXdict and the CDMA dictionary exist, was due to the ambiguity of the NeXus standard and the absence of tools to validate it. The process proposed was to translate HDF to XML, then validated against a schema. The java application nxvalidate is worthless without an application definition schema.

NAPI

Let's deprecate the NAPI. Facilities use HDF5 and databases to store archival data, not XML. Performance is critical.

Can the NIAC help facilities with database schema and tools? I think the answer is yes. Even if this is just a forum for information exchange.

It can help with HDF5 schema and tools.

I conditionally support replacing the NAPI with H5NX. The condition is performance tests on the H5NX API.

On the subject of performance, the NIAC should define performance benchmarks for the API, and have a test system to verify performance of API versions.

The NIAC has an obligation to XML users. We should survey how many instruments are using the NAPI to write XML and engage directly with them to determine a migration path.

If the reason detre for the NIAC is support and custodianship rather than definition of standards, we can then

OO NeXus

There already exist 2 object oriented APIs for NeXus file IO. These are netCDF using the common data model (CDM) http://www.unidata.ucar.edu/software/netcdf-java/CDM/ , and CDMA, which is an extension of the CDM model, adding a dictionary.