Workshop on HDF5 as hyperspectral data exchange and analysis format.
Grenoble, January 11th to January 13th, 2010.

Abstract

A workshop on HDF5(1) as a file format for hyper-spectral data exchange and analysis has taken place at the ESRF from January 11th to January 13th. This event was financed by the FP7-ELISA funded networking activity called VEDAC(2). The workshop enjoyed a very high participation in terms of both, quality and number of participants. The interest on the subject led to vivid discussions in an enthusiastic and cooperative atmosphere. The suitability of HDF5 and of NeXus(3) as an exchange format was clearly manifested by the fact files supplied by different participants were accessible by all of them in their favorite programming language or platform. Different acquisition and analysis approaches were discussed and working groups have been setup in order to propose complementary solutions either based on particular analysis techniques or more general based on building relational databases. There was general agreement to use the HDF5 compatible NeXus data format structure wherever possible, adding extensions when needed.
(1) http://www.hdfgroup.org/HDF5
(2) http://vedac.esrf.eu
(3) http://www.nexusformat.org

1. SESSIONS

1.1 Monday 11th

Monday sessions took place at the ESRF auditorium and were also followed by part of the ESRF staff.

Francesc Alted gave a nice introduction to the HDF5 capabilities to a very interested audience. There were numerous questions about HDF5 asked and answered during the talk

Mark Rivers presented part of his EPICS based acquisition system to handle N-dimensional detector data. The higher layers should be usable without need of an EPICS system. Informal discussions took place about implementations details between Mark and Laurent Claustre, head of the ESRF beamline control unit (formerly BLISS).

Mark Koennecke made a short introduction to NeXus and presented the concept of application definitions. These definitions are keyword fixed dictionaries specifying where to find the information to perform a particular analysis. This feature can answer many analysis problems but most of the audience was unaware of it because the NeXus web site seems to be quite behind the developments.

Alain Buteau and **Majid Ounsy** presented an approach looking to access the data as being organized into a database. The proposal produced divided feelings, not because of being or not being a good proposal, but because of seeming too ambitious. The ultimate idea is certainly appealing. It was already answering some of the issues appearing later during the workshop. SOLEIL is encouraged to continue exploring that path. Mark Koennecke said he had some code to extract information that might be used to index the files.

Darren Dale presented a measurement and analysis based approach to store the information in contrast to the NeXus instrument based approach. He underlined the very different situation between new facilities and established ones. The later need to have an evolution path and the proposed approach allows straightforward conversion from the SPEC file format to HDF5 while allowing NeXus compatibility if desired. In the discussions, it was recommended to the NIAC to

make sure that groups not being of any of the NXwhatever types/classes were ignored by the NeXus API and not considered in the file validation tests. According to the members of the NIAC present, that seems to be already the case.

Herbert Bernstein stressed in his talk the convenience of database approaches. He provided an example of embedding image CIF into a NeXus structure. A comment during the practical session of Tuesday was that it would seem more convenient to have the structure into a dedicated HDF5 group instead of having NXentry as the direct parent. Herbert said that was not a problem for him. He just waits for the NIAC recommendation about where to put it.

After the coffee break we had interventions from **Chris Jacobsen** and **Danielle Nuzillard**. They presented several examples of applications of multivariate analysis to very different problems.

Paul Dumas presented the needs of the infrared analysis community. By one hand they are bound to proprietary formats associated to their instruments and by other hand software able to treat the data and handle those formats is expensive. He said PyMca was open source and able to read one of the major formats but still far from their needs.

Gerd Wellenreuther presented his experience with non-negative matrix approximation (NNMA). The technique is quite helpful to interpret fluorescence maps.

Informal discussions and conversations continued during the Wine and Cheese session at the ESRF cantine.

1.2 Tuesday, 12th

Burkhard Kaulich presented the capabilities of the Elettra TwinMic station. He underlined their current ASCII based file format was not adapted to their needs.

Matt Newville presented some ideas about exchanging data for the EXAFS community. His talk was very rich in ideas. If data are just to exchanged, a simple ASCII file should be enough. However, if one needs to reprocess the data, more detailed information is needed because the measured data are not necessarily the exchanged data and full reprocessing can be impossible. If HDF5 can be used as a support for that information or as a support for a library of XAS spectra, then the situation is different. A relational database would certainly help to deal with all that information. A way to specify transformations of measured quantities into expected ones would also be needed.

Stefan Vogt presented X-ray fluorescence mapping and the clustering capabilities of MAPS.

Armando Solé made a demo about current HDF5 support in PyMca to illustrate the need for analysis programs to know more about the data they were dealing with. A short presentation followed.

Burkhard Kaulich presented results from data analyzed with aXis2000 and PyMca. Despite using totally different approaches (experimental references and theoretical spectra), both codes were giving the same results. The TwinMic station would like to have a simple way to use and to compare results obtained with both codes.

Workshop participants were offered a USB key with several HDF5 files containing experimental data. During the practical session people were exploring them with different codes. Nobody seemed to have real problems to access the files despite coming from different sources. This last point

should be retained as very encouraging.

Pierre Bleuet stressed the user need to be able to have everything together and to be able to handle large amounts of data. He showed examples of simultaneous fluorescence and diffraction tomography.

Pete Jemian, Peter Bösecke and **Javier Pérez** presented SAXS experimental stations and data approaches used at APS, ESRF and SOLEIL.

1.3 Wednesday, 13th

The session prior to the coffee break was divided into two different working groups. One based on Small Angle Scattering (SAS) and the other with the rest of participants. The second group tried to get a consensus on a-priori simple issues: a) how to exchange already treated data and b) to try to define an application definition for X-ray fluorescence analysis.

The structure provided by the NXdata field can already deal with common problems like units, labels and, while not being strictly necessary, can avoid many discussions when exchanging data. The TwinMic beamline is using PyMca and aXis2000 and could be a good test candidate.

It became clear that adding an attribute to the data to specify how had to be interpreted was needed. The main reason being that programs had to know if their were dealing with scalars, spectra, images, … Darren Dale made clear the most versatile way to store data was having one dimension more than the intrinsic dimension of the data. For instance, an array of images of r rows and c columns, should be stored as (n, r, c) While there was agreement on being the most versatile approach, each facility was already having its own way to store data depending on being acquired on regular grids or not. The adoption of the previously mentioned attribute to specify the type of data (scalars, spectra, images, ...) should solve part of the issue. Therefore, an additional problem was to know if those dimensions were the first or the last of the dataset. Herbert Berstein suggested to take a look at img_CIF and to propose a way to specify how all the dimensions were changing (fastest, slowest, ...). It was agreed that in the mean time, and in absence of other specification, C order would be assumed. Therefore, an dataset reported by HDF5 as being (dim0, dim1, dim2) and being tagged as an image, would correspond to an array of dim0 images of dim1 rows and dim2 columns.

Trying to provide an application definition proved not to be something to be solved in one hour. At the end we came to the conclusion we needed simple definitions that could be expanded. That feature seems to be available. Matt Newville manifested that one could measure something needed in an indirect way, therefore needing a way to specify transformations. Herbert Berstein seemed to have something related to it close to operational. Issue to be followed.

There were several **recommendations to the NIAC representative**s:

- Add an “intrinsic dimension” tagging attribute to the data generated by the detectors. It should be possible to use the same attribute on any dataset to let the programs know how to deal with them. Present NIAC representatives agreed.
- The NeXus subversion source repository should provide unrestricted read access to everybody.
- Add a version attribute to the definition of NXDL. Changed immediately added.
- Update Web site! Most of the presents knew nothing about the NeXus application definitions.

By the other hand, the SAS group reviewed and revised the NXsas definition. And decided to incorporate definitions of CIF into NeXus. NIAC to define where. Img_CIF will populate equivalent Nexus fields.

2. CONCLUSIONS

- We are free to use any version of HDF5. Whenever a higher version API reads a lower version data file, it should be possible to keep the lower version, no higher version features being used. We request the HDF-group to keep the maximum value of the version somewhere easily accessible in the file. Herbert provided the following information obtained from HDF5:

  "Per design in HDF5, version numbers are attached to each object in a very fine granulation, but not to the entire file format. So objects of different version numbers can be mixed in the same file.

  There is a set_version() call in the HDF5 API that limits the variability of versions that can be used with an HDF5 file. This functionality is already there, not sure if the command line utility support repack supports it yet. Would probably be easy to add such an option. "

  There was comment suggesting to allow saving/converting/downgrading an HDF5 file to a particular version. Something similar to what is commonly made in the text editors world. That would be very convenient, although we do not know if it is possible to implement.

- Sharing of data should be easily achieved by using the NXdata group. It will be tested at Elettra with TwinMic, aXis2000 and PyMca.

- While there is not endorsement of NeXus, wherever appropriate we will be using the NeXus definitions / conventions. No HDF4, no XML will be used except for the requirements to validate (no XML for data).

- Single measurements should be reproduced in a single NXentry. Processed results can go into the same NXentry or in a different one.

- We will not heavily rely on the use of links in the NX-part in order not to prevent the mapping of a NeXus-file to a relational database.

- Keep an eye on SOLEIL database oriented developments while working on technique specific application definitions.

- There is no possibility to define optional parameters in the NeXus application definitions. However, there is the possibility to derive application definitions from more basic ones thus providing the equivalent functionality. There can be more than one application definition in each entry.

- Attribute to tag data intrinsic dimension needed. Name to be defined.

- Application definitions should have a mandatory version attribute. This change was immediately adopted.

- Existing definitions of geometry in NXsas are not sufficient for SAS:

- Incorporate geometry definitions of CIF
- Additions to the detector description to identify the beam center
- Adopt at as much as possible from img_CIF and sasCIF (img_CIF: http://www-bernstein-plus-sons/software/CBF/doc/cif_img_1.5.4.html)
    - Q: Does NeXus time stamp descriptor allow for description of subseconds?
    - A: The NeXus time stamp descriptor (NX_DATE_TIME) uses ISO8601 which allows for seconds described by 1 or more digits of decimal fraction (e.g. http://www.w3.org/TR/NOTE-datetime)

- Establish the mapping of CIF terms into NeXus, highest-priority is img_CIF