Argonne NATIONAL LABORATORY

NeXus ? Comments from some heretics

Chris Jacobsen
Associate Division Director, X-ray Science Division, Advanced Photon Source
Professor, Physics & Astronomy, Northwestern University; Applied Physics; Chemistry of Life Processes Institute

My perspective:

- **Development of multiple x-ray microscopes, including coding of netCDF and HDF5 data storage**
- **Development of several analysis programs, including handling of data from electron and infrared microscopes**

Argonne NATIONAL LABORATORY

Advanced Photon Source

NORTHWESTERN UNIVERSITY

U.S. DEPARTMENT OF ENERGY

spectromicroscopy - Data analysis tool for spectromicroscopy - Google Project Hosting

http://code.google.com/p/spectromicroscopy/

Google

iGoogle NY Times Mobile Bills Chicago Argonne Northwestern Travel/Scuba Computer Science Stony Brook Personal

chris.j.jacobsen@gmail.com | My favorites | Profile | Sign out

spectromicroscopy

Data analysis tool for spectromicroscopy

Search projects

Project Home | Downloads | Wiki | Issues | Source

Summary | Updates | People

Project Information

Starred by 0 users

Activity: Medium

Project feeds

Code license

GNU GPL v3

Labels

spectromicroscopy, microscopy, xray1, Python

Members

mirna.lerotic

1 committer

Your role

Committer

Featured

Downloads

Mantis_1.05.zip

Show all »

Spectromicroscopy

Spectromicroscopy combines spectral data with microscopy, where typical datasets consist of a stack of microscopic images taken across an energy range. Due to the data complexity, manual analysis can be time consuming and inefficient, whereas multivariate analysis tools not only reduce the time needed but also can uncover hidden trends in the data.

Mantis

MANTiS is Multivariate ANalysis Tool for Spectromicroscopy developed in Python. It uses principal component analysis and cluster analysis to classify pixels according to spectral similarity.

http://tinyurl.com/3mhjvov

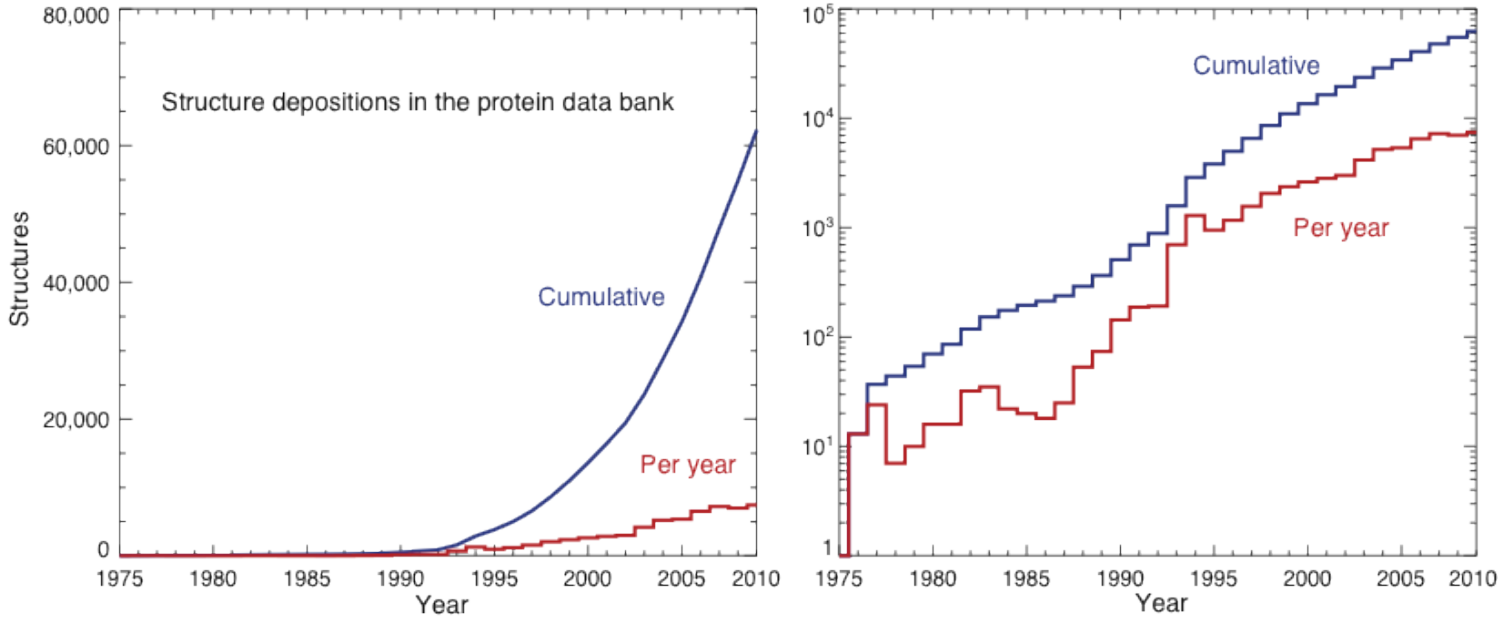
NeXus ? Comments from some heretics

There is one God

What a common file format provides:

- Users can read data from multiple experiments and facilities with ease.
- One analysis program can be used to read - and daresay even combine - data from different measurements of one sample.
- Programming is greatly simplified.
- The many benefit from the efforts of a few.

CCP4 in crystallography: 1994



There is *should be* one God...

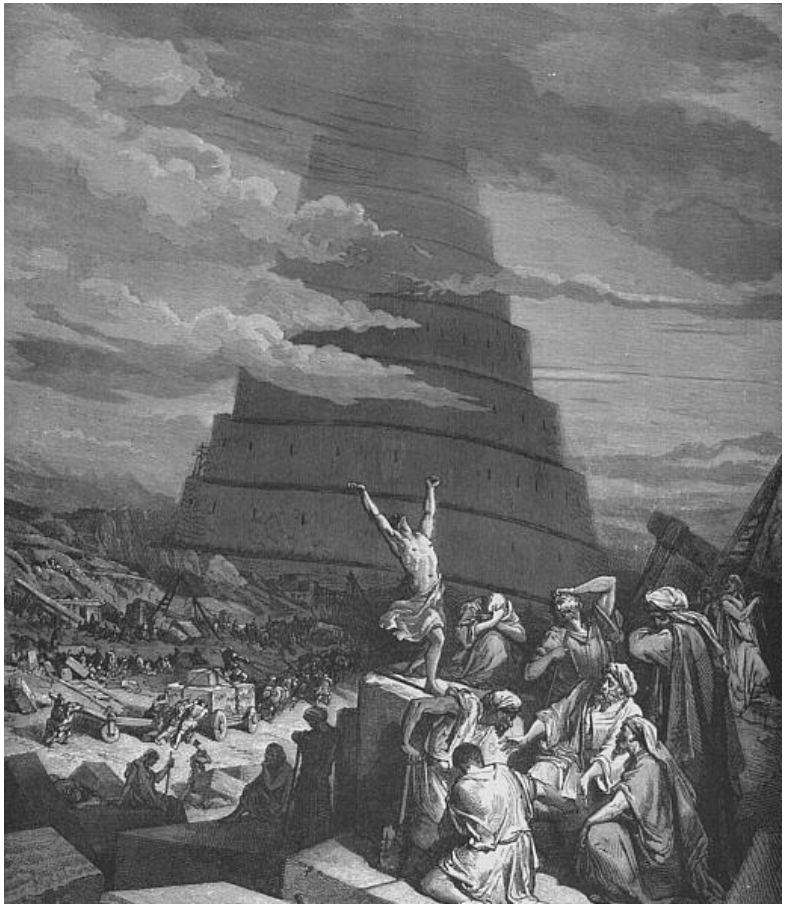
And His name is Yaweh...

And He is Triune...

And His prophet is Allah (peace be upon him)...

And she is disappointed in us...

- We have a plethora of homegrown data formats.
- Nobody can read each other's data files.
- Some experiments still store their numerical data in ASCII text format!



The Confusion of Tongues by Gustave Doré (1865) - from Wikipedia

"*The LORD said, 'If as one people speaking the same language they have begun to do this, then nothing they plan to do will be impossible for them.' So the LORD scattered them from there over all the earth, and they stopped building the city. That is why it was called Babel - because there the LORD confused the language of the whole world.*" *Genesis 1, NIV*

Our Creed

- Data should be binary for efficiency, yet platform independent.
  - Must include compatibility for parallel processor access.
- But binary files should include text that describes their structure.
- Often there's one key array (*e.g.*, signal versus position versus energy), with related subsidiary arrays (position scale, energy scale). It should be really easy to read this key array!
- One should distinguish between minimal mandatory components and optional elements, so as to require as little as possible for ease of data exchange.
- It's good to provide agreed-upon definitions for additional metadata, but programs should be able to gracefully handle the absence of some or all metadata.
- There should be ways to store additional "local" metadata, which may be too user-specific for there to be global agreement on its structure.
- We should be lazy, and benefit from tools developed by the larger computing community.

HDF5 satisfies our creed

- Binary, platform-independent, but with text tags describing all structures. Parallel I/O support. Compression.
- Easy to examine even by non-programmers, using `h5dump` or `HDFView.`
- One can define a top-level group to hold the key data array.
- One can add further groups to hold additional metadata using agreed-upon conventions. One can search for defined tags and gracefully handle their absence.
- One can add "local" groups to hold additional metadata before "universal" definitions exist, such as `/32id_oct2011_stuff`
- One can add additional metadata such for tracking provenance.
- Support is already built-in to IDL, Matlab, Python – no muss, no fuss, no extra libraries to install!
- A large community is behind its ongoing development.

So we agree! There is one God, and his/her name is HDF5

- HDF5 was released in 1998.
- HDF5 subroutine libraries can read HDF4 and netCDF.
- HDF5, being binary, is far more efficient than storing large data sets in XML.

So why is NeXus still bothering to support HDF4 and XML?
Why accept limitations due to HDF4 and XML history?

- Lack of support for 1D array of strings
- Lack of support for HDF5 dimension scale
- Lack of support for data compression (90-95% in fluorescence microscopy)
- Complications in target attribute due to XML storage support

But we disagree... Blasphemy regarding NeXus

- NeXus is based on an application programming interface (NAPI), which requires debugging on many platforms and in many languages.
  - Why do that, when HDF5 has already done this hard work?
  - Why not instead describe files in terms of HDF5 subroutine call sequences, instead of layering an API on top of the HDF5 library?
- NeXus tries to solve *all* storage of *all* data, which makes it very difficult to find the key data array in a file.
  - I'm told that NAPI produces segmentation faults if it encounters non-NeXus elements. Shouldn't it simply ignore what it doesn't understand?
- NeXus is complicated; who wants to read through 100 pages of documentation to understand how to read or write one array in a file?
- Rules for NeXus compliance are not well documented; is `nxvalidate` a finished product yet?
- How much has NeXus spread beyond neutron facilties? NeXus is now 8+ years old.
  - Not used by many at APS, ALS, NSLS.
  - Adoption as an import/export option by commercial instruments?
  - Unknown among electron microscopists?

Heretical philosophy

- Simplicity=portability
- The less we are *required* to specify in a file, the better. But we should have common definitions of additional *optional* information!
- Often an experiment's key data is contained in one multidimensional array. Let's make that easy to find!
- Let's make data exchange accessible to electron microscopy, light microscopy,…

> "Nobody should start to undertake a large project. You start with a small trivial project, and you should never expect it to get large. If you do, you'll just overdesign and generally think it is more important than it likely is at that stage. Or worse, you might be scared away by the sheer size of the work you envision. So start small, and think about the details. Don't think about some big picture and fancy design. If it doesn't solve some fairly immediate need, it's almost certainly over-designed. And don't expect people to jump in and help you. That's not how these things work. You need to get something half-way useful first, and then others will say 'hey, that almost works for me,' and they'll get involved in the project." – Linus Torvalds

A new revelation from God

- The files are written by basic HDF5 calls, making it easy for anyone to either look at an example file using `h5dump` or `HDFView`, or to look at example code in language X, and then create their own read and write routines in language Y.
- We provide an information and exchange definition that provides *just enough* mandatory information to share a multidimensional data array as simply as possible.
- We provide additional, optional, but clearly defined metadata components to the base definition.
- We provide technique-specific groups for tomography, spectromicroscopy, and fluorescence mapping applications. These involve separate HDF5 groups (like separate subdirectories within the file).
- Anyone can then add their own beamline-specific groups into the file, which a "standard" program can choose to ignore.

- https://confluence.aps.anl.gov/display/NX/Data+Exchange+Basics
- http://tinyurl.com/4xyo72v

Can’t we all just get along?

- Data Exchange provides simplicity, and easier readability.
- NeXus provides one structure for recording all metadata.

A hybrid HDF5 approach?

- Data Exchange at the start of a HDF5 file.
- Optional NeXus groups after, with pointers back to what’s in Data Exchange.
- Optional non-Data Exchange, non-NeXus groups afterwards for local use.

“Never doubt that a small group of thoughtful, committed citizens can change the world. Indeed, it is the only thing that ever has.” – Margaret Mead