NeXus and Synchrotrons: Challenges and Requirements

V.A. Solé – ESRF Software Group

NeXus Data Format Workshop, PSI, May. 2010

This talk

Synchrotron needs

NeXus challenges

High Data Rate Experiments: ID15

- Experiment:
  - Absorption and phase contrast tomography
- Detector:
  - 2k x 2k pixel 16 bit CMOS PCO camera
- Camera supported data rate:
  - > 1000 full frame images/s – > 8 Gbytes/second
  - 10^6 binned images/s Potentially a full tomography in less than 1s
- Bottleneck:
  - Data transfer and storage (current limitation)
  - Data analysis (if previous issue solved)
- Data access:
  - Most users come with their own disks to take data with them
  - Data analysis often performed on site

ID17 & ID19: The Malapa hominid study

The Malapa hominid study

Multiscale analysis performed on ID17 and ID19 in February 2010

Around 10 TB of basic data (radiographs)

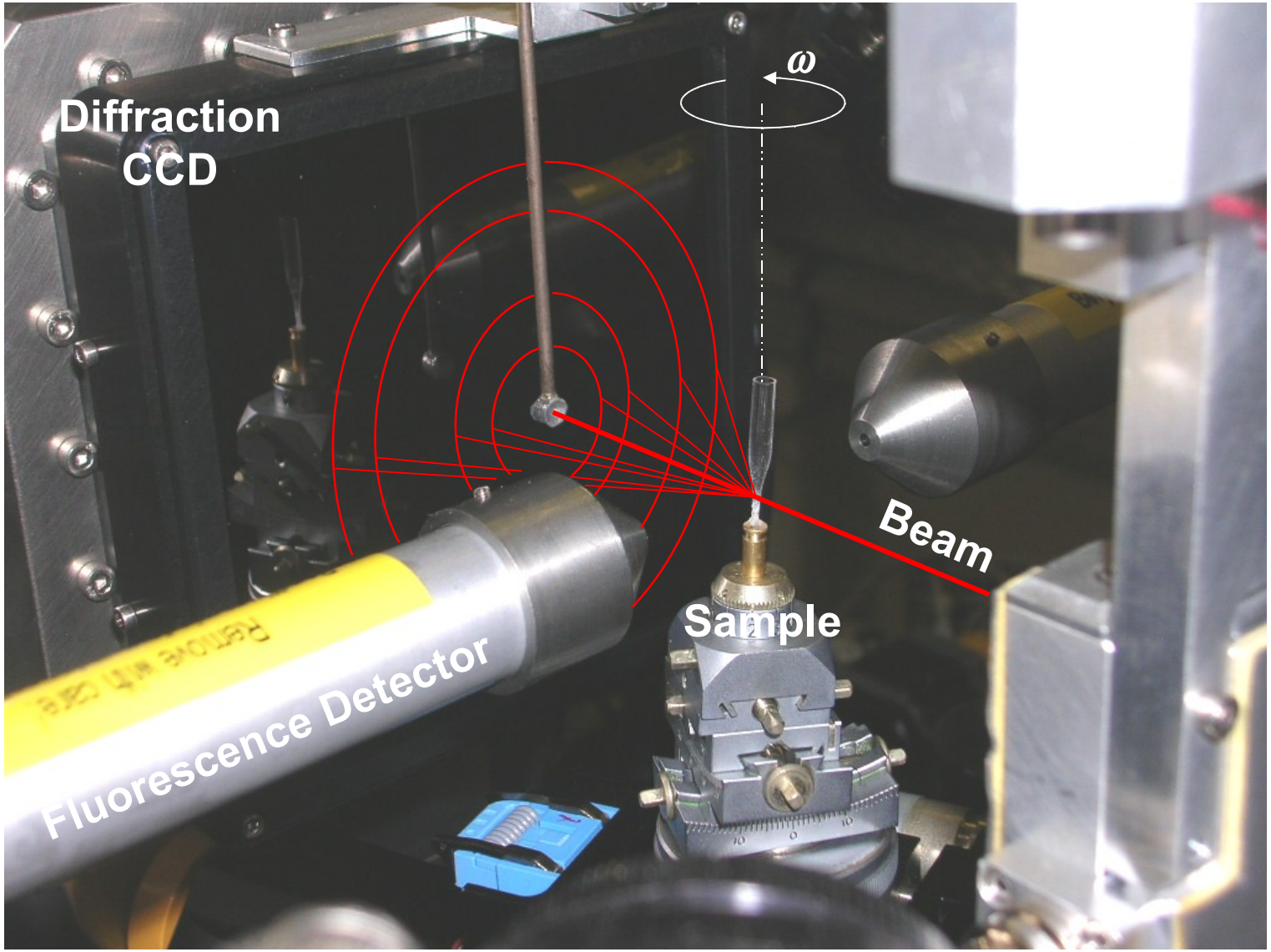
Around 15 additional terabytes of data after processing

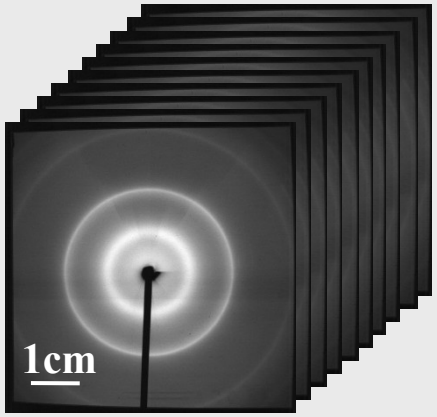
1 full month of calculation for reconstruction and artifact corrections with 100 processors, using at least 7 steps of processing.

Probably 1-2 years of analysis before having the first important results on dental development and internal anatomy

Acknowledgements : Paul Tafforeau, ESRF

ID22 – Fluorescence-Diffraction Tomography

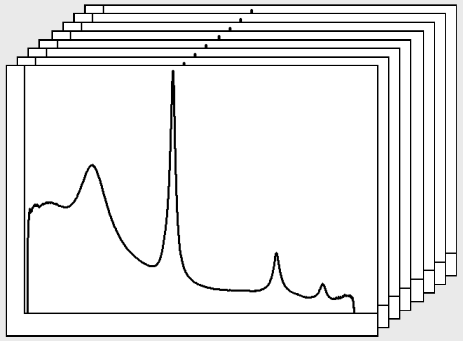




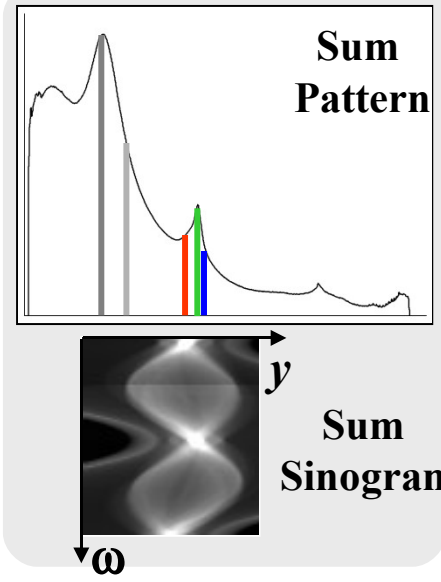
$N_y \times N_\omega$ Diffraction Images

Azimuthal Integrations

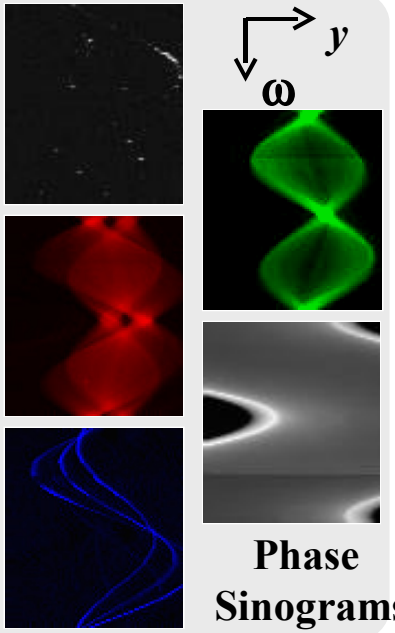
Fit2d software

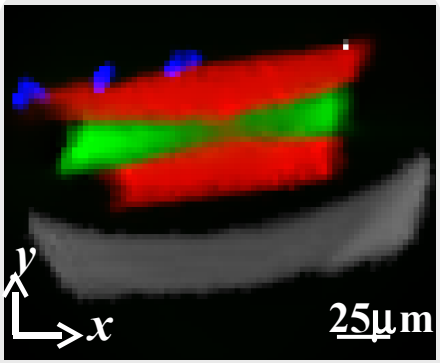


$N_y \times N_\omega$ Diffraction Patterns



PyMca software





Reconstruction

Capillary *Ferrite*
sp3
Calcite *Cubic*

Acknowledgements: Pierre Bleuet CEA - Grenoble

Data format issues

- Currently
  - Diffraction images in EDF or MarCCD format
  - Fluorescence data in EDF or SPEC file format
  - Scan information in SPEC file format
  - Result of azimutal integration on Fit2D .chi format
- Preparing to move to HDF5

Lesson learned:

Try to avoid inventing a new data format

Lesson NOT learned (yet?):

Forget about ASCII just because you want to look at the file

Synchrotron data format requirements

Efficient format to store different data types

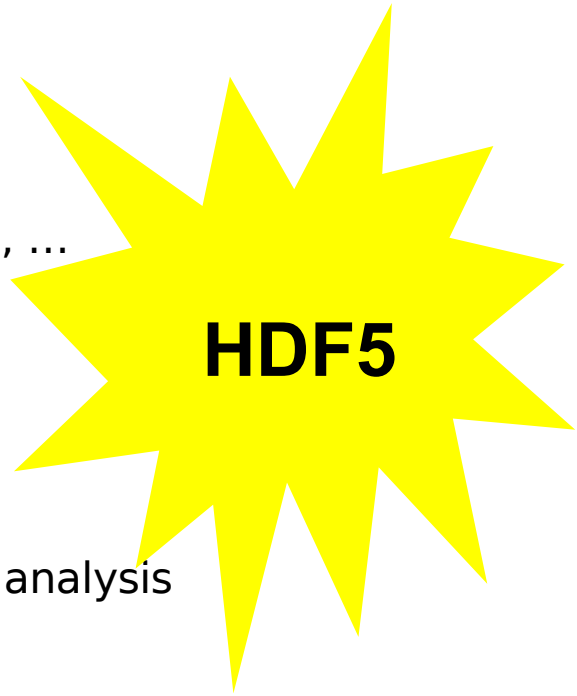
Keep together motors, counters, images, spectra, …

Compression support

Widespread support

Efficient and easy access to the data for visualization and analysis

Versatile enough to be a one-for-all format



Why HDF5?

It fulfills the previous requirements

It is supported on all common platforms and programming languages

Long term support seems assured

- Standard file format for storing data from NASA's Earth Observing System
- Petabytes of data stored in HDF5 (Global Climate Change Research Program)

Synchrotron needs in two words

Efficiency & Versatility

HDF5, what about NeXus?

What we like about NeXus

Well defined classes

Application definitions

A lot of endless discussions avoided

What we do not like about NeXus

A lot of endless discussions pending

Not many analysis applications supporting it

- At the ESRF only PyMca, others should follow.

No easy way to implement new needs

- Many synchrotrons are using NXdata as “data container” without setting signal and/or axis attributes and therefore not specifying a plot

- A new need should imply a new group

A beamline scientist friendly approach

NXroot

Top level. One per file.

NXentry

One group per measurement

Measurement

One group per measurement

Positioners

One group per Measurement
Ex. All motor positions when the command was issued.

ScalarData

One group per Measurement
Ex. Scanned motors and counters.

Spectrum

Several datasets per Measurement
Ex. 1 spectrum dataset per MCA device

ImageData

Several datasets per measurement
Ex. 1 image dataset per CCD device

Advantages

Simple to implement

Answers current scientists demands (keep measurement data together, compression, …)

Compatible with NeXus if desired (specific NeXus groups can be written at any time with links and the opposite is also true)

Can be seen/used as an intermediate step for not- yet-defined instruments or uses

Data Archival Challenge

The file should be able to provide EVERYTHING to perform the data analysis

If NeXus achieves that, then it is a must and the ultimate archival format

If it needs a complementary database, SOLEIL's approach (Database + plugins + file indexing) seems more appropriate (but it can be achieved with HDF5 instead of NeXus)

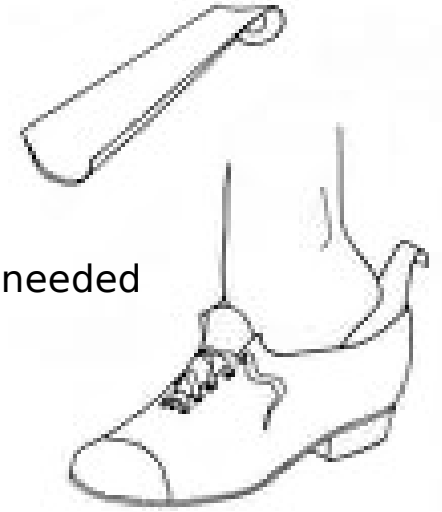
Data Analysis

Reduced and analyzed data sharing

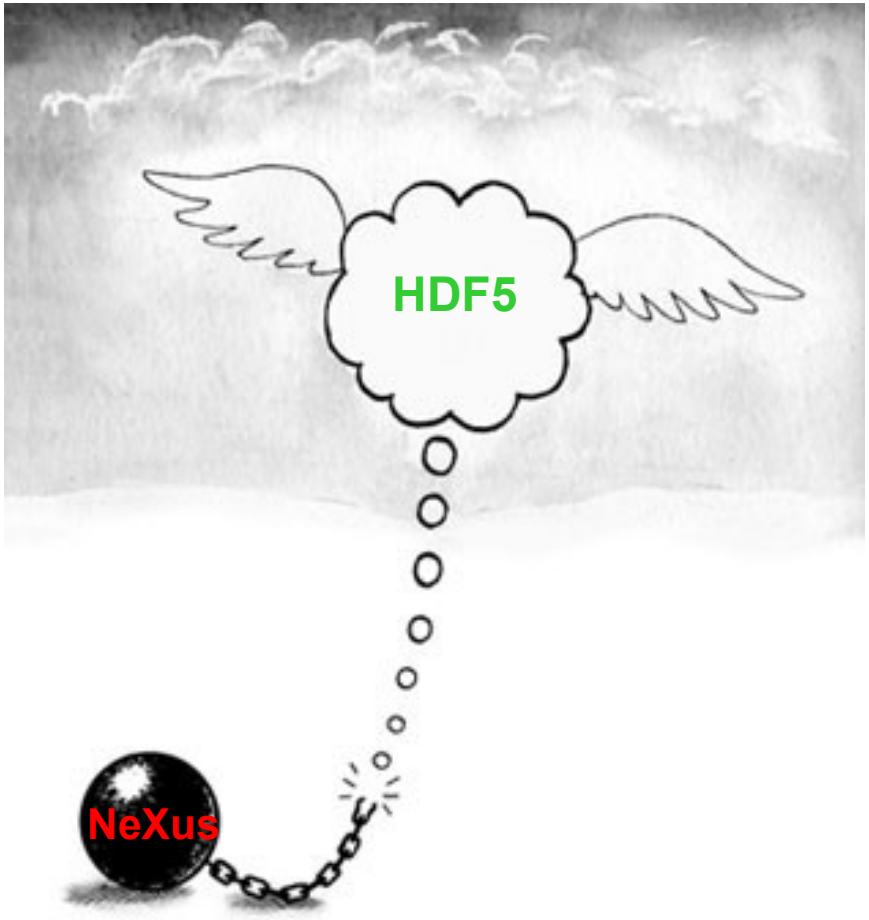
- NXdata seems very well suited to the job
  - Independent and dependent variables
  - Handling of units

Raw data analysis

- More NeXus application definitions needed
  - Should the actual applications provide them?
- Expressions to accommodate measured and requested data needed
  - Is NeXus going to do something in this way?
  - Should we join SOLEIL's plugin approach?



NeXus biggest challenge



Conclusion

HDF5 is a good data format choice

We expect to share data and analysis tools

- Conventions are certainly needed
- If they are not there, the analysis tools will impose them

Try to keep things as simple as possible

- NXdata provides the simplest form of data exchange
- NeXus application definitions are an excellent idea

Consider integration of Measurement group into NeXus

Validation tools are critical to validate non NeXus API produced files (MatLab, IDL, …)