HDF5 for synchrotrons

V.A. Solé – ESRF Software Group

ESRFUP Workshop, ESRF, Jan. 2011
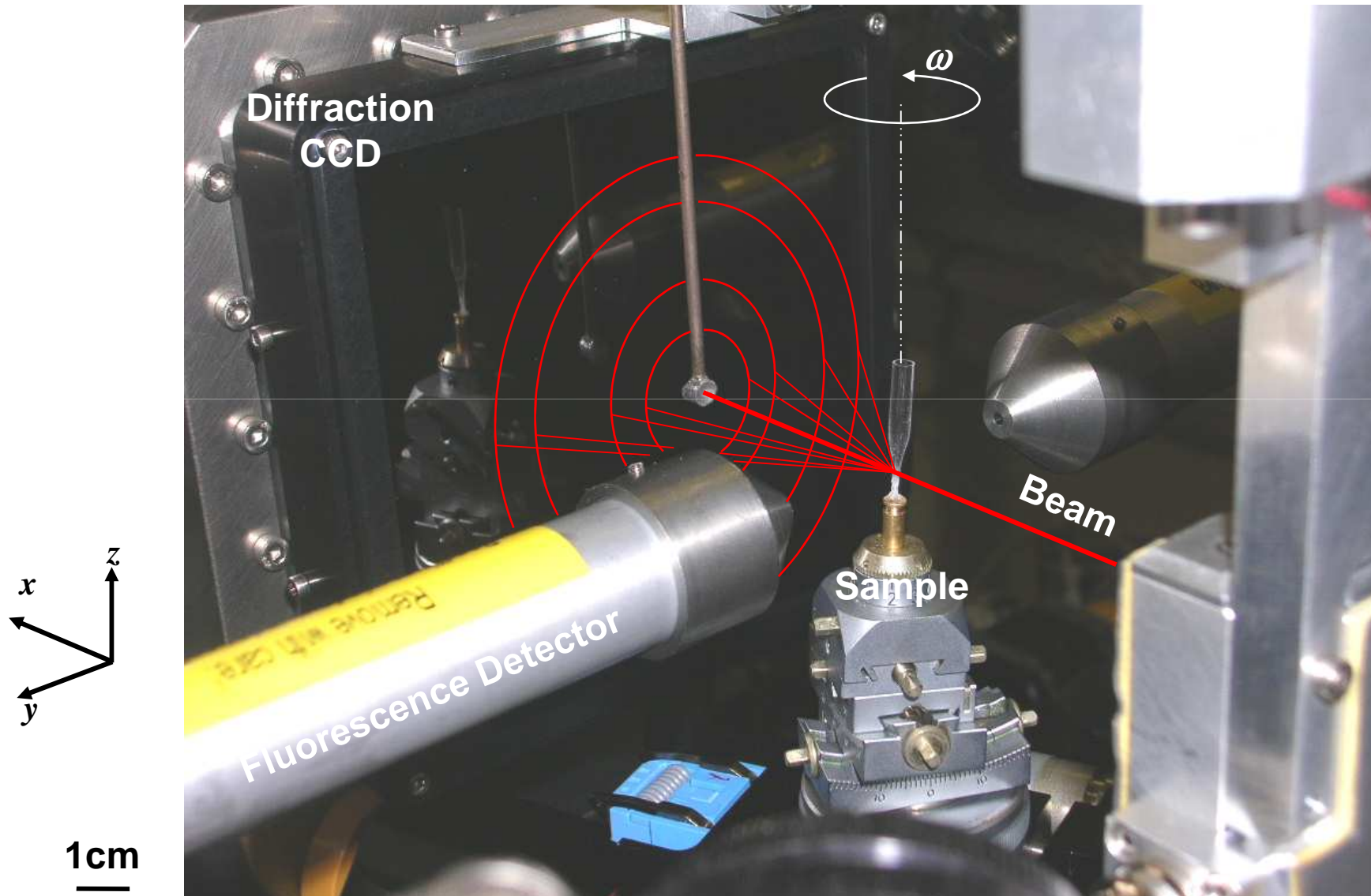
This talk

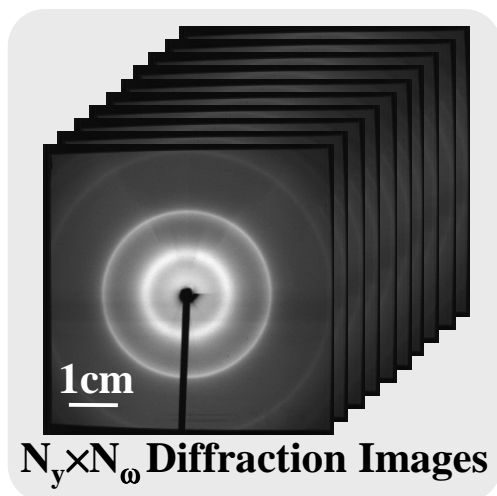Bla, bla, bla, …

HDF5 is great

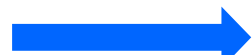Bla, bla, bla, …

Did I say loud enough how great HDF5 is?

Bla, bla, bla, …
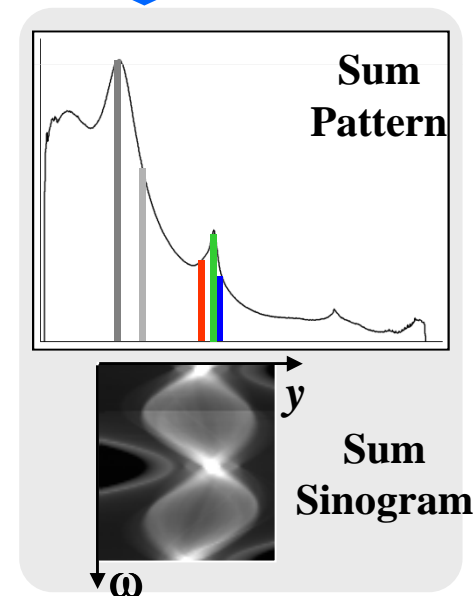
Conclusion

ID22 – Fluorescence-Diffraction Tomography

Azimuthal Integrations

Fit2d software

$N_y \times N_\omega$ **Diffraction Images**

$N_y \times N_\omega$ **Diffraction Patterns**

Sum Pattern

Sum Sinogram

PyMca software

Phase Sinograms

1cm

25μm

Reconstruction
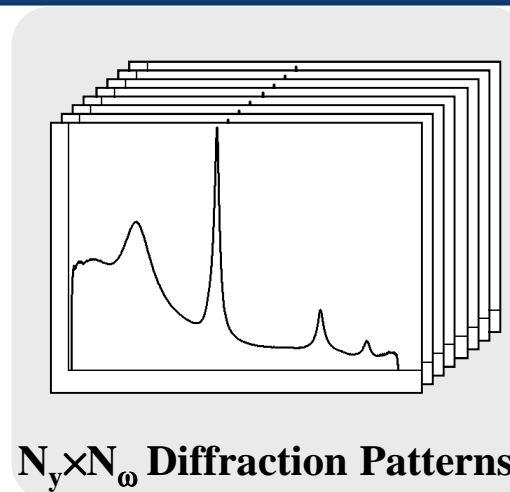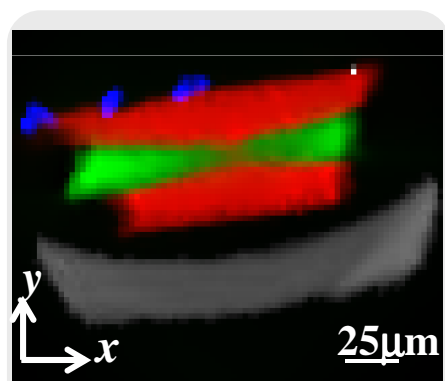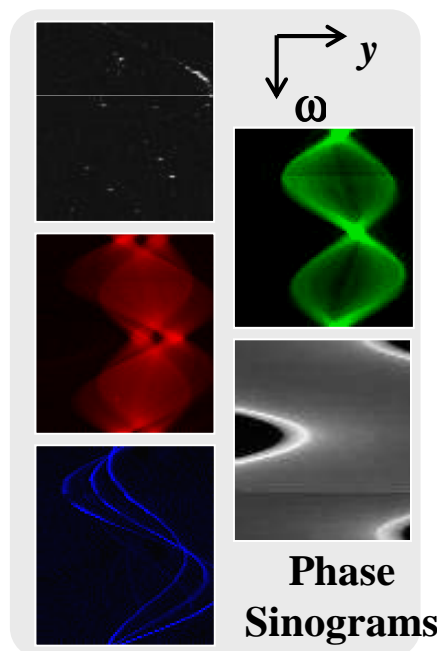
▭ *Capillary*	■ *Ferrite*	
▯ *Calcite*	■ *sp3*	
	■ *Cubic*	

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
　　　　Handle motors, counters, images, spectra, …

Versatile enough to be a one-for-all format

Compression support

Widespread support, not facility specific

Efficient and easy access to the data for visualization and analysis

Supported on all common platforms and programming languages

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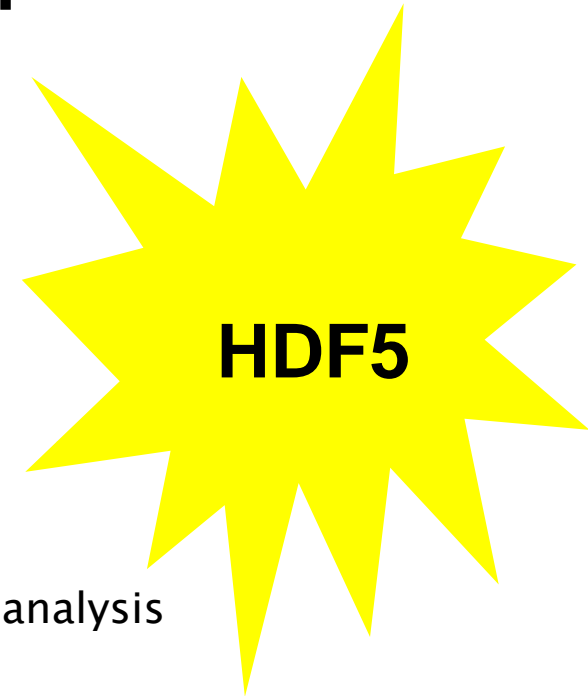
HDF5

Why HDF5?

The obvious answer

It fulfills all the previous requirements

The commercial answer

Manufacturers can understand "my format is HDF5" and use it

The political answer

It is recommended by our workshop sponsor

The HDF5 user and application developer answer

Why not? Are you suffering from the not-made-here syndrome?
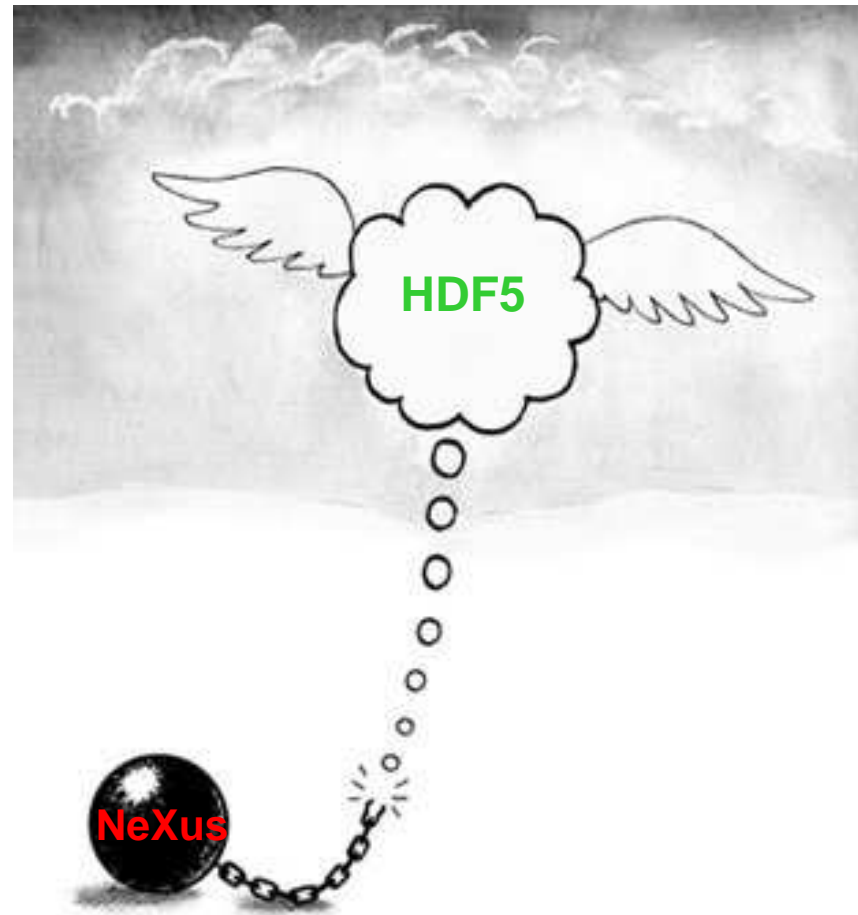
HDF5, not NeXus?

NeXus is NOT a data format

Take your favourite file system, you decide to write your images in a particular folder, and your other documents in other one, have you invented a new file system?
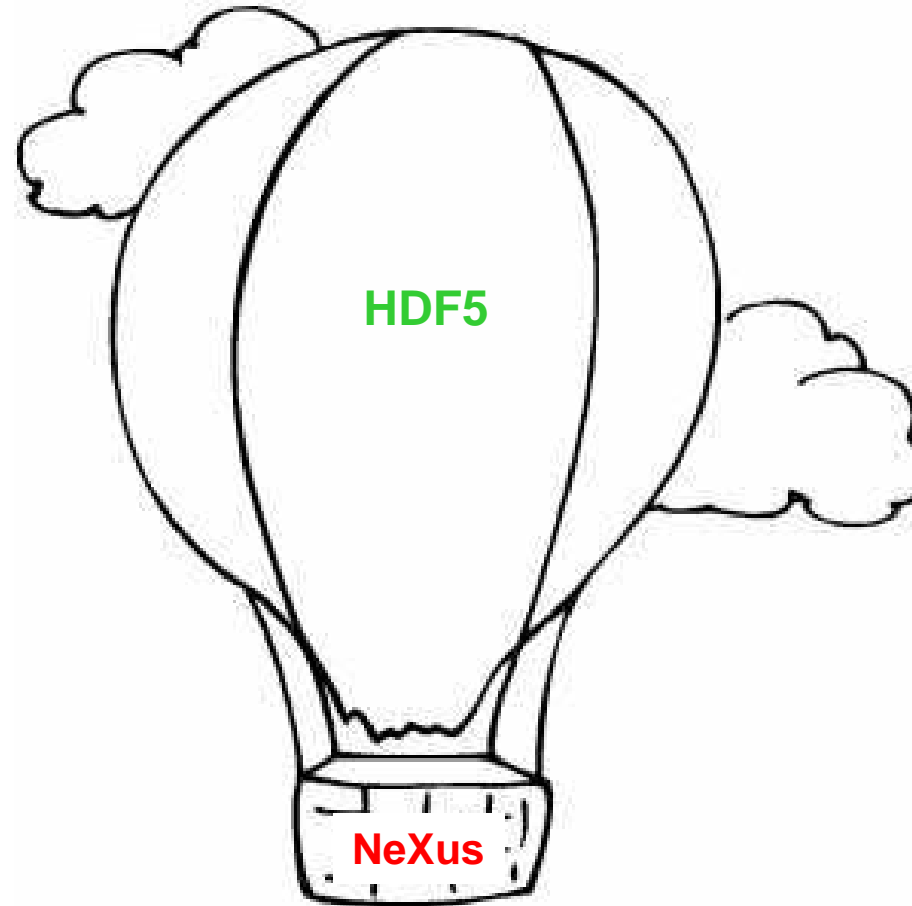
NeXus is a set conventions and an API

- to generate files in XML, HDF5 or HDF4
- to validate those conventions

PSI, May 2010: NeXus biggest challenge

After NIAC 2010

PSI, May 2010: Data Archival Challenge

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

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

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)

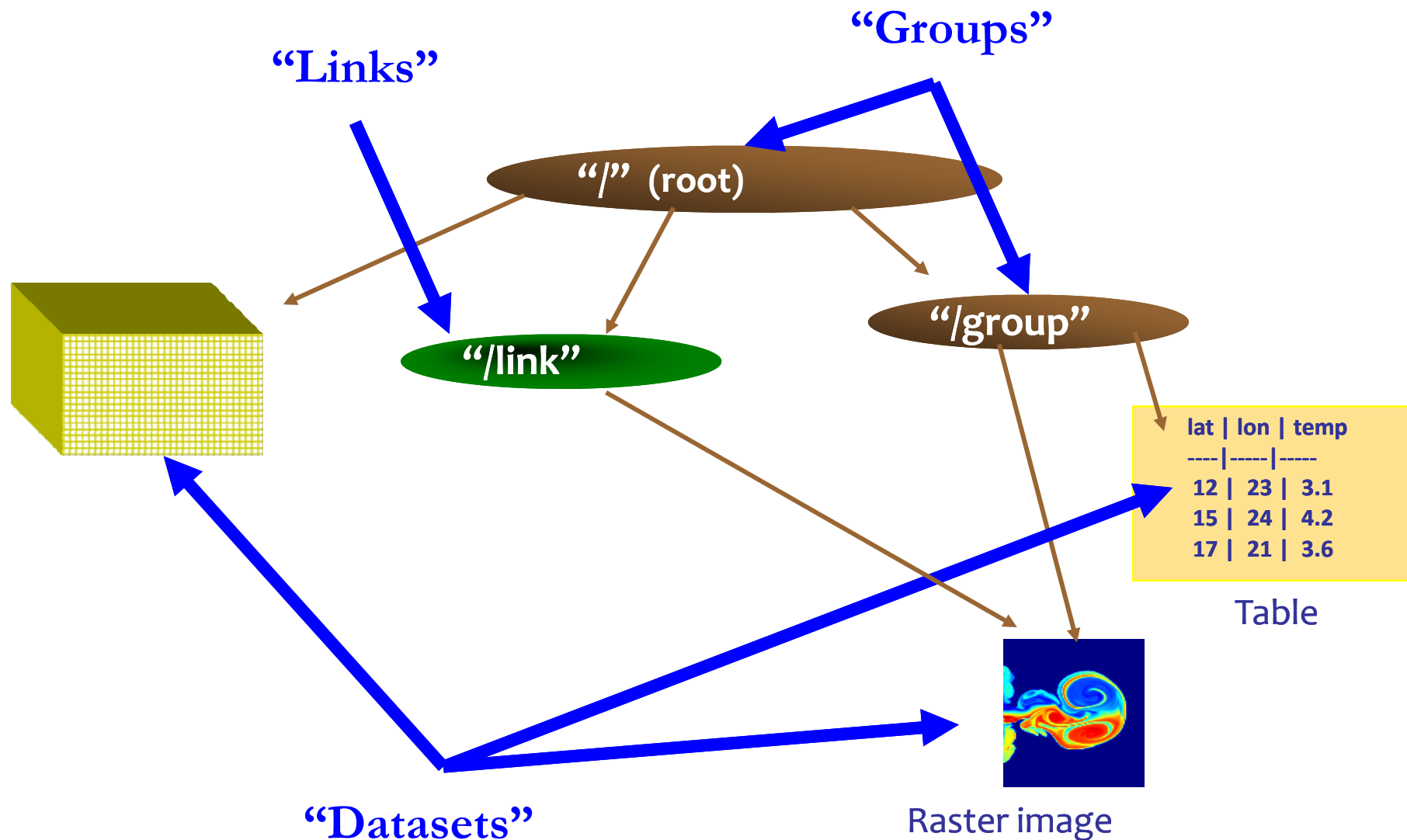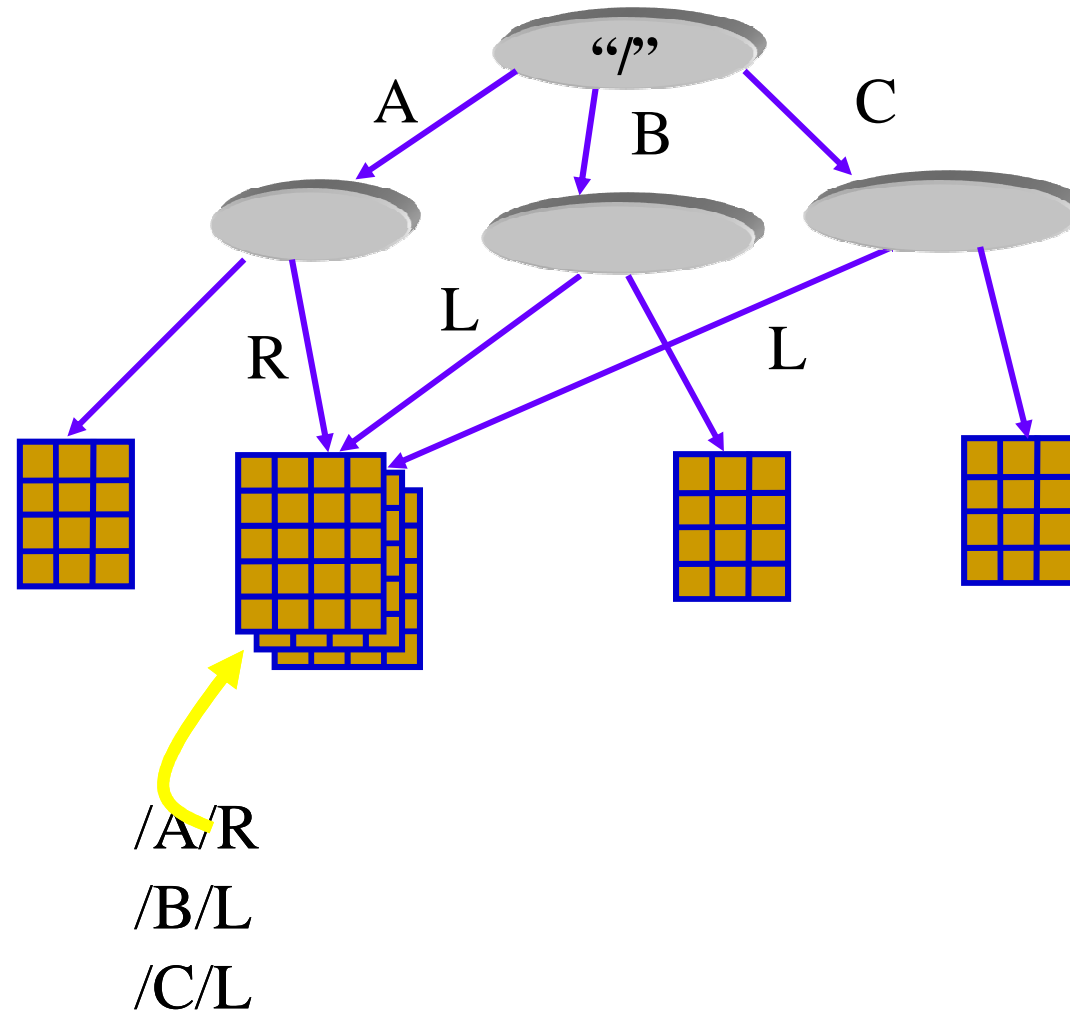# A slide from The HDF Group

Acknowledgements, Quincey Koziol, Elena Pourmal, The HDF5 Group

Structures to organize objects

"Links"

"Groups"

"/" (root)

"/link"

"/group"

lat	lon	temp
12	23	3.1
15	24	4.2
17	21	3.6

Table

"Datasets"

Raster image

HDF5 links

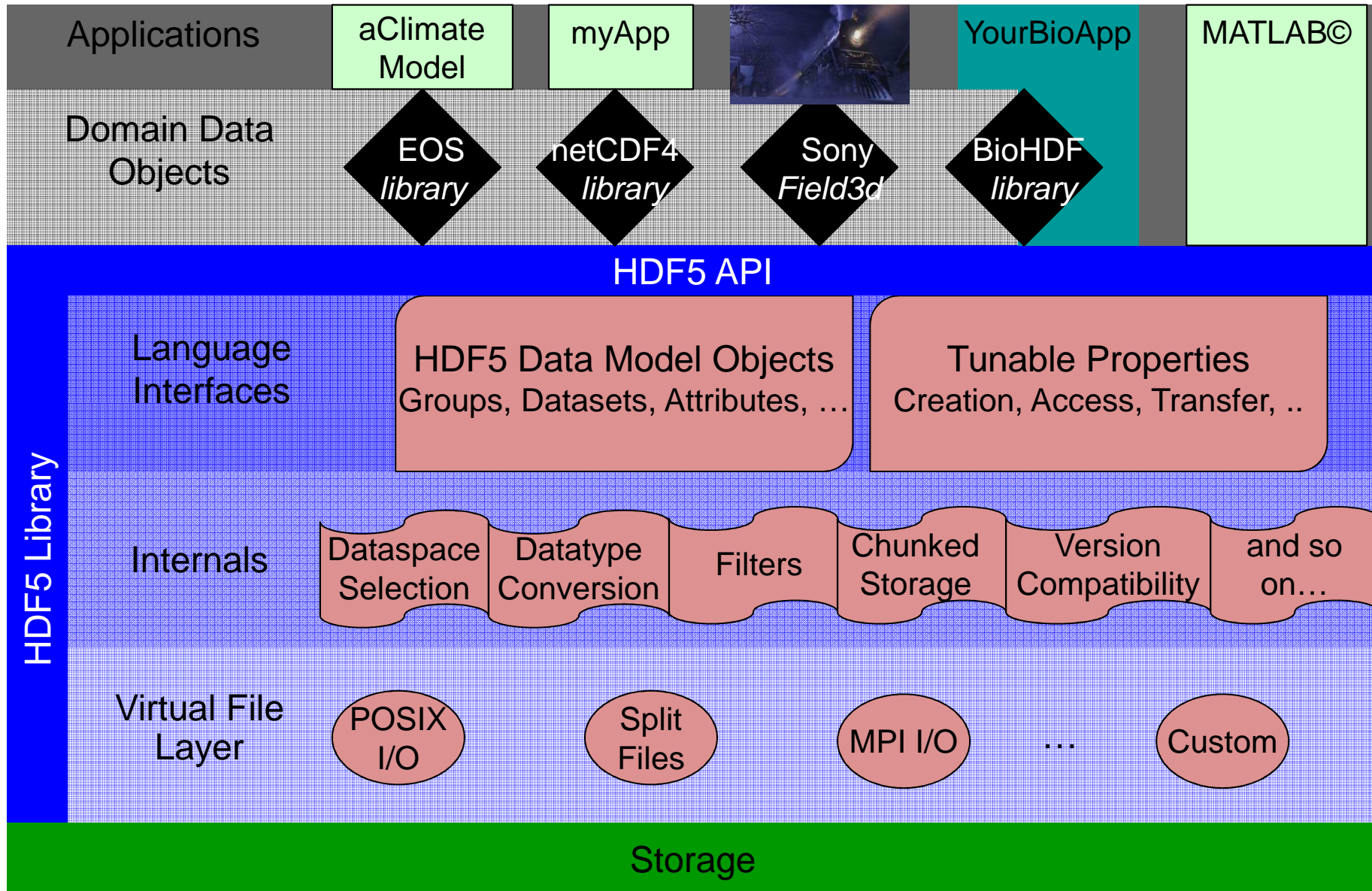# HDF External Links
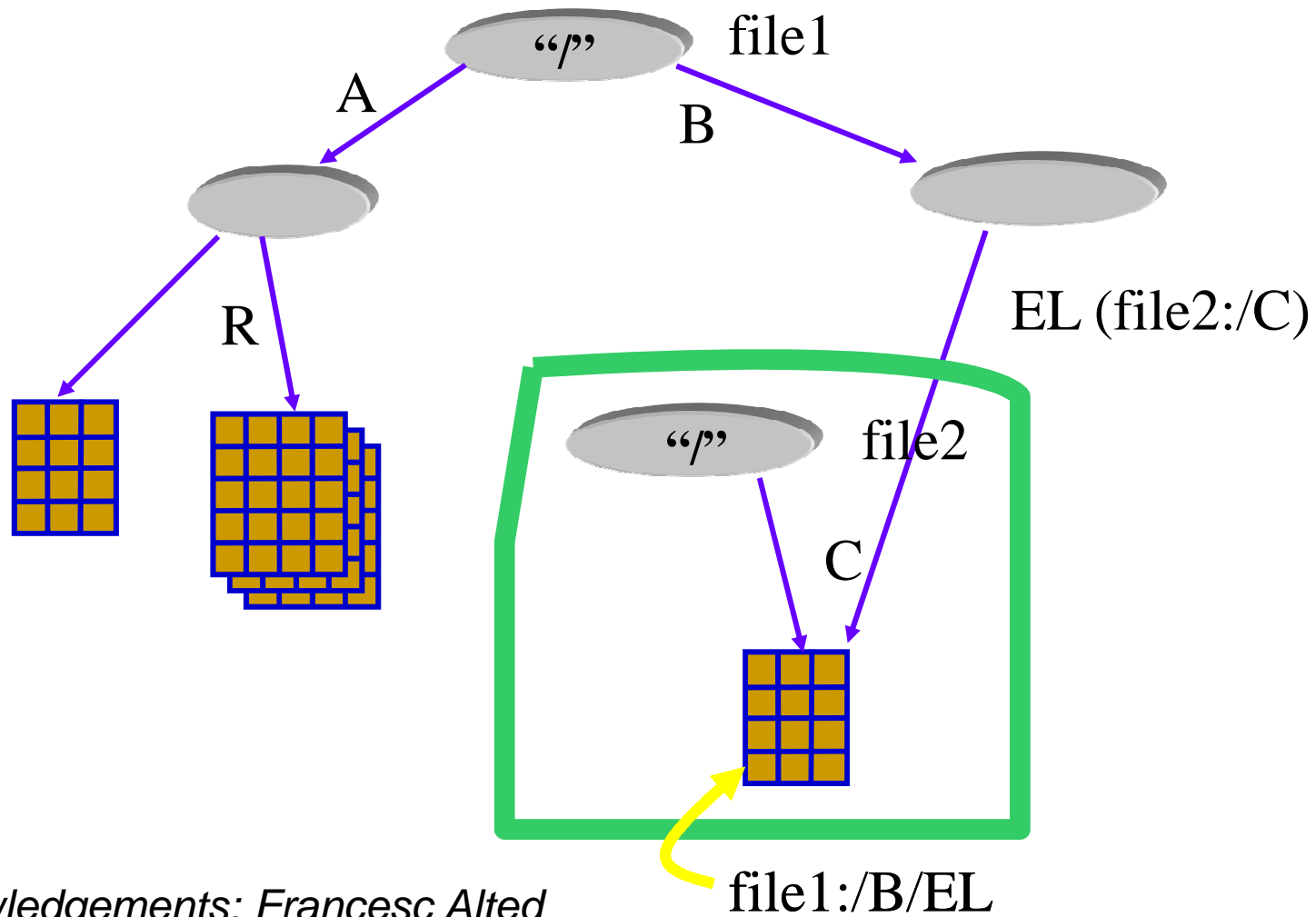
Acknowledgements: Francesc Alted

Where We'll Be Soon

- Release 1.10 - Overview
  - May make it into 2011, but no guarantee
  - Stopped adding major features, fleshing out our current efforts now
  - Major Efforts:
    - Improved scalability of chunked dataset access
    - Single-Writer/Multiple Reader (SWMR) Access
    - Improved fault tolerance

HDF5 API and Applications

Applications	aClimate Model	myApp		YourBioApp	MATLAB©

Domain Data Objects: EOS *library* — netCDF4 *library* — Sony *Field3d* — BioHDF *library*

HDF5 API

HDF5 Library

Language Interfaces

HDF5 Data Model Objects Groups, Datasets, Attributes, …	Tunable Properties Creation, Access, Transfer, ..

Internals

Dataspace Selection	Datatype Conversion	Filters	Chunked Storage	Version Compatibility	and so on…

Virtual File Layer

POSIX I/O — Split Files — MPI I/O — … — Custom

Storage

What Next?

- ## Release 1.12 - Overview
  - Special focus on HPC features & performance
  - Improvements to data model
  - "Virtual Object Layer"

What Next?

- Release 1.12 - Details
  - Improvements to data model:
    - Shared dataspaces
    - Attributes on dataspaces and datatypes
  - "Virtual Object Layer"
    - HDF5 without a file?
    - Introduce new layer in library focused on abstract data model
    - Working with Carlos Maltzahn @ UCSC to explore merging HDF5 with Ceph file system

ESRF, January 2011

Is HDF5 sufficient as the only answer for the whole synchrotron community?
Do we need another layer to integrate other synchrotron data formats e.g. CDM?

What would you answer if HDF5 itself would provide that layer?