HDF5/NeXus at ESRF

HDF5 Workshop

ESRF - Grenoble

September 2019

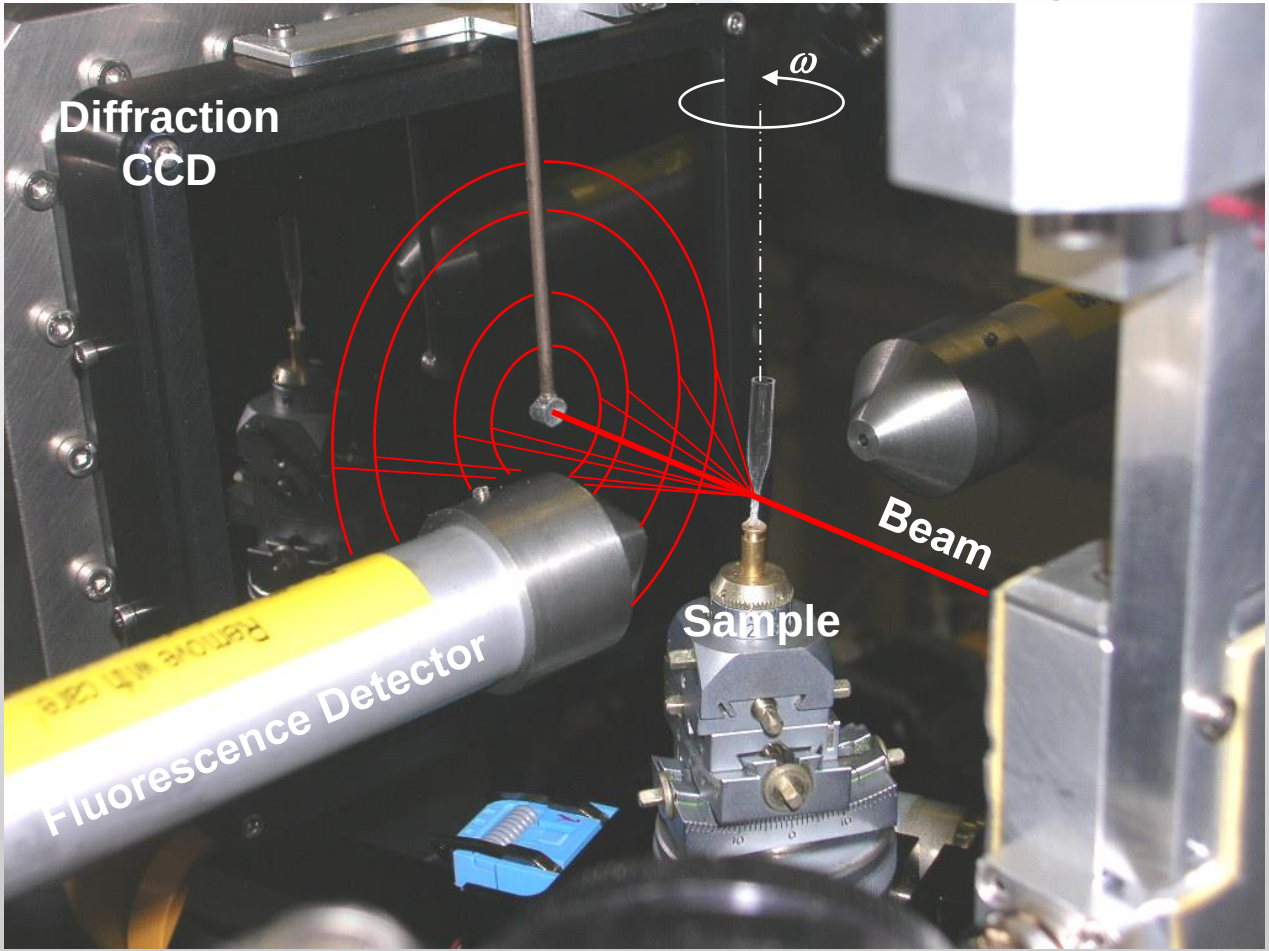
V. Armando SOLÉ

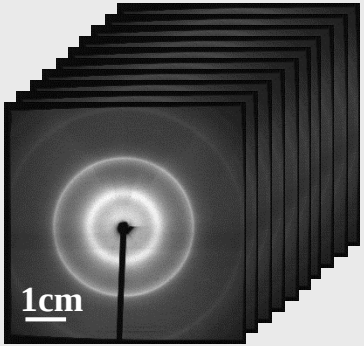
ESRF – Data Analysis Unit

This Talk

- ESRF Needs
- ESRF NeXus interpretation for raw data
- ESRF NeXus interpretation for processed data
- Other ESRF NeXus Uses: Metadata storage
- Status

ID22 – Fluorescence-Diffraction Tomography

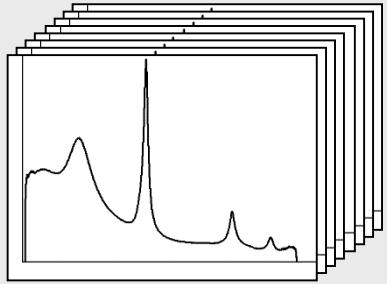




$N_y \times N_\omega$ Diffraction Images

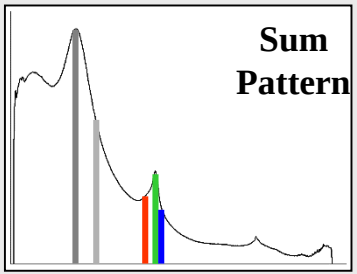
Azimuthal Integrations

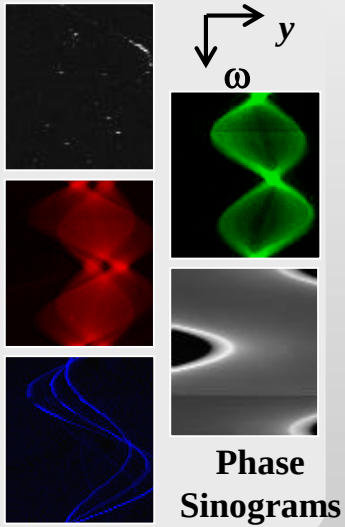
Fit2d software



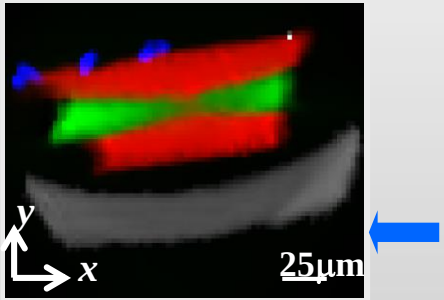
$N_y \times N_\omega$ Diffraction Patterns

PyMca software





Phase Sinograms



Reconstruction

Capillary
Calcite
Ferrite
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. Use an existing one

Lesson NOT learned (yet?):

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

ESRF Needs

- Preparing to move to HDF5 in 2010 (previous three slides)
  - The analysis programs have changed and/or adapted
  - The ID22 beamline has moved (now ID16A and ID16B)
  - The ESRF has been dismantled and re-built
  - But we do not acquire our data in HDF5 yet

Still not moved to HDF5 !!!

- ESRF 2010 needs still there in 2019 (but we are prepared 😂)
  - Single format to store different data types
    - Keep together counters, images, mca, …
  - Compression support
  - Widespread support
  - Efficient and easy access to the data for visualization and analysis

HDF5/NeXus – ESRF Interpretation for Raw Data

NXroot

Top level. One per file.

NXentry

One group per measurement

NXinstrument

Describe the instrument.
Only one per NXentry

Exclusive **Acquisition** Domain

measurement (@NXcollection)

Flattened view of everything measured

Only one per NXentry

Almost exclusive **Acquisition** Domain

sample (@NXsample)

Define the physical state of the sample during the scan

User/Scientist Domain

NXdata

The default data to be plotted.
One NXdata group per plot

User/Scientist Domain

user (@NXuser)

Details of a user, i.e., name, affiliation, email address, *etc*

Administrative Domain (GDPR? DOI?)

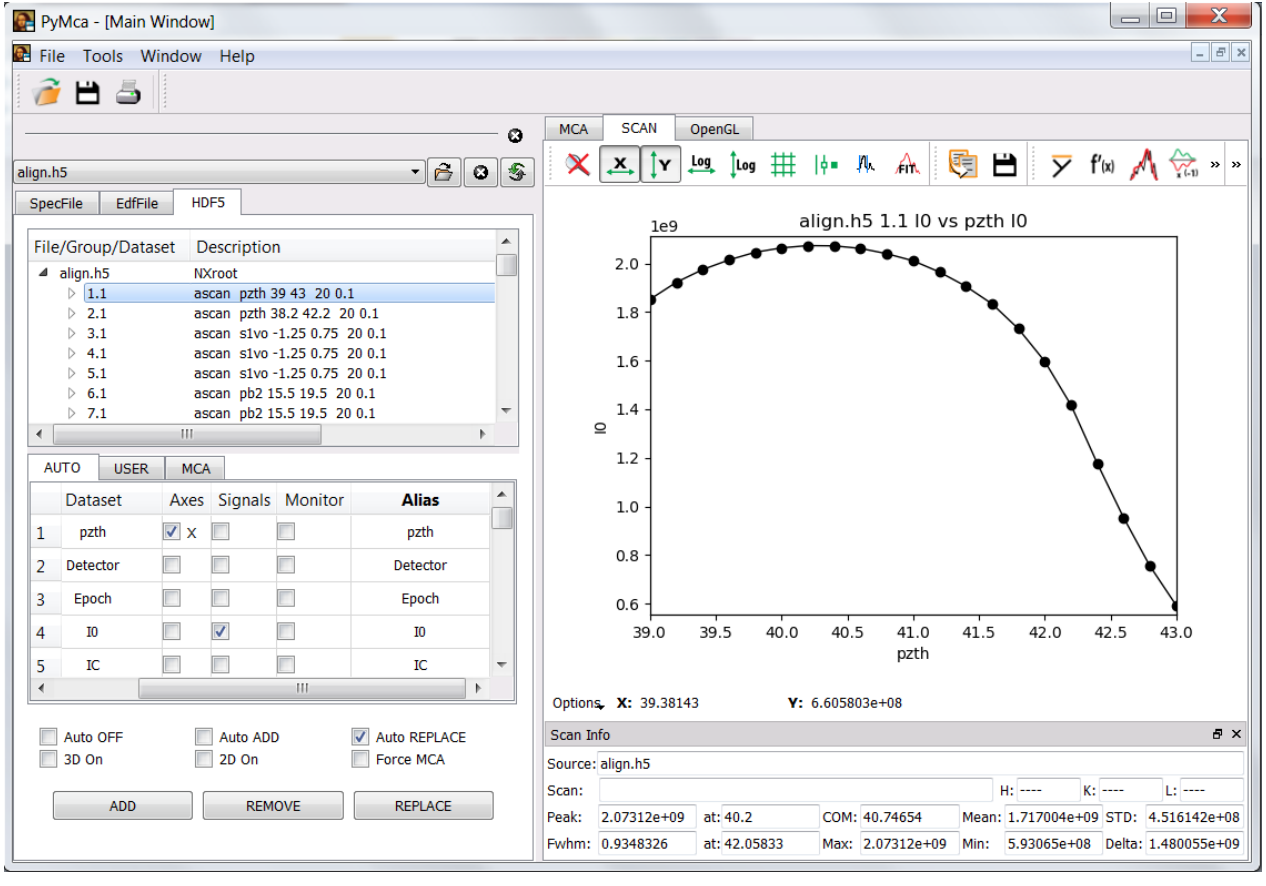
NXsubentry

Data or links to data for particular analysis

Analysis Domain

Measurement Group Convention

- Name-based convention followed by ESRF and Sardana (MAX IV, ALBA…)



- Targets interactive use
- Applications can profit

HDF5/NeXus: Requirements for Processed Data

- NeXus conventions are fairly clear in what concerns raw data
- How to store processed data in HDF5 files?
  - Needs
    - Program used
    - Configuration parameters
    - Results
    - Minimize file creation
      - More than one data treatment step into the file
      - Describe data treatment sequence

NeXus: ESRF Implementation for Processed Data (v1)

- Goals can be achieved with “extended” NXprocess groups

entry
    start_time
    end_time
    title
    process_name@NXprocess
        program_name
        version
        date
        sequence_index
        configuration@NXcollection if HDF5 supported by program
        results@NXcollection or NXdata if plot

Just a name based convention added to NXprocess

NeXus: ESRF Implementation for Processed Data (v2)

- A 100% pure NeXus way to specify the configuration: NXnote

entry
    process_name@NXprocess
        program_name
        version
        date
        sequence_index
        configuration@NXnote
            file_name
            type
            data
        results@NXcollection or NXdata if just a plot

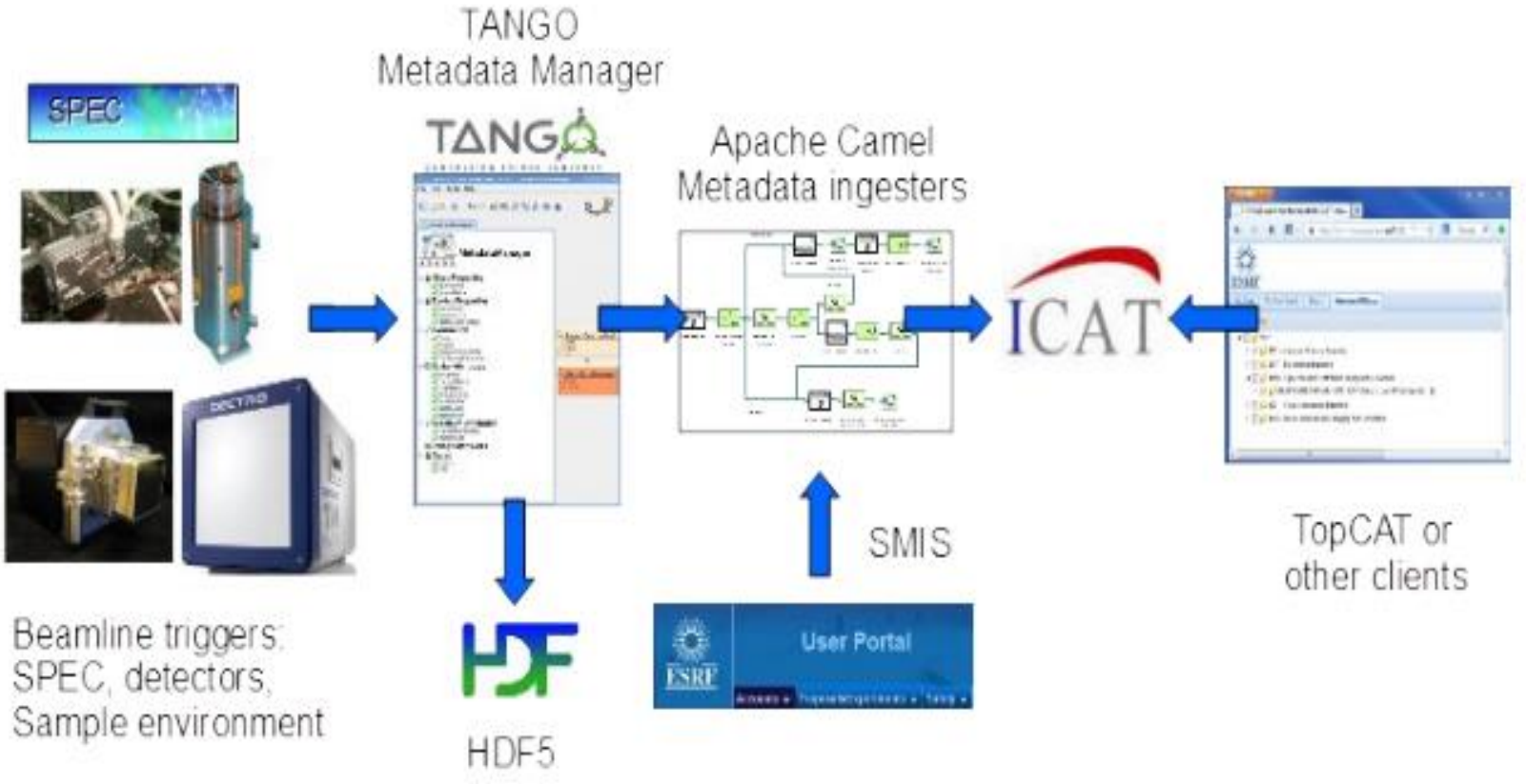
The key point is that the configuration can be used back. We have to be able to feed the original program with it.

Metada Storage: ICAT – NeXus Mapping

NeXus

ICAT

Metadata Collection Architecture



ICAT – NeXus Mapping

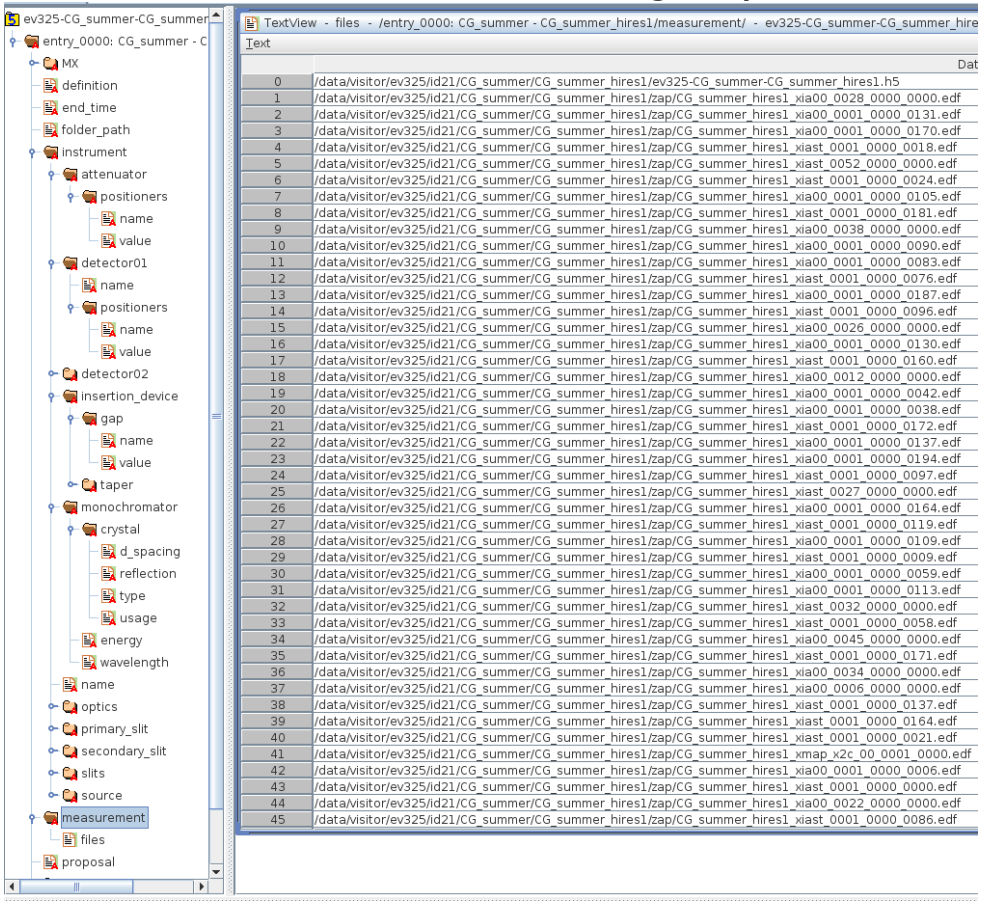
- Clear mapping from existing NeXus conventions to ICAT
  - ICAT key = Class1Class2Class3_dataset@attribute
  - Ex: NeXus current and mode in class Source inside class Instrument
    - InstrumentSource_current
    - InstrumentSource_mode

- Technique or beamline specific information as NXsubentry based keys

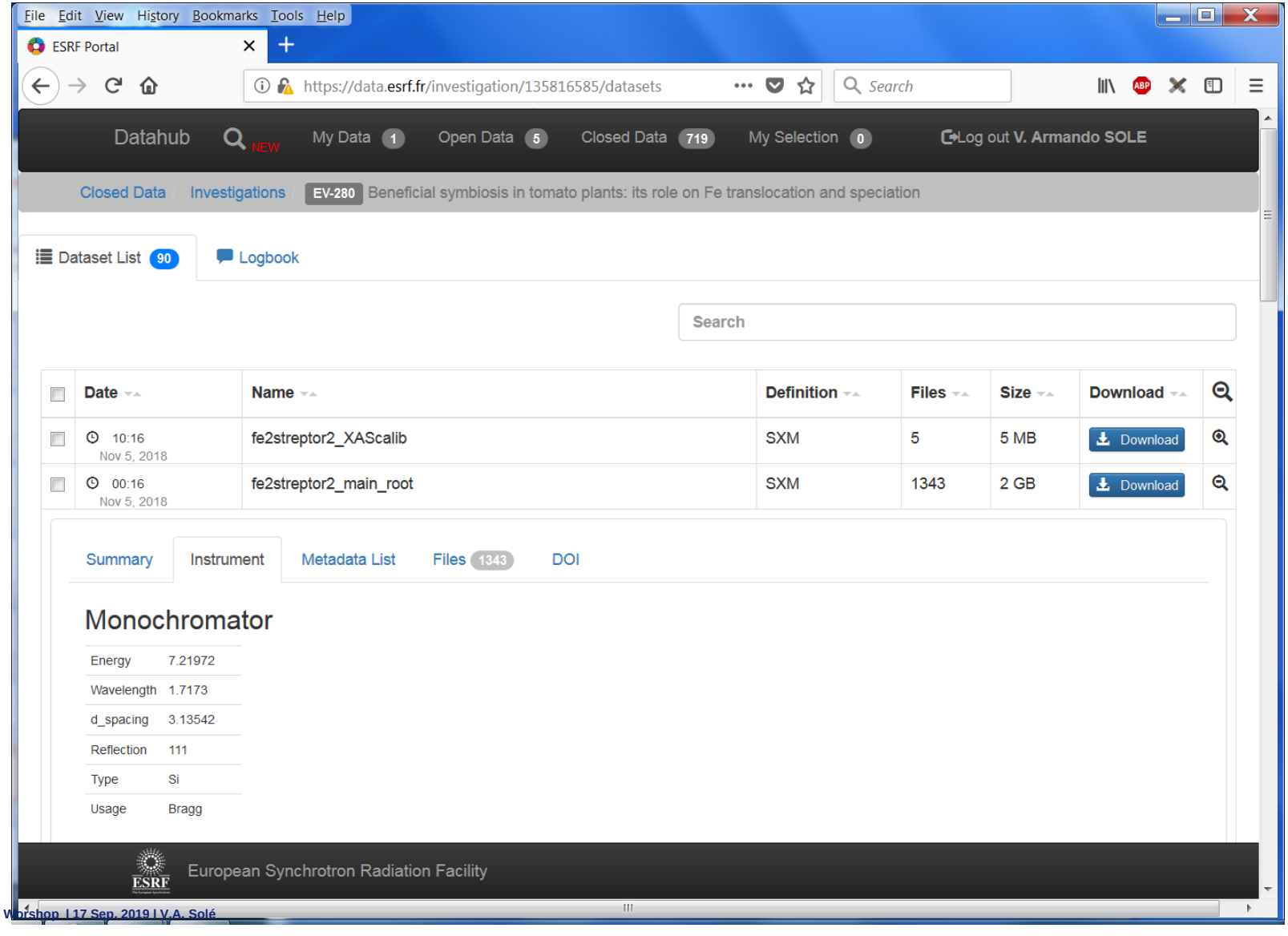
```
<group NX_class="NXsubentry" groupName="EM">
   <protein_acronym ESRF_description="Protein acronym" NAPItype="NX_CHAR">${EM_protein_acronym}</protein_acronym>
   <voltage ESRF_description="Voltage" NAPItype="NX_CHAR">${EM_voltage}</voltage>
   <magnification ESRF_description="Magnification" NAPItype="NX_CHAR">${EM_magnification}</magnification>
   <images_count ESRF_description="Number of images in movie" NAPItype="NX_CHAR">${EM_images_count}</images_count>
   <position_x ESRF_description="Position X" NAPItype="NX_CHAR">${EM_position_x}</position_x>
   <position_y ESRF_description="Position Y" NAPItype="NX_CHAR">${EM_position_y}</position_y>
   <dose_initial ESRF_description="Dose initial" NAPItype="NX_CHAR">${EM_dose_initial}</dose_initial>
   <dose_per_frame ESRF_description="Dose per frame" NAPItype="NX_CHAR">${EM_dose_per_frame}</dose_per_frame>
   <spherical_aberration ESRF_description="Spherical aberration" NAPItype="NX_CHAR">${EM_spherical_aberration}</spherical_aberration>
   <amplitude_contrast ESRF_description="Amplitude contrast" NAPItype="NX_CHAR">${EM_amplitude_contrast}</amplitude_contrast>
   <sampling_rate ESRF_description="samplingRate" NAPItype="NX_CHAR">${EM_sampling_rate}</sampling_rate>
</group>
```

ICAT – NeXus Mapping

- Collected files as list inside an NXcollection group named measurement



Aiming at using HDF5 files and external links


File Edit View History Bookmarks Tools Help
ESRF Portal
https://data.esrf.fr/investigation/135816585/datasets
Search
Datahub
NEW
My Data 1
Open Data 5
Closed Data 719
My Selection 0
Log out V. Armando SOLE
Closed Data / Investigations / EV-280 Beneficial symbiosis in tomato plants: its role on Fe translocation and speciation
Dataset List 90
Logbook
Search
Date
Name
Definition
Files
Size
Download
10:16 Nov 5, 2018
fe2streptor2_XAScalib
SXM
5
5 MB
Download
00:16 Nov 5, 2018
fe2streptor2_main_root
SXM
1343
2 GB
Download
Summary
Instrument
Metadata List
Files 1343
DOI
Monochromator
Energy 7.21972
Wavelength 1.7173
d_spacing 3.13542
Reflection 111
Type Si
Usage Bragg
European Synchrotron Radiation Facility

Status of HDF5/NeXus @ ESRF – Raw Data

- Acquisition
  - SPEC (legacy control system)
    - HDF5 not worth as native output. Use *silx convert* if desired
  - Bliss (new ESRF control system)
    - HDF5 native output operational
  - LImA (ESRF Library for Image Acquisition)
    - HDF5 native output operational
    - Using Direct Chunk Write for efficient multithreaded compression

Why that functionality is not offered by HDF5 itself?

Status of HDF5/NeXus @ ESRF – Data Analysis

- Data Analysis

✓ - Capability to read HDF5 files (preferred data analysis I/O format)

✓ - Unified (h5py-like) API to access all data formats

✓ - Support of NeXus NXdata I/O in viewers and analysis codes

✓ - Provide provenance via NXprocess (pyFAI, PyMca, PyNX,…)

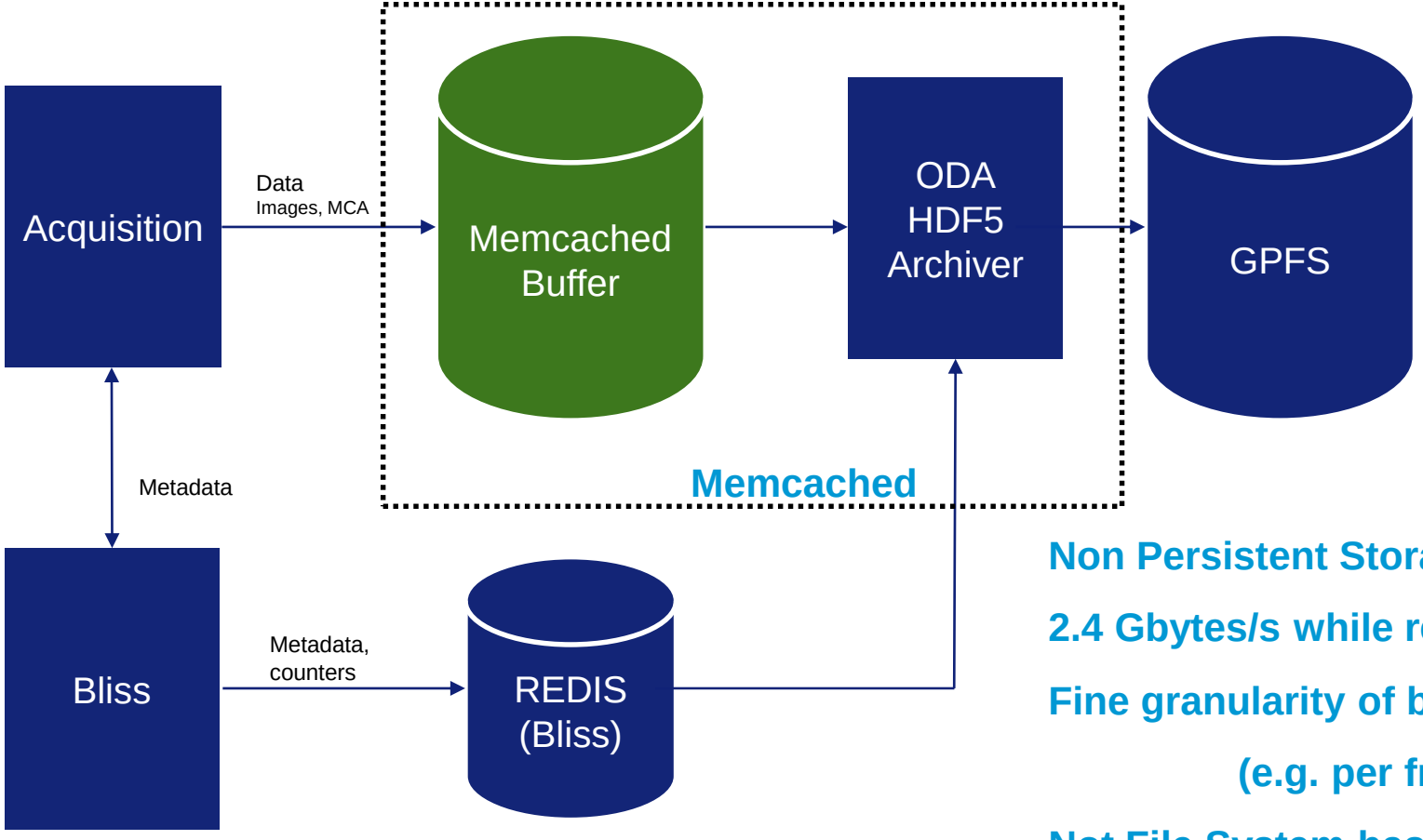
- Only one NeXus application definition supported (NXcxi)

Status of HDF5/NeXus @ ESRF – Online Data Analysis

- HDF5 is not appropriate for online data analysis
  - HDF5 Specific reasons
    - Potential concurrent access issues
      - SWMR is not enough (yet?): static file structure, flush…

      Ex. It can be OK for tomography but not for spectroscopy
  - Generic reasons
    - If we have to read from a file anything will be slow
      - Basically no data format is appropriate
      - Studying to externalize via REDIS + memcached
        - Writing a file becomes a particular case of ODA
        - File writing should be the last step, not the first one!

Online Data Analysis – HDF5 Writing a Particular Case of ODA



Non Persistent Storage

2.4 Gbytes/s while reading

Fine granularity of blocks

(e.g. per frame)

Not File System based

Status of NeXus @ ESRF - Metadata

- Data Policy and NeXus
  - Mirror ICAT and NeXus master file done
  - External links between master file and raw HDF5 files desirable

Ideally one should enable processing a dataset* from its master file

(*) A dataset in this context is not an HDF5 dataset but a collection of data

Summary

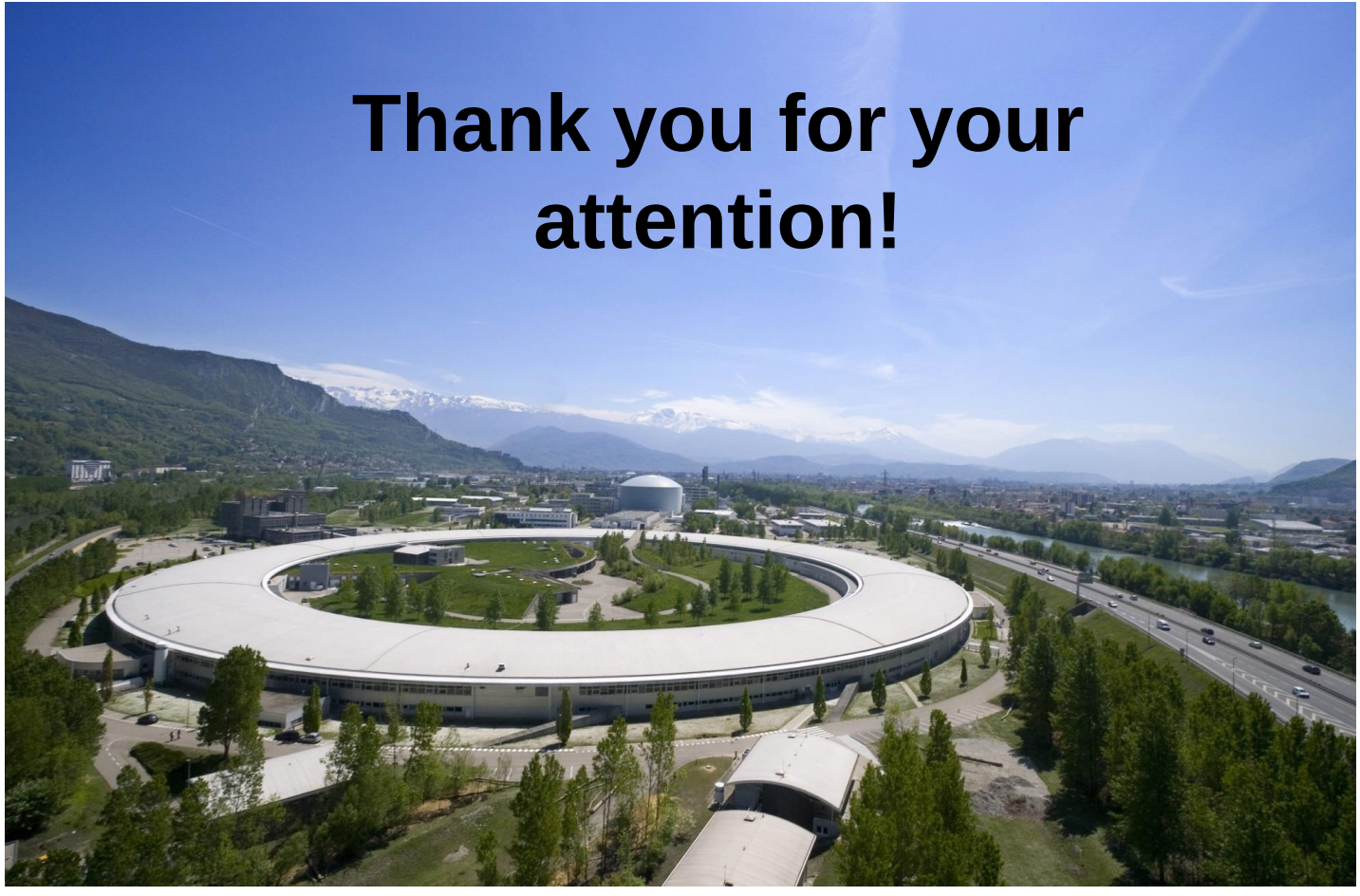
The Weight of Legacy

Adoption of HDF5/NeXus has been slower at the ESRF than at other synchrotrons due to the raw data being acquired in different formats. Detector output in HDF5 and the deployment of Bliss are speeding things up.

User experience with HDF5 files has to be better than with legacy formats HDF5 should not be the question but the answer.

Concerning data analysis, ESRF started to provide HDF5 support in 2009. Currently making convenient **use of the NeXus formalism as output and as integral part of the ESRF data policy**.

HDF5 and NeXus great for archival. Efficient **online data analysis needs to avoid the use of files as input** and HDF5 is no exception.


Thank you for your attention!