NeXus data formats at EuXFEL

Yury Kirienko

yury.kirienko@xfel.eu
Data Analysis Group

20 May 2021

Intro

In PaNOSC, we want to learn how to produce scientific results which [among other] are:

- open
- standardized
- reproducible
- self-descriptive

(FAIR principles, yep)

For proton and neutron data (PaN in PanOCS), we do have a format that provides all this, namely the NeXus format.

What is NeXus?

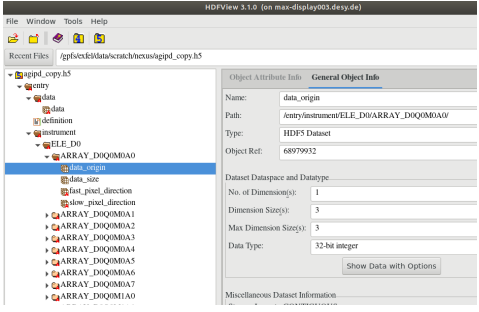
Nexus in a nutshell

- Nexus is a data exchange and archival format for neutron, X-ray and muon experiments (beamline data in general)
- NeXus specifies a dictionary of well-defined domain-specific field names.
- It is built on top of HDF5 and adds domain-specific rules for organizing data within HDF5 files.

NeXus environment

- NXDL definition language
  - generic Base definitions and specific Application definitions
  - support for different domains, experiment types (e.g. `NXmx`, `NXarpes`, `NXxas` etc)
  - reuse existing definitions (`NXdetector`, `NXtransformaton` etc. can be integrated in any application)
  - Definition can evolve (proper versioning enables the correct use)
- Data Format
  - HDF5 (Nexus can encapsulate EuXFEL Data files) or XML (small projects where text editor is enough)
  - Hierarchical entries with attributes connecting the data field to definitions
- Tools
  - Full version control on github
  - Automatic documentation generator based on rst
  - Data viewer (**hdfview**, **extra-data**, PaNOSC projects etc)
  - Data format validator

NeXus Definitions



Structure:

ENTRY: (required) NXentry

Note, it is recommended that file_name and file_time be included as attributes at the root of a file that includes NXmx. See NXroot.

title: (optional) NX_CHAR

start_time: (required) NX_DATE_TIME

ISO 8601 time/date of the first data point collected in UTC, using the Z suffix to avoid confusion with local time. Note that the time zone of the beamline should be provided in NXentry/NXinstrument/time_zone.

end_time: (optional) NX_DATE_TIME

ISO 8601 time/date of the last data point collected in

Competitors: CXI

See the details: `https://www.cxidb.org`

- CXI stands for Coherent X-ray Imaging
- May be treated as "relaxed" or "more simple" Nexus
- Design is mostly derived from Nexus

The main goal of the Coherent X-ray Imaging Data Bank is to address these problems by creating an open repository for CXI experimental data.

Nexus vs CXI vs HDF5

- Both Nexus and CXI use HDF5 as a data container, i.e.
  - Nexus is a subset of HDF5 (`.nxs` $\subset$ `.h5`)
  - CXI is a subset of HDF5 (`.cxi` $\subset$ `.h5`)
  - but neither of them is a subset of the other
  - however, there are rather simple transformations which allow to convert one into another
- CXI is more flexible but less descriptive
- Nexus is extremely complex and strict, but fully unambiguous and self-contained

NeXus @ EuXFEL

History and software

What we use:

- **cctbx** toolbox: `https://github.com/cctbx/cctbx_project`
  - The Computational Crystallography Toolbox (cctbx) is being developed as the open source component of the Phenix project.
  - Pro: powerful and mature project
  - Contra: huge, not easy to install, difficult to debug

NeXus @ EuXFEL

History and software

What we use:

- **cctbx** toolbox: `https://github.com/cctbx/cctbx_project`
  - The Computational Crystallography Toolbox (cctbx) is being developed as the open source component of the Phenix project.
  - Pro: powerful and mature project
  - Contra: huge, not easy to install, difficult to debug
- **EXtra-**packages: `https://github.com/European-XFEL/`
  - **extra-geom** for detector geometries
  - **extra-data** for data handling
  - **h5glance** for fast HDF5 access

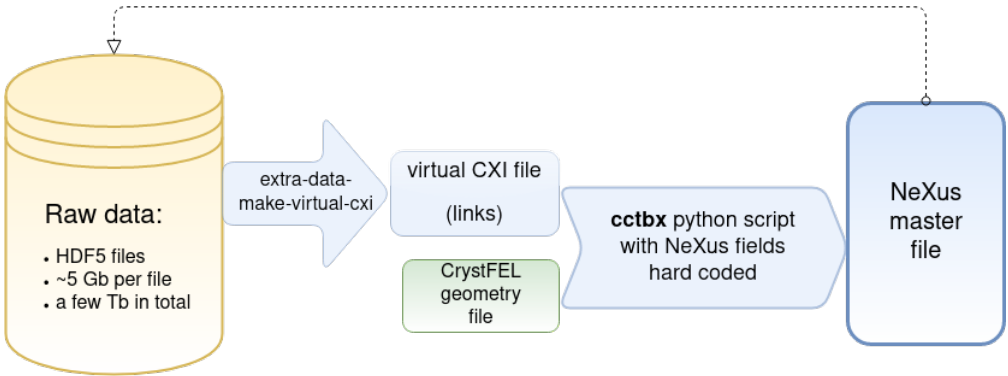
NeXus @ EuXFEL

History and software

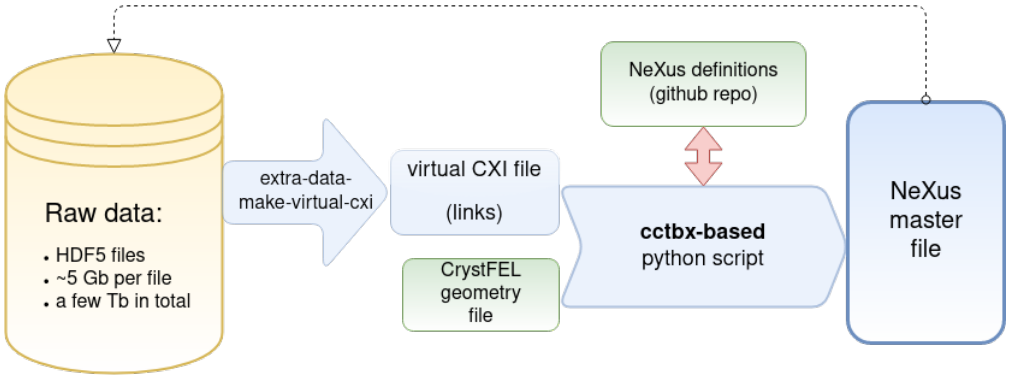
What we use:

- **cctbx** toolbox: `https://github.com/cctbx/cctbx_project`
  - The Computational Crystallography Toolbox (cctbx) is being developed as the open source component of the Phenix project.
  - Pro: powerful and mature project
  - Contra: huge, not easy to install, difficult to debug
- **EXtra-**packages: `https://github.com/European-XFEL/`
  - **extra-geom** for detector geometries
  - **extra-data** for data handling
  - **h5glance** for fast HDF5 access
- **CrystFEL** software suite: `https://www.desy.de/~twhite/crystfel/`

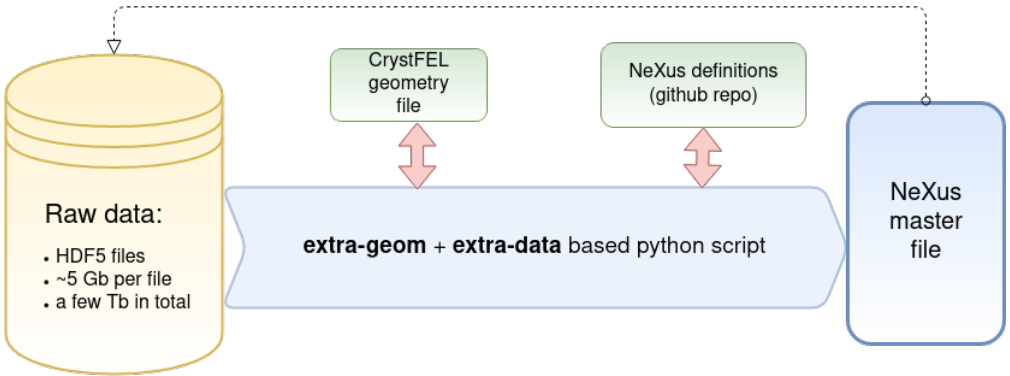
Initial NeXus pipeline



Current NeXus pipeline

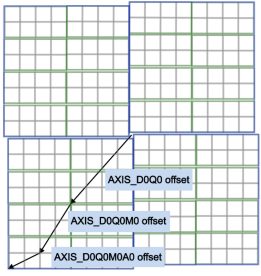


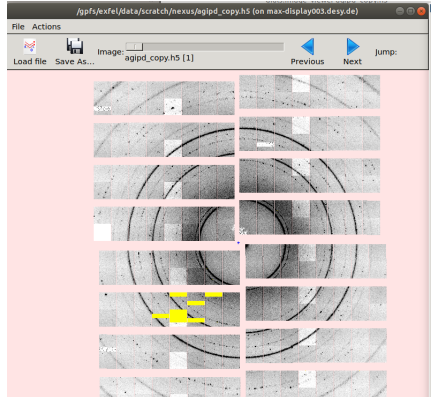
Future NeXus pipeline



Results

How we store geometry and how it looks like in the end





NeXus for Simulations

In EuXFEL, we also work on simulations (PaNOSC WP5, Juncheng E)

- **extra-geom** for detector description
- **Simex** for simulations
- **extra-data** and **h5glance** for visualisation
- **NeXus** as a final result

Problems with NeXus and possible solutions

If NeXus file format is *that* beautiful why it is not *equally* spread?

- it is complicated

Problems with NeXus and possible solutions

If NeXus file format is *that* beautiful why it is not *equally* spread?

- it is complicated
- lack of software

Problems with NeXus and possible solutions

If NeXus file format is *that* beautiful why it is not *equally* spread?

- it is complicated
- lack of software
- it's now what scientists really need right now (publications)

Problems with NeXus and possible solutions

If NeXus file format is *that* beautiful why it is not *equally* spread?

- it is complicated
- lack of software
- it's now what scientists really need right now (publications)
- more templates, better documentation

Problems with NeXus and possible solutions

If NeXus file format is *that* beautiful why it is not *equally* spread?

- it is complicated
- lack of software
- it's now what scientists really need right now (publications)
- more templates, better documentation
- more open software, better coordination

Problems with NeXus and possible solutions

If NeXus file format is *that* beautiful why it is not *equally* spread?

- it is complicated
- lack of software
- it's now what scientists really need right now (publications)

- more templates, better documentation
- more open software, better coordination
- integration with publishing systems? ¯_(ツ)_/¯