NeXus: A DataFormat for x-ray, n and muon Scattering

Mark Könnecke, on behalf of the NIAC

Paul Scherrer Institute
Switzerland

August 7, 2014

The Predicament of the Traveling Scientist

- A different data format wherever she goes

NeXus

The Predicament of the Traveling Scientist

- A different data format wherever she goes
- Waists time converting formats or writing readers

The Predicament of the Traveling Scientist

- A different data format wherever she goes
- Waists time converting formats or writing readers
- Waists time loading data from inefficient data formats

NeXus

The Predicament of the Traveling Scientist

- A different data format wherever she goes
- Waists time converting formats or writing readers
- Waists time loading data from inefficient data formats
- Needs multiple files in different formats, local knowledge, lab books etc to analyse data

The Predicament of the Traveling Scientist

- A different data format wherever she goes
- Waists time converting formats or writing readers
- Waists time loading data from inefficient data formats
- Needs multiple files in different formats, local knowledge, lab books etc to analyse data
- Cannot read her collaborators data

NeXus Mission

- Definition of a standard data format
  - Rules
  - Validation tools
- Promotion of NeXus
  - Documentation
  - NeXus API
  - Outreach to the scientific community

NeXus Design

- Complete data for typical use
- Full Beamline Description (FBD)
- Extendable, add additional data as you please
- Self describing
- Easy automatic plotting
- Platform independent, public domain, efficient
- Suitable for a wild variety of applications
- Express validatable standards

NeXus Levels

1. Physical file format and API for accessing files
2. Rules for storing data in files
3. Component and application definitions
4. NeXus Utilities

Physical File Format

- Portable, self describing, extendable, public domain
- HDF-5
- NeXus adds names and rules to HDF-5
- Historical: HDF-4, XML, NeXus-API
- I/O to NeXus files with either HDF-5 or the NeXus-API

HDF-5

- Efficient binary format
- grouping support
- on the fly compression
- reading/writing subsets of data
- Dimensions appendable
- Public domain C, F77 access library
- Well supported by many scientific tools
- Simple to read, see Michael Rissi talk for an example
- Used by: NASA, Boeing, Deutsche Bank, HPC, the weathermen, ...
- Supported and maintained by the HDF group

NeXus

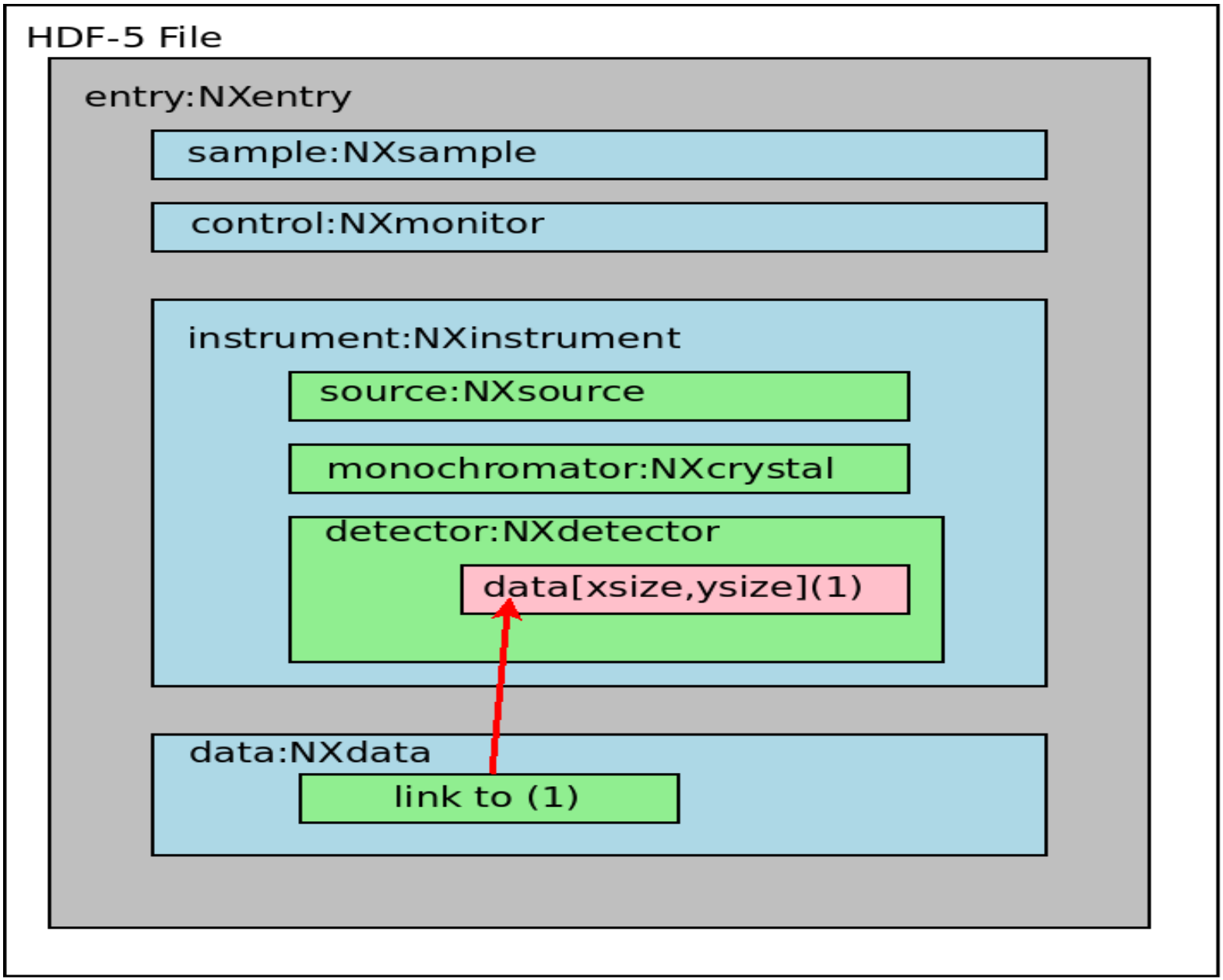
NeXus Objects

- Files
- Groups identified by name and a classname
- Scientific data sets
- Attributes
- Links

Coordinate Systems

- McStas Coordinate System
- NeXus stole the CIF way of storing translations, rotations and dependencies
- CBFlib will be updated to work with this

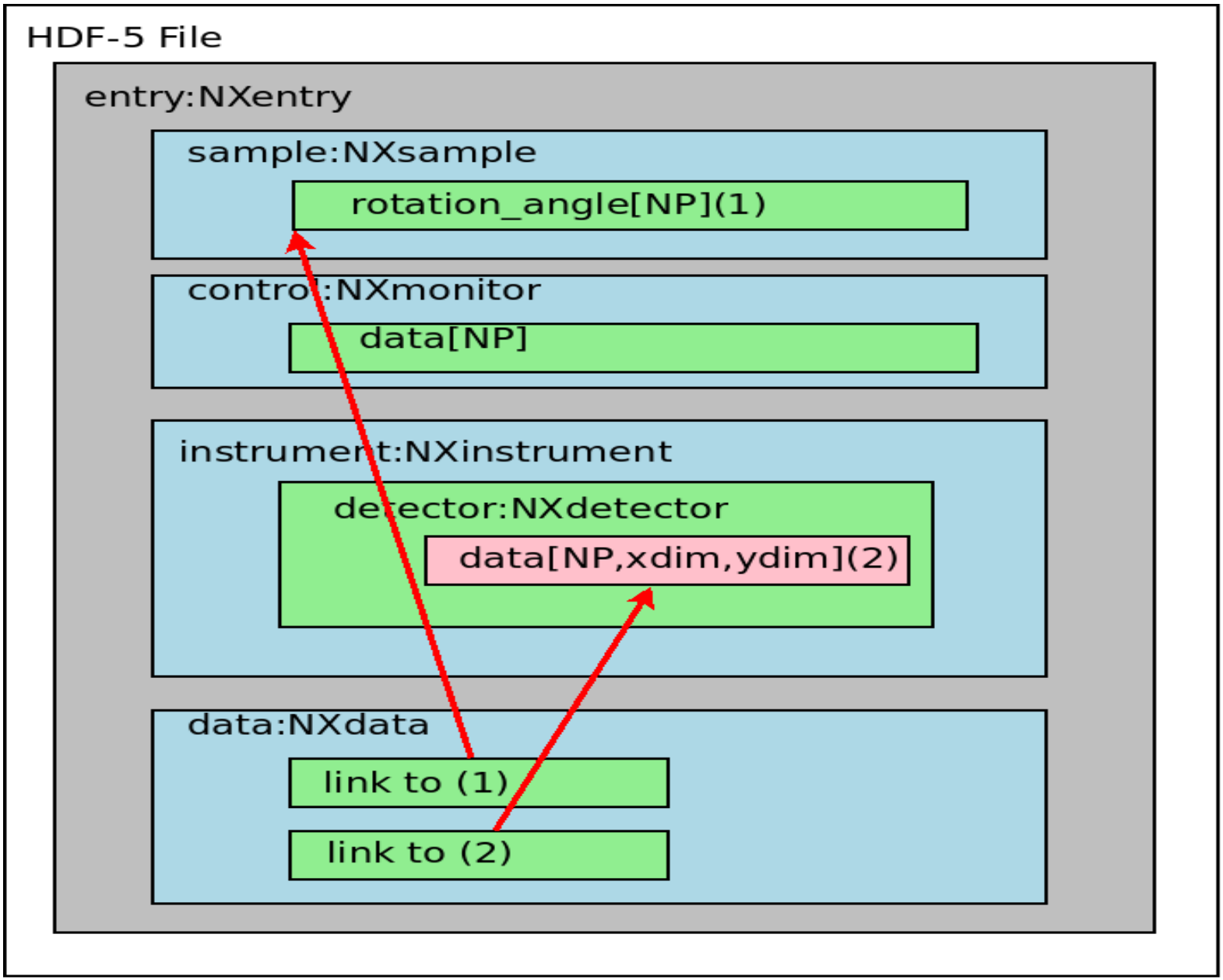
NeXus Raw Data File Structure



Scans and Rasterisation

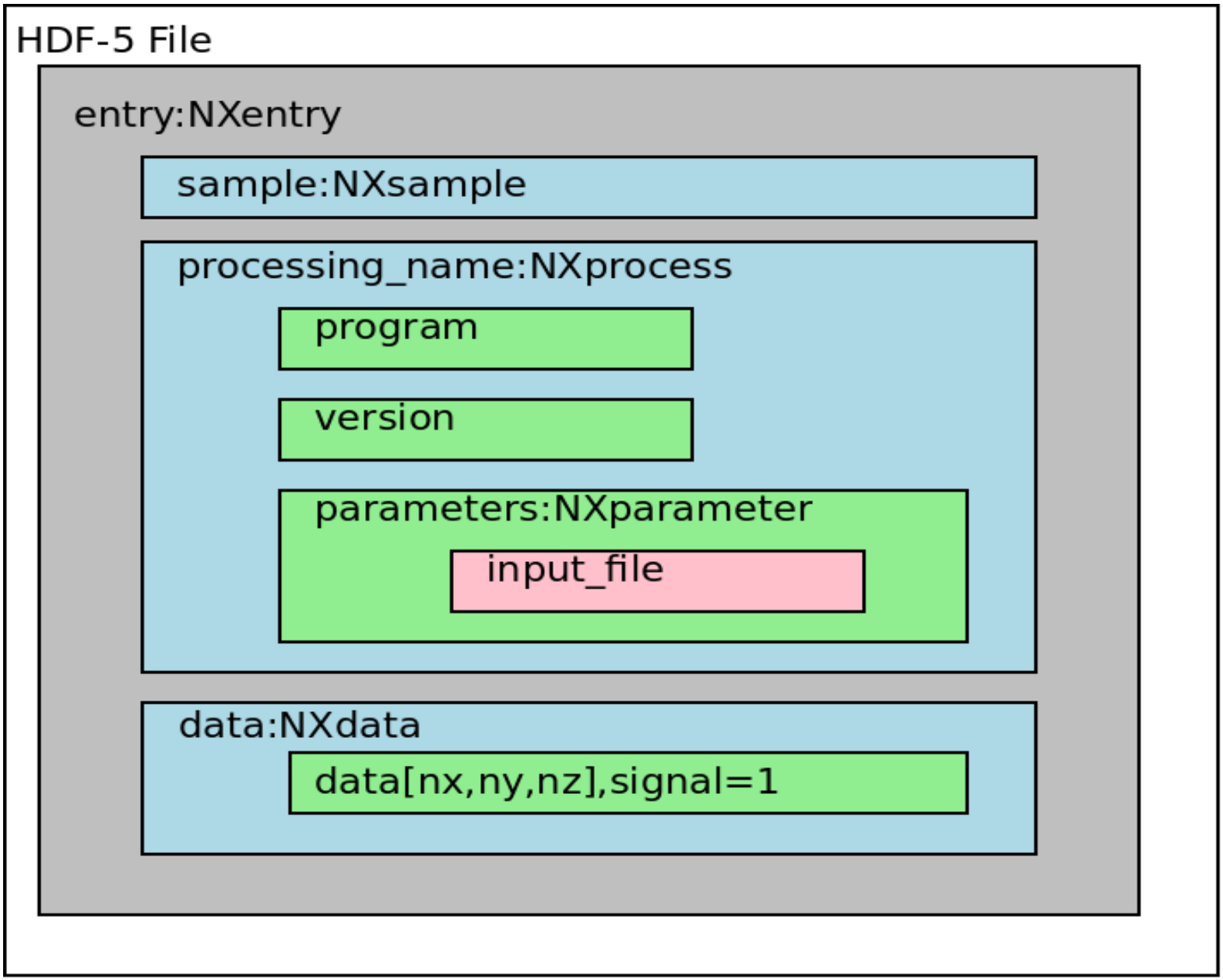
- Come in all shapes and sizes
- Captured by rules:
  - NP is the number of scan points
  - Store all varied parameters as arrays of length NP at the appropriate place in the NeXus hierarchy
  - For area detectors, NP is always the first dimension
  - In NXdata: create links to counts and varied variables
- Rasterisation is treated similar to scans

Scan Example: sample rotation, area decetor



NeXus

Processed Data Structure



NeXus

Why the Hierarchy??

- Supports self description and allows short names in components

Why the Hierarchy??

- Supports self description and allows short names in components
- Name, classname pair allows for multiple components of the same type

Why the Hierarchy??

- Supports self description and allows short names in components
- Name, classname pair allows for multiple components of the same type
- NXentry allows for multiple datasets in the same file

Why the Hierarchy??

- Supports self description and allows short names in components
- Name, classname pair allows for multiple components of the same type
- NXentry allows for multiple datasets in the same file
- NXdata supports automatic plotting

Why the Hierarchy??

- Supports self description and allows short names in components
- Name, classname pair allows for multiple components of the same type
- NXentry allows for multiple datasets in the same file
- NXdata supports automatic plotting
- Hierarchy makes namespace manageable for full beamline descriptions

Rules for Storing Data Items

- Store physical values in C storage order; exceptions allowed
- Use NeXus components and dictionary names
- Missing names will be quickly accepted by the NIAC
- Units must be specified
- Rules associate axis with data

NeXus Component and Application Definitions

- Component definitions: Dictionaries of allowed field names for the various NeXus groups
- **Application Definitions**
  - **Define what has to be in a NeXus file for a certain application**
  - **Defines standards**
  - **Another view: Contract between file producers and users about what has to be in a NeXus file for a well defined purpose**
  - **Validation by NXvalidate**
- Written in NeXus Definition Language, NXDL

Availability of Component and Application Definitions

- Component definitions for all the standard NeXus classes and many beamline components exist
- Missing fields or component definitions can be added quickly
- Suggested application definitions for many techniques exist

NeXus Tools

nxbrowse	CLI NeXus browser
nxtree	prints NeXus tree
NXmeta	dumps all NeXus meta data
nxtranslate	transforms into NeXus
nxvalidate	validates NeXus files against a NXDL application definition
nxextract	converts from NeXus to ASCII and binary
nxplot	plots any NeXus file

NeXus Benefits

Benefit 1 By using a discoverable data format like NeXus, HDF-5, people can at least figure out what is in the data file.

NeXus Benefits

Benefit 1 By using a discoverable data format like NeXus, HDF-5, people can at least figure out what is in the data file.

Benefit 2 Using predefined names from a dictionary gives meaning to the data in a file.

NeXus Benefits

Benefit 1 By using a discoverable data format like NeXus, HDF-5, people can at least figure out what is in the data file.

Benefit 2 Using predefined names from a dictionary gives meaning to the data in a file.

Benefit 3 Using a shared API reduces learning costs and increases application stability.

NeXus Benefits

Benefit 1 By using a discoverable data format like NeXus, HDF-5, people can at least figure out what is in the data file.

Benefit 2 Using predefined names from a dictionary gives meaning to the data in a file.

Benefit 3 Using a shared API reduces learning costs and increases application stability.

Benefit 4 With NeXus, HDF-5, plus professional programming techniques a DA application can read any file which contains the required data.

Benefit 5 Storing as much data as possible increases the likelihood that the needed data is actually on file, even for unforeseen uses.

More Benefits

Benefit 5 Storing as much data as possible increases the likelihood that the needed data is actually on file, even for unforeseen uses.

Benefit 6 Storing as much data as possible allows to track down causes of funny results

More Benefits

Benefit 5 Storing as much data as possible increases the likelihood that the needed data is actually on file, even for unforeseen uses.

Benefit 6 Storing as much data as possible allows to track down causes of funny results

Benefit 7 Storing as much data as possible helps to protect against scientific fraud

NeXus

More Benefits

Benefit 5 Storing as much data as possible increases the likelihood that the needed data is actually on file, even for unforeseen uses.

Benefit 6 Storing as much data as possible allows to track down causes of funny results

Benefit 7 Storing as much data as possible helps to protect against scientific fraud

Benefit 8 Application Definitions make all data handling problems go away

NeXus

NeXus Usage

- Soleil: 20 out of 26 instruments do NeXus, 2 mill files
- PSI-SINQ: 11 from 16 instrument on NeXus, 1.4 Mill files
- Lujan/LANL: 11 instruments, no change, 1 million files
- ANSTO: 7 going to 10
- KEK: 10, 6 planned
- SNS: 14,3 in the pipeline
- DESY: 0, 11 in 2 Jahren
- Diamond: 7 NeXus only, 17 writing, moving to 18 as primary format
- ISIS: 8 using, 20 writing, planned: 20 using
- Less intense users:
- PSI-SLS: 2 planned,
- ESRF: 2 beamlines, limited to NXentry, NXcollection, NXdata, moving to 4
- HZB: 3 Neutron, 1 synchrotron, 3 planned
- Muons: 4 instruments

Dispelling NeXus Myths

Inmature No longer true

Complex Everything in NeXus is there for a reason, often compiled from differing and conflicting requirements. Every other format with the same scope as NeXus will be as complex

Unresponsive NeXus is inclusive. Volunteer effort: join the NIAC!

Not used NeXus is used

Must store useless stuff Typical application definition: 10-30 values. Read only what you need!

Not Standardized Enough NeXus allows to define and validate real standards

Conclusion

- NeXus/HDF-5 is a mature and capable data format
- There is no other standard then NeXus on the horizon
- HDF-5 appears to be a consensus in the community
- New things are developed with NeXus everywhere, uptake at established sites is slow
- You are invited to join the NIAC and contribute to NeXus
- Work underway to collaborate/merge with CIF
- More information: http://www.nexusformat.org