The NeXus Data Format

Mark Könnecke,[1] Frederick Akeroyd,[2] Herbert J. Bernstein,[3] Aaron S. Brewster,[4] Stuart Campbell,[5] Bjørn Clausen,[6] Stephen Cottrell,[2] Jens Uwe Hoffmann,[7] Pete R. Jemian,[8] David Männicke,[9] Raymond Osborn,[10] Peter F. Peterson,[5] Tobias Richter,[11] Jiro Suzuki,[12] Benjamin Watts,[13] Eugen Wintersberger,[14] and Joachim Wuttke[15]
(NeXus International Advisory Committee)
[1)] *Laboratory for Development and Methods, Paul Scherrer Institute, 5232 Villigen-PSI, Switzerland*
[2)] *ISIS, Rutherford Appleton Laboratory, UK*
[3)] *imgCIF, Dowling College, USA*
[4)] *Lawrence Berkeley National Laboratory, USA*
[5)] *Spallation Neutron Source, USA*
[6)] *Los Alamos National Laboratory, USA*
[7)] *Helmholtz Zentrum Berlin, Germany*
[8)] *Advanced Photon Source, Argonne National Laboratory, USA*
[9)] *ANSTO, Australia*
[10)] *Materials Science Division, Argonne National Laboratory, USA*
[11)] *Diamond Light Source, UK*
[12)] *KEK, Japan*
[13)] *Swiss Light Source, Paul Scherrer Institute, 5232 Villigen-PSI, Switzerland*
[14)] *Deutsches Elektronen-Synchrotron DESY, 22607 Hamburg, Germany*
[15)] *Forschungszentrum Jülich, JCNS at MLZ, 85747 Garching, Germany*

(Dated: 2 October 2014)

NeXus is an effort by an international group of scientists to define a common data exchange format for neutron, X-ray, and muon experiments. NeXus is built on top of the scientific data format HDF5 and adds domain-specific rules for organizing data within HDF5 files in addition to a dictionary of well-defined domain-specific field names. The NeXus data format has two purposes. First, NeXus defines a format that can serve as a container for all relevant data associated with an experiment. Second, NeXus defines standards in the form of *application definitions* for the exchange of data between applications. NeXus provides structures for raw experimental data as well as for processed data.

Keywords: NeXus, data format, HDF5, X-ray, neutron, data analysis, data management

I. INTRODUCTION

Increasingly, major neutron and X-ray facilities have chosen to store data using the NeXus data format[1,2]. Since 2006, NeXus has undergone substantial refocusing, refinement and enhancement as described in this paper.

Historically, neutron and X-ray facilities have chosen to store data in a plethora of home-grown data formats. This scheme has a number of drawbacks addressed by NeXus:

- It makes the life of traveling scientists unnecessarily difficult as they must deal with multiple files in different formats, file converters, *etc.*, in order to extract scientific information from the data.
- An unnecessary burden is imposed on data analysis software producers to accommodate many different formats.
- The whole idea of open access to data is sabotaged if the data is in a format that cannot be easily understood.
- Scientific integrity is jeopardized if the data cannot be understood or important elements are missing.
- Modern high speed detectors produce data at such a high rate that many older single image storage schemes have become impractical and an efficient container format is a necessity.

The primary necessity for a data format is to define a physical file format: how is the data written to disk? Rather than invent yet another format, NeXus chose HDF5[3] as the physical file format. HDF5, a binary file format in the public domain that is well supported by both commercial and free software tools, is efficient, self-describing, and platform-independent.

NeXus adds to HDF5:

- Rules for organizing domain-specific data within a HDF5 file
- Features to enable rapid data visualization
- A dictionary of documented domain-specific field names
- Definitions of standards that can be validated

The development of NeXus is overseen by a committee, the NeXus International Advisory Committee (NIAC)[4].

II. DESIGN PRINCIPLES

NeXus data files are built using basic HDF5 storage elements, *i.e.*, groups (like file system folders), fields (such as strings, floats, integers, and arrays), attributes (additional descriptors of groups and fields), and links (like file system links), but follow certain design principles to make it easy to navigate even the most complex of HDF5 files. Data and associated metadata are stored as fields within groups that have a logical (and often physical) association with the experiment (see FIG. 1). HDF5 attributes are used to define the types, or classes, of these groups. For example, sample information is stored in a group of class `NXsample`, instrumental information in a group of class `NXinstrument`, *etc.* The beamline components that form the instrument, such as monochromators, collimators, and detectors, are stored as sub-groups within the `NXinstrument` group. This hierarchical structure makes NeXus extremely flexible, capable of accommodating new types of instrument as they are developed, and extremely scalable, capable of storing data from single point-detectors to complex multi detector configurations. It can also, just as easily, contain processed data or even theoretical simulations to be stored alongside the experimental results.

NeXus groups are contained within a root-level group with class `NXentry`. The `NXentry` group contains all the data from a single measurement, which could represent data collected in a certain configuration or in a scan, so multiple measurements can be stored in separate `NXentry` groups within a single file if needed. Each NeXus file is required to contain at least one `NXentry` group.

Each `NXentry` group should contain at least one `NXdata` group, which contains the measured (or processed or simulated) data along with the other information required to plot it, *e.g.*, the plotting axis or axes. The NeXus design allows default plots of `NXdata` groups to be generated without any prior knowledge of the type of measurement. This feature was implemented in NeXus before HDF5 introduced dimension scales, which provide similar functionality.

As well as defining a logical group structure, NeXus provides a dictionary of names that can be used to define specific fields within each class of groups. For example, if the sample temperature is stored, the NeXus standard specifies that it should be called `temperature` and stored in the `NXsample` group. These names are documented in the NeXus base class definitions (Sect. V). It should be stressed that it is not necessary for a particular NeXus file to contain every item defined for each base class; the base classes just define the names that should be used when they are present. However, certain applications may require particular items to be present for specific types of data analysis. For each of these different applications, a specific subset of the standardized NeXus entities (data groups and fields) are standardized in the NeXus application definitions (Sect. VI).

The combination of a well-defined hierarchy of groups with a comprehensive and well-documented dictionary of data and metadata names ensures that NeXus files are self-describing. It should be possible for another scientist to understand the contents of a NeXus file without consulting documentation specific to any one facility or beamline. By enabling the storage of comprehensive metadata, the NeXus format facilitates the sharing of data between collaborators and long-term data curation.

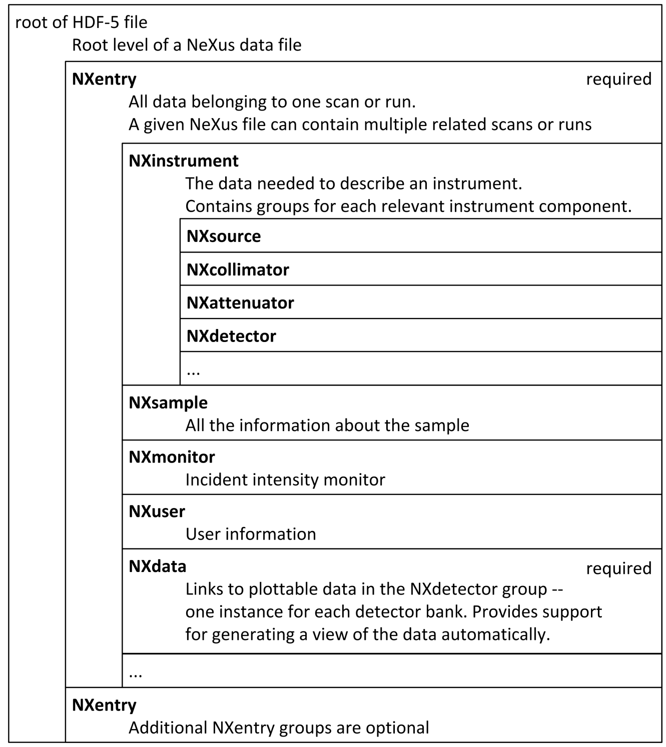

FIG. 1. Overview of the structure of a NeXus raw data file. Note that only a small part of this structure (the first `NXentry` group and the first `NXdata` group) is actually required. The other content is optional.

III. FILE HIERARCHIES

A. Raw Data File Hierarchy

A major focus of NeXus has been the recording of *raw* experimental data, *i.e.*, information taken directly from the experimental equipment or processed only as required to provide physically meaningful values. The NeXus raw data file hierarchy is the consequence of some practical considerations. An overview of the NeXus data file structure for raw experimental data is shown in FIG. 1.

When looking at a beamline, it is easy to discern different components: beam optic components, sample position, detectors, *etc.* It is quite natural to replicate this physical separation with a logical arrangement, in which metadata from each component are stored a separate group. This approach explains the list of beamline components in the `NXinstrument` group presented in FIG. 1. As there can be multiple instances of the same kind of equipment, like slits or detectors, in a given beamline, it becomes necessary to add type information to the group name. This type information, the NeXus class name, is provided by a HDF5 attribute. By convention, NeXus class names start with the prefix `NX`. Each NeXus group describing a beamline component contains further groups and fields describing the component. A field can contain a single number, a text string or an array, as appropriate to the data to be described.

The requirement to store multiple related scans or runs in the same file or to capture a complete workflow in a file causes the beamline component hierarchy to be pushed one level deeper into an `NXentry` group in the hierarchy. The `NXentry` group thus represents one scan or run (or a processed data entry, as will be discussed later). The `NXentry` group also holds the experiment metadata, such as the date and time at which it was performed.

To enable default visualization of the experimental data, at least one `NXdata` group should be provided at the `NXentry` level. It contains the plottable data, or links to the data, which often reside in the `NXdetector` group (links are supported by HDF5 and work like symbolic links in the Unix file system). It also contains information about plot axes, using attributes to define what the data should be plotted against.

The `NXentry` group may also contain one or more `NXmonitor` groups, containing data from beamline monitors. Since they may also contain plottable data, it uses the same attribute scheme to associate the monitor data with its plotting axes. Its location in the `NXentry` group facilitates quick inspection for beamline diagnostics.

Most NeXus files will also contain a `NXsample` group containing information about the sample being measured in the experiment, *e.g.*, its chemical composition, mass, unit cell parameters, *etc.* It may also contain information about the sample environment, such as temperature or pressure. If one or more of these parameters is varied in an experiment, these could be used as scanned variables (see Section III.A).

A special base class, `NXcollection`, exempts its contents from validation and thereby allows inclusion of whatever data in arbitrary non-NeXus formats.

1. Multiple Method Instruments

Some instruments, particularly at X-ray sources, offer multiple techniques that can be used in parallel. For example small-angle scattering and powder diffraction can be measured simultaneously at a SAXS/WAXS beamline. We recommend storing the data from all methods in *one* file, in a *single* `NXentry` hierarchy (FIG. 2). All information from detectors, logs, *etc.*, are collected in this one `NXentry` group to keep the data together. Information that is particular to one experimental technique is linked into a `NXsubentry`. The `NXsubentry` follows the hierarchy of `NXentry`, but it will typically only link to the data required by the application definition for the specific experimental technique. The point of this scheme is that both humans and computerized users can easily locate method-specific data while maintaining the full view of the experiment.

2. Scans

Scans come in all shapes and sizes. Almost anything can be scanned against anything. An additional consideration is that, in practice, the final number of scan points in the scan cannot be known in advance since it is possible that a scan may be interrupted or terminated before its planned number of observations. Thus, it is a challenge

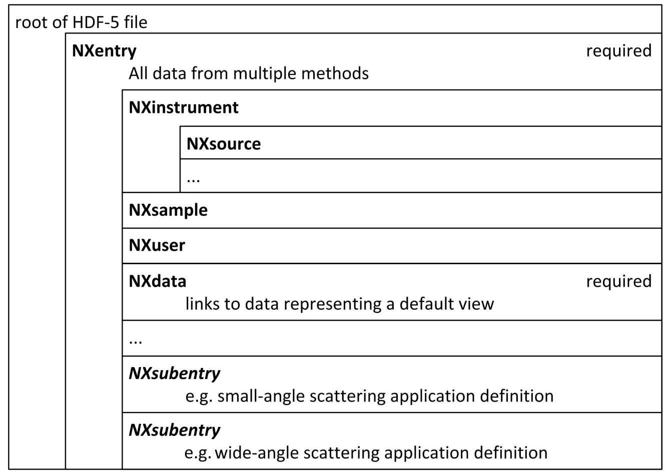

FIG. 2. Overview of the structure of a NeXus raw data file for an instrument with multiple methods.

to standardize a scan. NeXus addresses this challenge through use of HDF5 *unlimited dimensions* and additional conventions as described below. With the HDF5 unlimited dimensions feature, one axis of the data is allowed to expand without limit. Thus, the size of a data array does not need to be declared in advance. Data can be appended to an array along the unlimited dimension as required.

Scans are stored in NeXus following these conventions:

- Each variable varied or collected in the scan is stored at its appropriate place in the NeXus beamline hierarchy as an array. The array's first dimension is the scan axis. This is the unlimited dimension in the implementation and data at each scan point is appended to the array.

- The `NXdata` group holds links to all the variables varied or collected during the scan. This creates something equivalent to or better than the tabular representation people are used to for scans. The main detector data can be plotted against any scanned parameter as well as against everything that was deemed worth recording in addition to that. The necessary data is all gathered together in the `NXdata` group either directly or via links, so that other groups do not normally have to be searched to do this plotting.

NeXus allows multi-dimensional scans too. This makes it very simple to produce meaningful slices through data volumes, whether the software is designed for NeXus (*e.g.*, NeXpy[5]) or NeXus-agnostic (*e.g.*, HDFView[6]).

B. Processed Data

At the request of the user community, NeXus created a simplified structure for storing the result of data processing: be it reduction or analysis. An overview of the NeXus structure for processed data is given in FIG. 3.

The hierarchy is much reduced as it is not important to carry all experimental information in the data reduction.

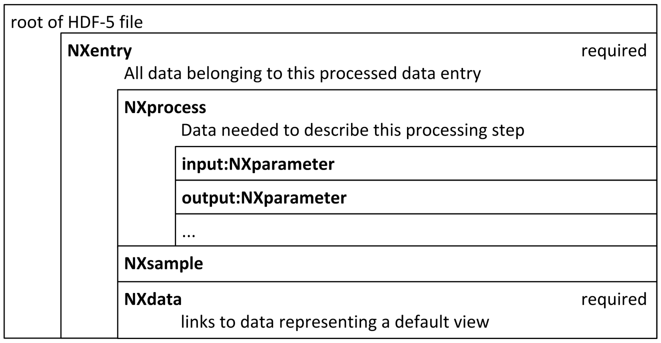

FIG. 3. Overview of the structure of a NeXus processed data file.

In contrast to the raw data file structure, `NXdata` in the processed file structure is the place to store the results of the processing, together with its associated axis or axes.

In addition to the `NXdata` and `NXsample` groups, the `NXprocess` group provides structure to store details about the processing, such as the program used, its version, the date of processing, and other metadata. The `NXprocess` group can hold additional `NXparameter` groups which are containers for storing the input and output parameters of the program used to perform the processing.

IV. COORDINATE SYSTEMS, POSITIONING OF COMPONENTS AND FURTHER RULES

For data reduction, it is often necessary to know the exact position and orientation of beamline components. The first thing needed is a reference coordinate system. NeXus chose to use the same coordinate system as the neutron beamline simulation software, McStas[7].

For describing the placement and orientation of components, NeXus stores the same information as the Crystallographic Interchange Format (CIF)[8]. CIF (and NeXus) stores the details of the translations and rotations necessary to move a given component from the zero point of the coordinate system to its actual position. As coordinate transformations are not commutative, the order of transformations must also be stored.

Detailed consideration has been given to the composition of the NeXus format, with a comprehensive set of rules for creating a NeXus file now being well defined. For the sake of brevity here, a complete description of the format is provided elsewhere in the NeXus manual[9] where a full listing of the rules is presented. However, to illustrate the attention to detail given in the definition of the format, the rules include these considerations:

- How axes are associated with data
- That units must be given with the data
- How data is to be stored in NeXus fields
- How to describe array data which is not in ANSI C storage order

V. BASE CLASSES

As can be seen from the discussion of the NeXus file hierarchy, NeXus arranges data in groups which have a type descriptor and a NeXus base class name associated with them. Technically, the class name is the value of the HDF5 attribute `NX_class`. The term *base class* is not used in the same sense as in object-oriented programming languages; in particular, there is no inheritance. The NeXus base classes provide a comprehensive dictionary of terms that can be used in each class. The terms in the dictionary comprise concepts and names common to the topic of the base class. The expected spelling and definition of each term is specified in the base classes. It is neither expected nor required to provide all the terms specified in a base class. Terms with other names are permitted but might not be recognized by standard software. These terms designate data fields that can be stored within a group. A data field can have a simple type (like integer, float, date/time, binary), or it can be a NeXus subgroup. The base class definition also contains informal annotations about the semantics of each field.

At base class level, NeXus has no mechanism to mark some fields as obligatory. *All allowed fields are optional.* Which of them are written into data files must be decided according to application needs. These decisions can be standardized in the form of application definitions (see below, Sect. VI).

The NeXus base classes are defined in XML files using the NeXus Description Language (NXDL)[9]. NXDL files may be parsed either by people or by software and used to validate NeXus files for syntax and content. GUI tools have been prepared[10] to perform such validation.

VI. APPLICATION DEFINITIONS

An application definition, expressed in NXDL, specifies a data structure for a given application domain such as a scientific technique or a specific type of instrument. The data structure consists of a hierarchy of NeXus groups. For each group, a *minimum* content is specified. Application definitions are therefore different than base class definitions, which specify a comprehensive dictionary of terms that can be used according to the context.

Historically, an application definition addressed one type of instrument, like an X-ray reflectometer or direct-geometry neutron time-of-flight spectrometer. Thus, application definitions were originally named *instrument definitions.* However, the same instrument can be used for different types of analysis that require different experimental variables; *e.g.*, a powder diffractometer could be used for Rietveld refinements or pair-distribution-function analysis. The more generic term *application definition* has been adopted to signify what data are required for each type of data analysis.

VII. CONTRIBUTED DEFINITIONS

The process of drafting and ratifying application definitions is ongoing (see also below, Sect. VIII). Currently, scientists representing both the NeXus International Advisory Committee and the IUCr Committee on the Maintenance of the CIF standard are nearly finished with a NeXus application definition for macromolecular crystallography. CBFlib[11] is being extended to work with NeXus-MX format. This work will be published in another paper. Work on another NeXus application definition for reduced small-angle scattering data is also in progress[12] by scientists representing canSAS, NeXus, and the IUCr Commission on Small-Angle Scattering.

All such proposals from the scientific community to extend NeXus with new application definitions and base classes are added to NeXus, initially, as contributed definitions either in incubation or as a special case not for general use. The NIAC is charged to review any new contributed definitions and provide feedback to the authors before ratification and acceptance.

VIII. GOVERNANCE

The development of NeXus is overseen by the NeXus International Advisory Committee (NIAC)[4]. The NIAC seeks a balanced representation of the international community. Most major neutron, X-ray, and muon facilities have appointed delegates. Other facilities are invited to join.

The NIAC reviews proposed amendments to the NeXus base classes and application definitions, and holds online votes to ratify changes. A great number of candidate NeXus application definitions exist which were derived from our understanding of the technique described. For each of these, the NeXus team seeks community approval.

IX. UPTAKE OF NEXUS

NeXus is already in use as the main data format at many facilities including Soleil, Diamond, SINQ, SNS, Lujan/LANL and KEK. Other facilities including ISIS, DESY, and the μSR community are in the process of moving towards NeXus as their data format. At LBNL, NeXus is currently being adapted for XFEL serial crystallographic data. The APS is using it for some techniques. The EPICS[13] area detector software has a plug-in to write acquired images into NeXus data files. Also, some commercial manufacturers of area detectors now write acquired images into NeXus data files.

The adoption of NeXus has taken some time. The reason is that partly NeXus is often chosen whenever a facility starts operation or undergoes major refurbishments. For those facilities where there is an existing and working pipeline from data acquisition to data analysis, the resources are usually lacking to move towards NeXus as the only data file format.

This is reflected in the experience of the muon community. For the ISIS source, the move to a Windows PC-based data acquisition system in 2002 required a new data format, providing an ideal opportunity to exploit the emerging NeXus standard[14]. In contrast, sources at PSI, TRIUMF and KEK continued to make good use of existing formats and software. More recently, funding from the EU has enabled the community to develop the Application Definition as a common exchange format for muon data[15].

Whether used as the main or an intermediate format, users are able to produce compatible NeXus files for data written across all facilities, enhancing the uptake of NeXus within the community.

X. BACKWARDS COMPATIBILITY

Historically, NeXus supported reading and writing data files in HDF4, HDF5 and XML formats by use of the NeXus Application Programming Interface (NeXus-API or NAPI). The NAPI is still available, but frozen except for bug fixes. After consultation with the community, the currently recommended use of NeXus is solely in terms of the HDF5 file format, using standard HDF5 tools. That is expected to remain the basis for NeXus software development and file creation in the future.

XI. SUMMARY

NeXus has matured considerably over the last ten years and is now in use in many facilities. NeXus is flexible enough to accommodate a wide variety of instruments and scientific applications, yet efficient enough to handle the data coming from modern high speed detectors. More information, including a full PDF manual, can be found on the project website[16]. Members of the NIAC[4] always welcome correspondence concerning the development of the NeXus data format.

ACKNOWLEDGMENTS

The NIAC acknowledges the support of all the institutions contributing to NeXus and their respective funding agencies, most notably DOE, NIH and NSF in the US. Work at Argonne was supported by the Scientific User Facilities Division and the Materials Science and Engineering Division, Basic Energy Sciences, Office of Science, U.S. Department of Energy. The development of the muSR NeXus data format was partly funded by the European Commission within the sixth Framework Programme.

[1] M. Könnecke, Physica B **385-386**, 1343 (2006).
[2] P. Klosowski, M. Könnecke, J. Tischler, and R. Osborn, Physica B: Physics of Condensed Matter **241**, 151 (1998).
[3] *HDF-5* (2014 (accessed July 2014)), `http://www.hdfgroup.org/HDF5/`.
[4] *NeXus International Advisory Committee (NIAC)* (2014 (accessed July 2014)), `http://wiki.nexusformat.org/NIAC`.

[5] *NeXpy: A Python GUI to Analyze NeXus Data* (2014 (accessed September 2014)), http://nexpy.github.io/nexpy/.
[6] *HDFView* (2014 (accessed September 2014)), http://www.hdfgroup.org/products/java/hdfview/index.html.
[7] P. Willendrup, E. Farhi, and K. Lefmann, Physica B **350**, 735 (2004).
[8] S. Hall and B. McMahon, *International Tables for Cystallography Volume G: Definition and exchange of crystallographic data* (Wiley, 2006).
[9] *NeXus Manual* (2014 (accessed July 2014)), http://download.nexusformat.org/kits/definitions/nexus-manual-3.1.0.tar.gz.
[10] *NXvalidate: a java GUI tool to validate NeXus data files* (2014 (accessed July 2014)), https://github.com/nexusformat/code/tree/master/applications/NXvalidate.
[11] H. J. Bernstein and P. J. Ellis, in *International Tables For Crystallography*, Vol. G: Definition and Exchange of Crystallographic Data, edited by S. R. Hall and B. McMahon (Springer, Dordrecht, NL, 2005) Chap. 5.6, pp. 544 – 556, see http://sf.net.projects/cbflib and http://www.bernstein-plus-sons.com/software/CBF/.
[12] *Description of the canSAS2012 data format* (2014 (accessed July 2014)), http://www.cansas.org/formats/canSAS2012/1.0/doc/.
[13] M. Rivers, *areaDetector: EPICS software for area detectors* (2014 (accessed September 2014)), http://cars9.uchicago.edu/software/epics/areaDetector.html.
[14] P. K. D. Flannery, S.P. Cottrell, Physica B **326**, 238 (2003).
[15] S. P. Cottrell, F. Pratt, A. Hillier, P. King, F. Akeroyd, A. Markvardsen, N. Draper, Y. Yao, and S. Blundell, Physics Procedia **30**, 20 (2012).
[16] *NeXus A common data format for neutron, x-ray and muon science* (2014 (accessed July 2014)), http://www.nexusformat.org/.