The Design of HDF5 Data Formats for HEPS

HAOFAN WANG, QI LUO, HAO HU
Computing Center, Institute of High Energy Physics, CAS, 19B Yuquan Road, Shijingshan
District, Beijing, China, (hfwang@ihep.ac.cn)

Introduction

The High Energy Photon Source (**HEPS**) encompasses a variety of experimental types, ***including diffraction, scattering, imaging, and spectroscopy***. The data generated from these experiments are highly dimensional, uncertain, and computationally intensive. Considering the users' needs for interoperable data analysis and high-performance I/O processing, it is necessary to organize and manage the data and metadata efficiently. We have adopted HDF5 and NeXus as the unified data format standards for HEPS. By sequentially ***interfacing with the first phase of 14 beamlines***, we have organized the data formats and summarized common requirements, eventually ***developing three generic data formats***. These serve as ***foundational templates for the subsequent phases involving 90 beamline stations***.

We introduce the characteristics of experimental data types at HEPS. From the perspectives of experimental users and beamline station managers, we detail the organization and specification requirements of various experimental data HDF5 formats designed according to NeXus. Finally, we summarize some challenges and future plans concerning HEPS data formats.

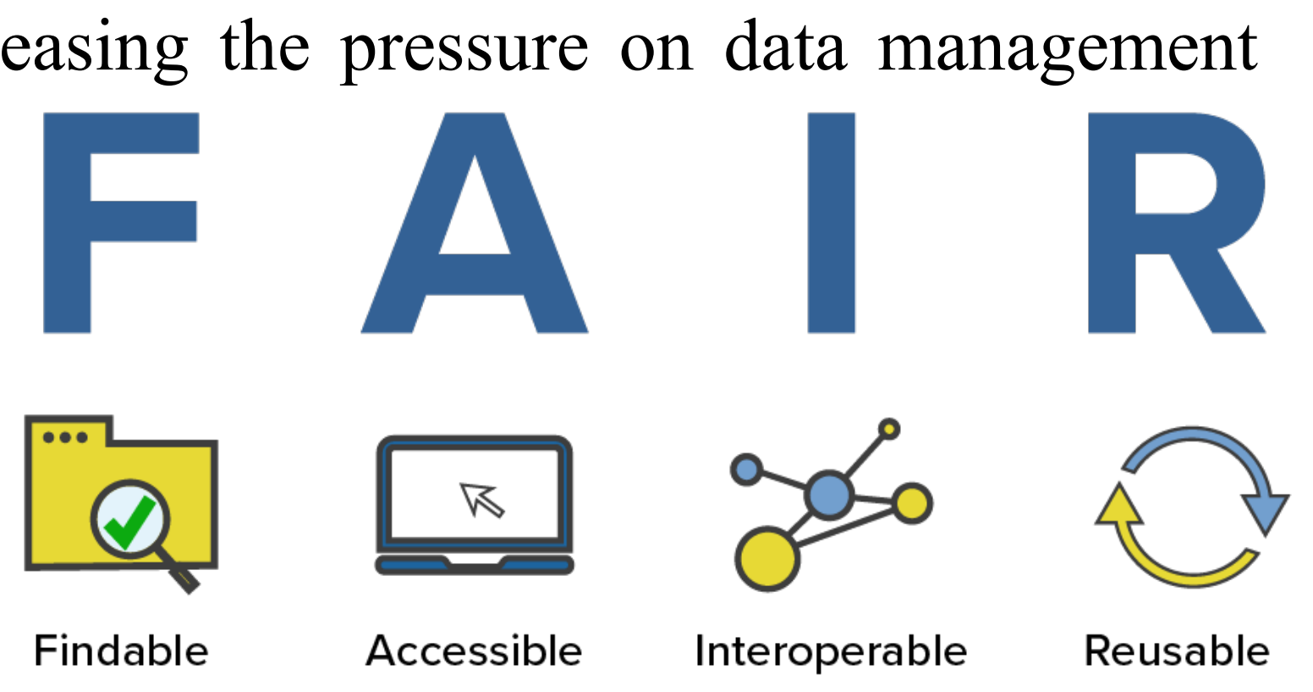
Why does HEPS need a uniform data format?

In the past, there exist many formats of data files produced from experiments carried out at beamlines because of the discipline preference and the limit of detector outputs. This creates several issues for the management and utilization of scientific data:

- Scientific data in multiple formats is difficult to use directly and requires extensive processing.
- Accessing these scientific data requires multiple software tools, making data sharing and utilization very challenging.
- Designing various extractors is necessary, increasing the pressure on data management efforts.

We need to adopt a unified data format that adheres to the FAIR principles, aiming to achieve:

- Promote the standardization of scientific data.
- Enable efficient and precise data management.
- Facilitate the utilization and sharing of data.



Why HDF5 and NeXus?

■ Features of HEPS Beamline Data

- **Diverse Experimental Types**: Including diffraction, scattering, imaging, and spectroscopy, and features high throughput, ultra-fast frequencies, dynamic loading, and diverse experimental methodologies.
- **Rich and diverse metadata**: Covering diverse aspects such as sample details, in-situ environments, and experimental conditions, with the organization of data varying depending on the type of experiment.

■ Features and Advantages of HDF5

- **Efficient Storage and Management**
- **Efficient Data Access**
- **Data Sharing and Interoperability**
- **Rich Metadata Description**

■ Features and Advantages of NeXus

- **Standardized Data Format**
- **Predefined Data Structures**
- **Wide Community Support**
- **……**

HDF5 and NeXus offer advantages in efficient storage, flexible structure, efficient access, data sharing, and rich metadata description, making them ideal choices for data management at HEPS.

Implementation Roadmap

■ Roadmap

1. The data format for beamline stations is collaboratively organized by the data management system, beamline stations, DAQ system, and experimental processing system.

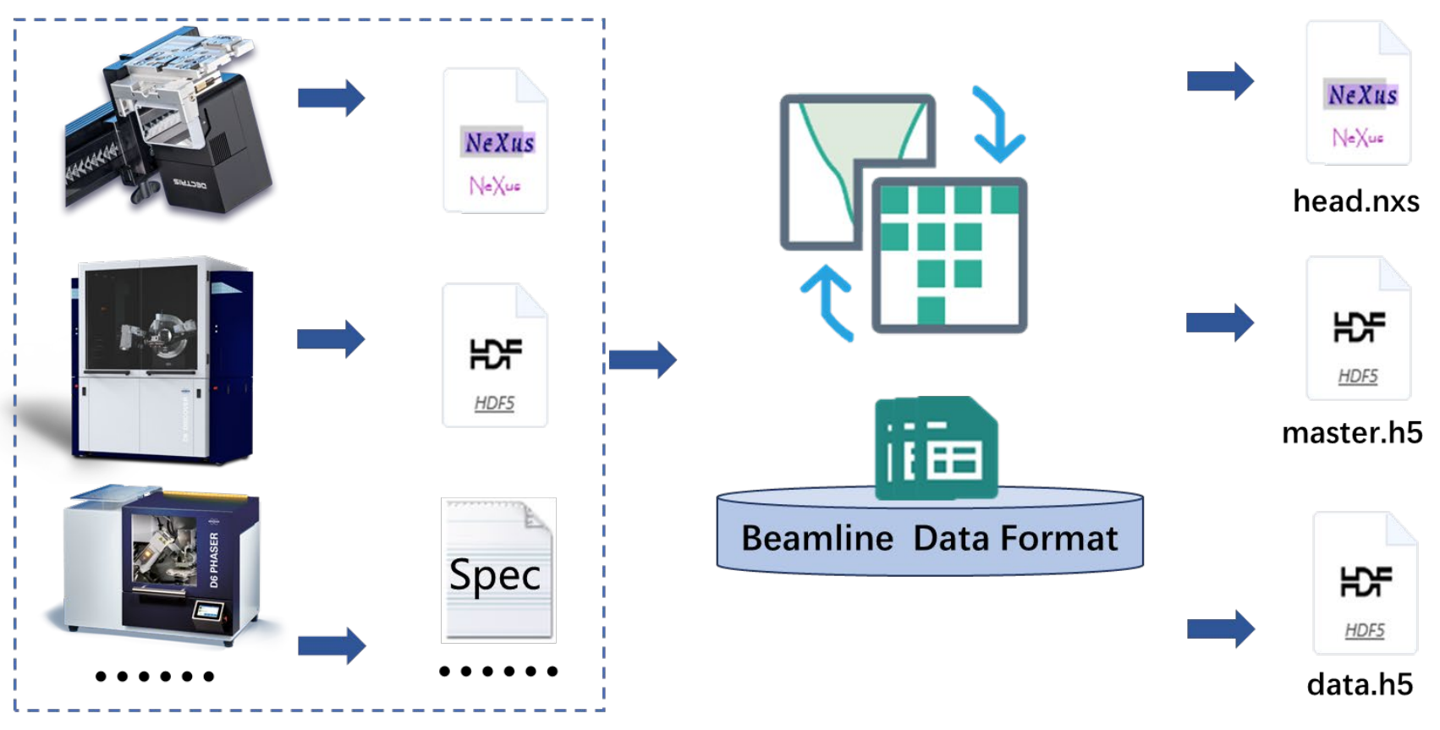
2. Starting with the 14 beamline stations from the first phase of HEPS, which involve methodologies such as diffraction, scattering, imaging, and spectroscopy, we design data formats for each station and summarize common elements to create a general data format template, providing a foundation for future beamline stations.

3. Continuously adjust and update the data format versions based on the equipment and needs of the beamline stations.

4. Design two types of data formats: "Raw Data" and "Processed Data".

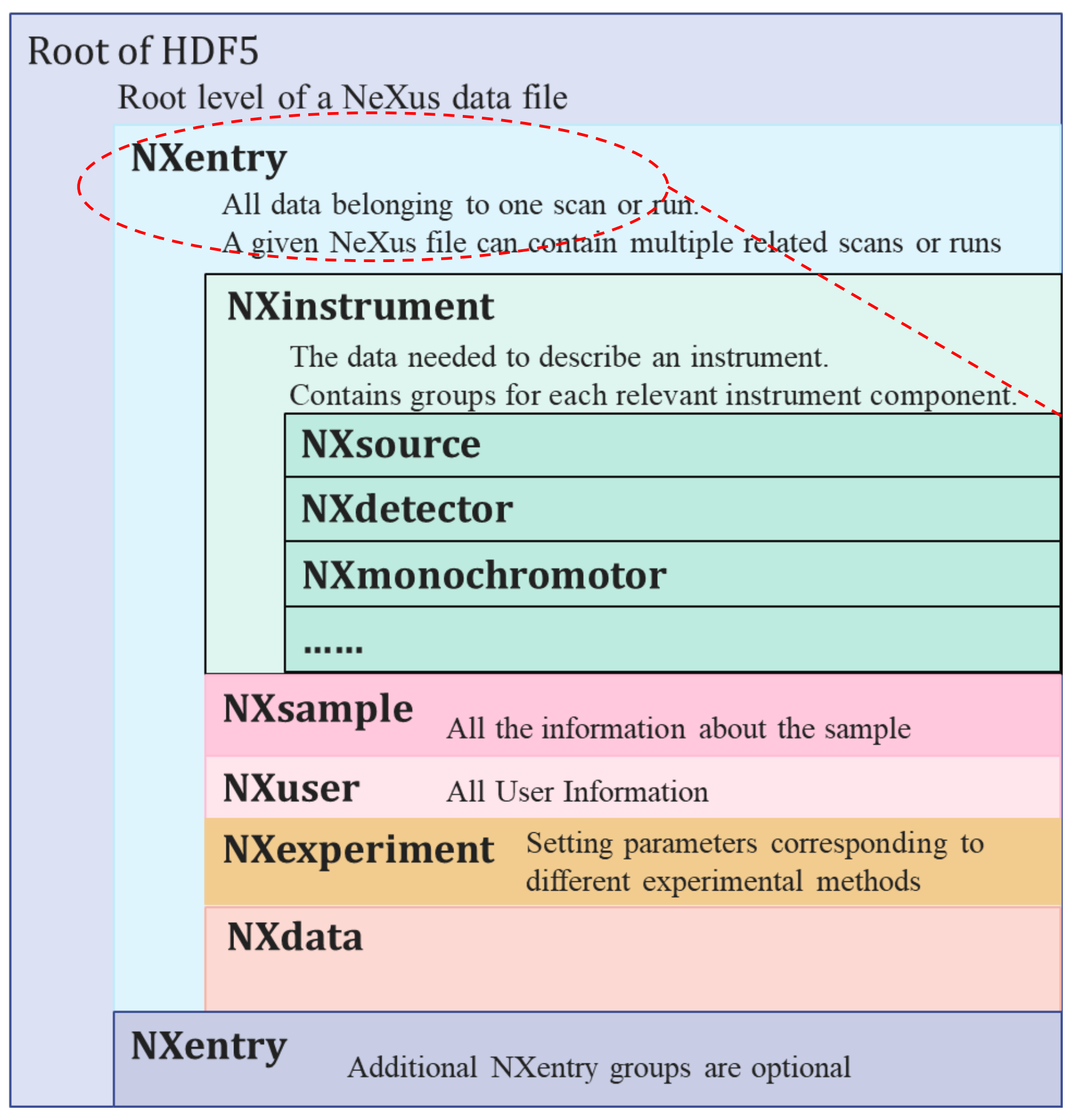
■ Principles for Organizing Raw Data

- Commercial Detectors: Reorganize the generated data files to construct "header.nxs + detector_master.h5 + detector_data.h5".
- In-House Developed Detectors: Directly generate the required "header.nxs + data.h5".



Data Format Design

■ Data Format


Root of HDF5
Root level of a NeXus data file
NXentry
All data belonging to one scan or run.
A given NeXus file can contain multiple related scans or runs
NXinstrument
The data needed to describe an instrument.
Contains groups for each relevant instrument component.
NXsource
NXdetector
NXmonochromotor
……
NXsample
All the information about the sample
NXuser
All User Information
NXexperiment
Setting parameters corresponding to different experimental methods
NXdata
NXentry
Additional NXentry groups are optional

(a) Overview of HDF5 Structure

- The designed HDF5 structure includes information related to the equipment, sample, detector, experiment, and users.
- Under the entry, it primarily contains relevant identifiers and basic information to discern the meaning of the datasets.

beamline	NX_CHAR	Beamline ID
entry_identifier	NX_CHAR	RUN ID
experiment_identifier	NX_CHAR	Proposal ID
experiment_description	NX_CHAR	Experimental methods, e.g BCDI, SAXS, CDI ……
beamtimeID	NX_CHAR	beamtimeID
data_format	NX_CHAR	Data format File name and version
start_time	NX_DATE_TIME	Start Time
end_time	NX_DATE_TIME	End Time
user	NXuser	User Information
instrument	NXinstrument	Instrument Information
sample	NXsample	Smaple Information
experiment(optional)	Nxcollection	Experiment Information

(b) NXentry Information

□ NXinstrument :

- Based on the differences in beamline equipment and experimental station equipment, as well as the distinctions among diffraction, imaging, and spectroscopy types, different sub-levels are designed.
- These may include information on instrument status, source, filter, chopper, CRL, beam attenuator, Laue monochromator, slit, XBPM, sensors (such as vacuum pumps, vacuum gauges, radiometers, oxygen, ozone, and temperature monitors), and hutch environment.

□ NXexperiment :

- It may also include information on the sample stage, detector, sample environment, and experimental methods. However, this type of information may vary with different experiments.

□ NXdata :

- It includes data files and closely related metadata information.
- If some data entities already exist in other files, there is no need to store the information again; access can be achieved through linking.

Summary / Outlook

- ■ Designed the HEPS data format architecture based on HDF5 and NeXus.
- ■ Created data formats for each beamline station, summarized common parameters, and developed universal templates.
- □ Developed data format orchestration software to simplify construction.
  - Quickly select and recommend appropriate NeXus.
  - Manage data format versions.
  - Automatically generate HDF5 example files.

Acknowledgements

We wish to acknowledge the support and assistance provided by the beamline scientists, experiment control and data acquisition experts, and experiment users at HEPS for their contributions to the data format work.

DOMAS, or Data Organization Management Access Software, is a comprehensive framework designed to streamline the complex challenges of data management for advanced light source facilities. It offers a suite of four core modules: metadata catalogue, metadata acquisition, data transfer, and data service, which together automate the organization, transfer, storage, and distribution of scientific data. By providing common basic modules and universal interfaces, DOMAS allows for the rapid establishment of a data management system with just parameter configuration and minimal code development. This not only addresses the common demands and unique characteristics of advanced light sources but also offers a significant advantage by reducing the time and resources required to build a customized data management scheme.