Report on the Integration of HDF4 Mapping Schema with Existing Standards

Ruth Aydt (aydt@hdfgroup.org)

Peter Cao (xcao@hdfgroup.org), Ruth Duerr (rduerr@nsidc.org),
Mike Folk (mfolk@hdfgroup.org), Christopher Lynnes (christopher.s.lynnes@nasa.gov)

As part of a NASA-funded project to improve long-term preservation of EOS data by independently mapping HDF4 data objects, a study of existing schemas was conducted to determine what existing schema, if any, would be appropriate for use in the project. METS, PREMIS, MIX, XFDU, ESML, CSML, and NcML were reviewed as possible candidates for adoption or inspiration. This report summarizes the overall conclusions as well as specific observations of these standards/schemas as they relate to the HDF4 Independent Mapping Project.

1 Overall Conclusions

After careful study, it was concluded that the existing schemas were developed for purposes different from those of the HDF4 Independent Mapping Project. While there was some overlap in goals, and consequently in the information included in the XML documents described by the schema, *none of the existing schemas met all the needs of the mapping project*. Furthermore, it was not clear that one could start with an existing schema and extend it in a way that satisfied the needs of the mapping project without paying a significant cost in terms of clarity and ease-of-use. With this in mind, *a new HDF4 File Content Map Schema was designed that is tailored to the project goals, should harmonize with PREMIS[1], and leverages terminology and approaches from the schemas that were studied.*

2 Project Requirements and Goals

A number of requirements and goals enumerated for the HDF4 Independent Mapping Project were relevant in the evaluation of the existing standards / schemas. They are summarized here:

1. The XML documents described by the schema must provide complete access to the content in the binary HDF4 files held by NASA's Earth Observing System. To the extent possible given the time constraints of the project, and without introducing excessive complexity, the XML documents described by the schema should provide access to all content originally stored in binary HDF4 files, including content stored using objects, representations, or compression schemes that are not found in NASA's EOS collection of binary HDF4 files.

[1] A review of Version 1.0.0 of the HDF4 File Content Schema should be done by a PREMIS expert to make sure that it does indeed harmonize with PREMIS.

2. The XML documents described by the schema should be usable by a person twenty or more years in the future who is interested in the content originally stored in the binary HDF4 files. The person:
    - will have access to the binary HDF4 files
    - will not necessarily know anything about the HDF4 data model or file format, and may not have access to any HDF4 documentation or software
    - will have very basic knowledge of XML
    - may not have access to the schema that describes the XML documents
3. A present-day data repository may extract preservation metadata about binary HDF4 files from the XML documents described by the schema.

As a consequence of these requirements and goals, the XML documents produced by the mapping project and described by the schema must contain preservation, structural, technical, and descriptive metadata about the companion binary HDF4 file that is being mapped.

More concretely in relationship to goals 1 and 2, there must be sufficient information in the XML document for the person in the future to:

- understand the data objects in the HDF4 file and their relationship to each other
- retrieve and correctly interpret bytes in the binary HDF4 file representing data that has not been migrated into the XML document

While the focus of the project is on making the contents of the binary HDF4 file accessible in the future, the XML document produced will be of little use if the companion binary HDF4 file that is mapped becomes corrupt. For that reason, a minimal amount of information about the binary HDF4 file (as opposed to information about the file content) is also needed in the XML document.

3 Existing Standards / Schemas

Several existing schemas (METS, PREMIS, MIX, XFDU, ESML, CSML, and NcML) were investigated. Most of these are currently in use by the Earth science community, and could therefore offer the benefits of having an established user base that is familiar with the standards.

3.1 METS (Metadata Encoding & Transmission Standard)

3.1.1 URL and Brief Description

http://www.loc.gov/standards/mets/

"The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library…"

3.1.2 Summary of Findings

The focus of the METS standard is on the exchange of digital materials, and not on preservation of those materials or on how to interpret their contents. While it does allow for extensions in the techMD (technical metadata) section, there is not much to build on for expressing the HDF4 data objects, their relationships to each other, and the data they hold.

Two extensions to METS that have been endorsed are PREMIS and MIX.

3.1.3 Concepts applied to HDF4 File Content Map Schema

No concepts from METS were applied directly.

3.2 PREMIS (Preservation Metadata: Implementation Strategies)

3.2.1 URL and Brief Description

http://www.loc.gov/standards/premis/

"The PREMIS Data Dictionary defines what a preservation repository needs to know." (*Understanding Premis* by Priscilla Caplan for the Library of Congress, available at http://www.loc.gov/standards/premis/understanding-premis.pdf)

3.2.2 Summary of Findings (based on version 2.0)

Many of the concepts and terms used in the PREMIS Data Dictionary, especially those related to Intellectual and Object Entities, were helpful for thinking about the activities and designing the products of the mapping project.

PREMIS Object Entities include Files and Bitstreams—both relevant for the mapping project. However, the PREMIS Data Dictionary and schema did not include elements to naturally describe the HDF4 data objects and their relationships to each other, or to express the combination and interpretation of different bitstreams in an HDF4 file to retrieve data values. This is not surprising, as PREMIS explicitly excludes metadata that apply to only one file format or class of formats.

PREMIS does provide an extensibility mechanism that allows for format-specific technical preservation-related metadata. MIX used this mechanism.

3.2.3 Concepts applied to HDF4 File Content Map Schema

Several PREMIS semantic units for the Object Entity are reflected in the HDF4 File Content Map Schema. In particular, the elements that provide information about the HDF4 file and external files referenced by the HDF4 file include name, location (type and value), size, and MD5 checksum. The primary purpose of these elements in the context of the map schema is to provide a rudimentary mechanism for a reader of the map file to locate the files referenced by the XML File Content Map document and verify that they are not corrupt. Our intent was that these elements and values conform to the requirements and constraints associated with the corresponding semantic units in the PREMIS Data Dictionary without introducing unnecessary complexity into the HDF4 File Content Map documents.

The PREMIS bitstream concept formed the underpinning for the byteStreamT and byteStreamSetT types in the HDF4 File Content Map schema.

3.3 MIX (NISO Metadata for Images in XML Schema).

3.3.1 URL and Brief Description

http://www.loc.gov/standards/mix/

"… an XML schema for a set of technical data elements required to manage digital image collections."

3.3.2 Summary of Findings (based on version 2.0)

Although not related to Earth Science data, the MIX schema was studied as part of our investigation as an extension to METS and PREMIS. It was especially helpful because it defines format-specific metadata, while harmonizing with PREMIS and being endorsed by METS.

3.3.3 Concepts applied to HDF4 File Content Map Schema

Several concepts and element names from MIX were adapted for use in the HDF4 File Content Map schema in an attempt to adopt familiar terminology whenever possible. These include fileSize, byteOrder, Compression, imageWidth and imageHeight, and the "filepath" designation for location type. The overall structure of the MIX schema also influenced the design of the HDF4 File Content Map Schema.

3.4 XFDU (XML Formatted Data Unit)

3.4.1 URL and Brief Description

http://public.ccsds.org/publications/archive/661x0b1.pdf

XFDU is a standard being developed by the Mission Operations and Information Management Area Data Archive Ingestion Working Group (MOIMS-DAI) of the Consultative Committee for Space Data Systems (CCSDS).

"This Recommended Standard defines a technique for the packaging of data and metadata, including software, into a single package (e.g., file or message) to facilitate information transfer and archiving. … This Recommended Standard leverages the wide community acceptance and usage of XML technologies by making the packaging manifest an XML document defined by the XML Schema specified in the document. "

3.4.2 Summary of Findings

The XFDU schema is targeted at a coarser granularity of packaging than that of the HDF4 Independent Mapping project. Within the dataObject element, some concepts such as byteStream and transformObject (used for compression, as one example) do have relevance in the context of retrieving bytes from a file and manipulating them to get the correct data values. The XFDU schema does not address objects within a data model and their relationships to each other.

3.4.3 Concepts applied to HDF4 File Content Map Schema

No concepts from XFDU were applied directly. However, the Layered Information Model presented in ANNEX E of CCSDS 650.0-B-1, "Reference Model for an Open Archival Information System (OAIS), available at http://public.ccsds.org/publications/archive/650x0b1.pdf, was very helpful in thinking about the layers of information that must be conveyed by a HDF4 File Content Map. Basically, information at the Object, Structure, and Stream Layers was all relevant to the HDF4 Independent Mapping project.

3.5 ESML (Earth Science Markup Language)

3.5.1 URL and Brief Description

http://esml.itsc.uah.edu/

"ESML is an interchange technology that enables data (both structural and semantic) interoperability with applications without enforcing a standard format within the Earth science community. Users can

write external files using ESML schema to describe the structure of the data file. Applications can utilize the ESML Library to parse this description file and decode the data format."

3.5.2 Summary of Findings (based on version 3.0)

The ESML schema focuses on describing the data in a file but does not include sufficient information for a reader of an ESML XML document to retrieve and process the data from the file without reliance on the ESML library. EMSL's BinDatum types align well with HDF4's simple datatypes, but there is no provision for non-IEEE floating point, middle-endian data, or compressed values in the current schema. ESML's Array and Field elements could be used to represent HDF4's SDS and Vdata, although it was not clear how Vdata with multiple entries per Vdata field would be represented, or how HDF4's Rasters, Palettes, and Attributes would be represented. The HDF4 Vgroup did not have a counterpart in ESML.

Overall, we felt that representing the HDF4 data model and storage structures would require significant extensions to the fairly generic element definitions in the existing ESML schema, and may result in schema that was more cumbersome than necessary. While the ESML authors were responsive to questions, there appears to have been little active work on the project for several years and the size of its existing user community is uncertain, limiting the potential gains achieved by adopting an existing standard.

3.5.3 Concepts applied to HDF4 File Content Map Schema

The ESML notion of a field as the lowest (atomic) level of data was very pertinent to HDF4, and EMSL's BinDatumDef complex type is reflected heavily in the HDF4 File Content Map's datumTypeT. The term Array is more intuitive than "SDS", and was adopted in the Mapping project, for objects that contain data of uniform type.

3.6 CSML (Climate Science Modelling Language)

3.6.1 URL and Brief Description

http://csml.badc.rl.ac.uk/

"CSML is a standards-based data model and GML (Geography Markup Language) application schema for atmospheric and oceanographic data with associated software tools developed at the Rutherford Appleton Laboratory."

3.6.2 Summary of Findings (based on version 2)

CSML has two basic components. The first provides a conceptual model of feature types such as Points, Grids, and Swaths and the second is a mechanism for mapping file-based data into the features. They are connected through the use of XLink (http://www.w3.org/TR/xlink).

While the concepts of the first component are relevant to HDF EOS data, they are at a higher level than the HDF4 data model and not presented explicitly as part of the HDF4 Independent Mapping project. CSML does not offer "features" that align well with the HDF4 data model objects (Vdata, Groups, SDSes, Rasters, Palettes).

The CSML StorageDescriptor does have overlap with the project, and contains information such as arraySize, numericType, and fileExtract. However, as was the case with ESML, the schema does not include sufficient information for a reader of a CSML XML document to retrieve and process the data from the file without reliance on a software library that is file-format aware.

3.6.3 Concepts applied to HDF4 File Content Map Schema

CSML's idea of values being presented "inline" in the XML document is reflected in the name of one of the HDF4 File Content Schema map documents.

While there is some segregation of model and storage in the mapping project, it was decided that for the most part having storage information co-located with model object (feature) instances would be easier for the user of the future, especially given that we are dealing with only one underlying file format. CSML's distinct handling of these two reinforced the belief that they are fundamentally different types of information.

3.7 NCML (NetCDF Markup Language)

3.7.1 URL and Brief Description

http://www.unidata.ucar.edu/software/netcdf/ncml/

"NcML is an XML representation of netCDF metadata, ... NcML is similar to the netCDF CDL (network Common data form Description Language), except, of course, it uses XML syntax."

http://www.unidata.ucar.edu/software/netcdf/ncml/v2.2/AnnotatedSchema.html

"An NcML document represents a generic netCDF dataset, i.e. a container for data conforming to the netCDF data model… An NcML document therefore should not necessarily be thought of as a physical netCDF file, but rather the 'public interface' to a set of data conforming to the netCDF data model. "

3.7.2 Summary of Findings (based on version 2.2)

The netCDF data model is widely used in the Earth Science community, so the use of the NcML schema is very attractive from a user-familiarity point of view.

NcML presents the netCDF data model, which has considerable overlap with HDF4, but which is not identical.

- The NcML netcdf element corresponds to the HDF4 file
- The NcML group element corresponds to the HDF4 Vgroup object. However, NcML only allows hierarchical trees while HDF4 allows cycles. (Cycles have not been seen in any EOS files)
- The NcML variable element is a container for data and can contain other variables to build more complex structures. Variable elements, with some well-chosen attributes, could probably be used to describe HDF4 Vdata, SDS, Raster, and Palette objects.
- The NcML attribute element is similar to HDF4 Attributes and Annotations, although it can only be associated with NcML's netcdf, group, and variable elements, while HDF4 also allows Attributes and Annotations on dimensions as well.

The Basic NcML Tutorial (http://www.unidata.ucar.edu/software/netcdf/ncml/v2.2/Tutorial.html) states:

> The ***NetCDF Markup Language (NcML)*** is an XML dialect that allows you to create CDM datasets. An ***NcML document*** is an XML document that uses NcML, and defines a CDM dataset. Commonly, the NcML document refers to another dataset called the ***referenced CDM dataset***. The purpose of NcML is to allow:

1. Metadata to be added, deleted, and changed.
2. Variables to be renamed, added, deleted and restructured.
3. Data from multiple CDM files to be combined (aka "aggregated").

The emphasis of NcML on enabling changes to the referenced dataset, and aggregations of data from multiple files, results in many elements in the schema that are not relevant to the HDF4 Independent Mapping Project.

While the NcML schema provided a good basis for understanding the objects in the HDF4 file and their relationship to each other, it did not provide sufficient information for a reader of an NcML instance document to retrieve and correctly interpret bytes in the binary HDF4 file. Rather, it relied on software (The NetCDF-Java/CDM library) that is file-format aware to perform that function.

3.7.3 Concepts applied to HDF4 File Content Map Schema

Several of the element names in the NcML schema are reflected in the HDF4 Schema. These include Group, Attribute, Dimension, and values (for data values included inline in the XML file).

On the other hand, we felt that having unique elements for the HDF4 Vdata (Table), SDS(Array), Raster, and Palette objects would make the HDF4 File Content Maps clearer for the future reader than relying on a single Variable type with special attributes to distinguish among them. This was especially true given that these objects can have different underlying storage representations in the HDF4 file.

3.8 Summary of Existing Standards and Schema

Of the standards examined, METS, PREMIS, and XFDU are designed more for sharing and preserving files (or bytes in files) than for describing data models that might give structure to those files (or bytes). The MIX schema focused on a technical metadata for a different domain (Digital Still Image), but was instructive in its specification and integration with METS and PREMIS. ESML, CSML, and NcML reflected underlying data models that overlapped with the HDF4 data model to varying degrees, but all relied on file-format-aware software to extract and interpret bytes from the binary (or ASCII) files and did not include sufficient information in the Schema for this to be done without the software. Since all of these schemas provided a common model view to a number of underlying file formats (including HDF4), this reliance on file-format-aware software is not surprising.

4 Additional Information

Additional information about the HDF4 Independent Mapping Project, including the HDF4 File Content Map Schema, is available at: http://www.hdfgroup.org/projects/h4map/

Acknowledgements

This work was supported by a Cooperative Agreement with the National Aeronautics and Space Administration (NASA) under NASA grant NNX06AC83A. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NASA.