

EXPRESS Data as HDF5 Mapping Specification Version 0.2

This document defines the mapping of a subset of EXPRESS-driven data into HDF5.

Table of contents

1 Scope.....	2
2 Normative References.....	2
3 HDF5 in a Nutshell.....	3
4 EXPRESS Data Represented as HDF5.....	4
4.1 Approach.....	4
4.2 HDF5 Path Names.....	5
4.3 The Population of EXPRESS Schemas as HDF5.....	5
4.4 EXPRESS Schema Information as HDF5.....	7
4.5 EXPRESS Entity Instances as HDF5.....	10
4.6 EXPRESS Datatype Values as HDF5.....	10
4.7 EXPRESS Array Datatype Values as HDF5.....	11
4.8 EXPRESS Aggregate Datatype Values as HDF5.....	12
5 Examples.....	13
5.1 test1_line example.....	13
6 EXPRESS-HDF5 Mapping Issues.....	13
6.1 Initial Issues from V 0.1 development.....	13
6.2 Issues Against V 0.1.....	13
7 Relationship to ISO 10303-21 Part 21 Clear text encoding.....	14
8 Bibliography.....	14

Warning:

Version 0.2 is simply a second initial proposal. The mapping, including its scope, may change significantly as the EXPRESS community gains HDF5 experience. However, the scope did not change between Version 0.1 and Version 0.2.

1. Scope

The following are within the scope of this specification. The scope will be expanded in future revisions of the mapping.

- The mapping of entity instances into HDF5.
- The mapping of Entity and Attribute into HDF5.
- The mapping of Enum into HDF5.
- The mapping of all simple datatypes into HDF5 except binary.
- The mapping of defined types that specialize other defined types or EXPRESS simple types.
- The mapping of n-dimensional arrays of a single type into HDF5.
- The mapping of simple single-dimensional lists, sets and bags of a single type are included into HDF5.

The following are outside the scope of this specification.

- rules/functions
- abstract entity
- complex numbers
- The mapping of inverse attributes into HDF5.
- The mapping of derived attributes into HDF5.
- The mapping of some redeclaration into HDF5.
- The mapping of Defined types of arrays into HDF5.
- The mapping of Select (some include both Entity and DT of array), into HDF5.
- The mapping of multidimensional lists, sets and bags into HDF5.
- The mapping of arrays, lists, sets and bags not of a single type into HDF5.
- The mapping of the EXPRESS binary data type.

2. Normative References

The following normative references are applicable for this specification.

ISO 10303-11:2004, Industrial automation systems and integration - Product data representation and exchange - Part 11: Description methods: The EXPRESS language reference manual.

HDF5 Release 1.6.4, March 2005 [cited 2005-04-09]. Available from World Wide Web: http://hdf.ncsa.uiuc.edu/HDF5/doc/RM_H5Front.html.

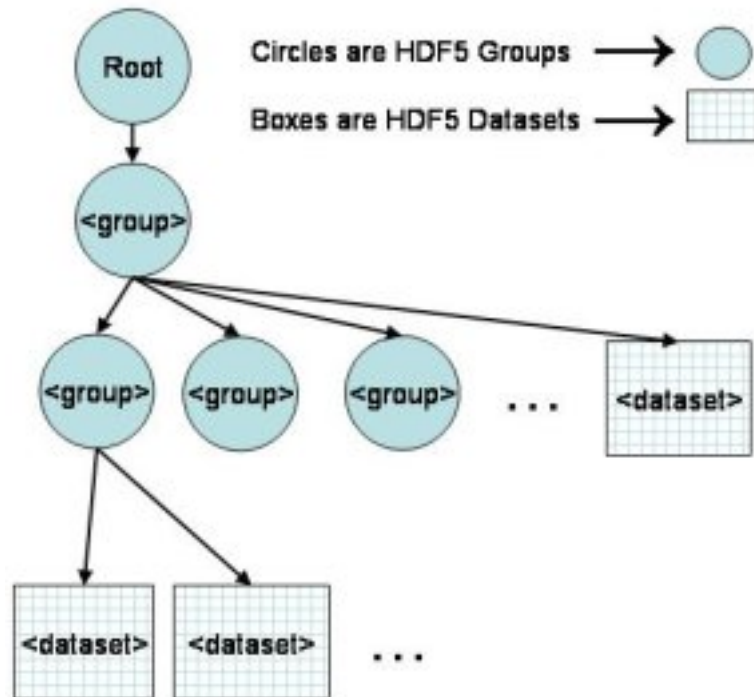
HDF5 User's Guide, HDF5 Release 1.6.4, March 2005 [cited 2005-10-07]. Available from World Wide Web: <http://hdf.ncsa.uiuc.edu/HDF5/doc/UG/index.html>.

3. HDF5 in a Nutshell

For those less familiar with HDF5 the following, it's perhaps simplest to start ignoring some of the complexity and focusing on a simple use of its basic architecture as follows:

- An HDF5 file can contain data and objects of many types (logical groupings of data, datatype definitions, bitmap images, large arrays of data, etc).
- An HDF5 Group is a logical structure within an HDF5. HDF5 Groups contain other HDF5 Groups and HDF5 Datasets.
- The topmost HDF5 Group is called the Root Group and there is one Root Group per HDF5 file.
- An HDF5 Dataset contains the data in the file. An HDF5 Dataset contains the definition of HDF5 Datatypes, the definition of the dimensions of the array it contains called an HDF5 Dataspace and the data array itself.
- An HDF5 Datatype can be a simple type (e.g. integer) or a compound type similar to a C struct and can be named.
- An HDF5 Dataspace is the definition of the dimensions of the array and also contains information about how it is stored (e.g. is it compressed or not).
- HDF5 Groups and HDF5 Datasets can have HDF5 Attributes attached to specify some characteristic of the Group or Dataset.

The above statements are incomplete but give the reader some hint of how HDF5 works. The following diagram shows these concepts and a convention used for diagrams in this specification. HDF5 Groups appear as circles and HDF5 Datasets appear as boxes. The arrows point from the containing concept to the contained concept (e.g. Groups can contain Datasets).



HDF5 Concepts Diagram

A key resource for the understanding of the HDF5 technology and how it relates to this specification is the HDF5 Users Guide: Chapter 1 HDF5 Data Model. Implementors and reviewers are encouraged to read that document before continuing with this specification.

4. EXPRESS Data Represented as HDF5

This section describes the representation for EXPRESS-driven data represented using HDF5. This representation is specified by relating EXPRESS data concepts to HDF5 data concepts, not by specifying any particular application programming interface (API). HDF5 implementations are available in multiple programming languages. This specification does not specify the use of any particular HDF5 API.

A summary presentation of the mapping is also available : EXPRESS/HDF5 Mapping Version 0.2 Specification Summary.

4.1. Approach

A set of basic propositions underly the approach for representing EXPRESS-driven data using

HDF5 in this specification. Those basic propositions follow.

- HDF5 software manages arrays well both in memory and on disk to satisfy the EXPRESS-driven data requirements.
- Maximizing the use of HDF 5 structures to enable the use of its optimizations is needed to satisfy performance and file size requirements.
- Only the datatypes used for the writing of the data on disk are specified, nothing is stated about the representation of that data in memory.
- It is preferable to use the HDF5 pre-defined data types where possible.
- It is preferable to allow flexibility of representation which puts somewhat more burden on developers of software that reads HDF5 files.
- Cross-platform interoperability (e.g. write in Unix C and read in Windows Fortran) needs to be supported.

The general approach for representing EXPRESS-driven data using HDF5 is to treat instances of EXPRESS entity types and EXPRESS array instances in a similar manner. The set of instances of an EXPRESS entity type along with any non-array attribute values are treated as a dataset based on an HDF5 compound data type. EXPRESS Array-valued attributes are represented using an HDF5 reference to an HDF5 dataset for each array instance.

4.2. HDF5 Path Names

HDF5 Groups, Datasets and Named Data Types used in this mapping are identified by an HDF5 Path. For HDF5 constructs mapped directly from EXPRESS constructs, the EXPRESS identifiers are used as part of the construction of the names of the HDF5 Paths. The case of the EXPRESS identifiers is preserved. The dot (".") character is used to separate EXPRESS names.

For example,

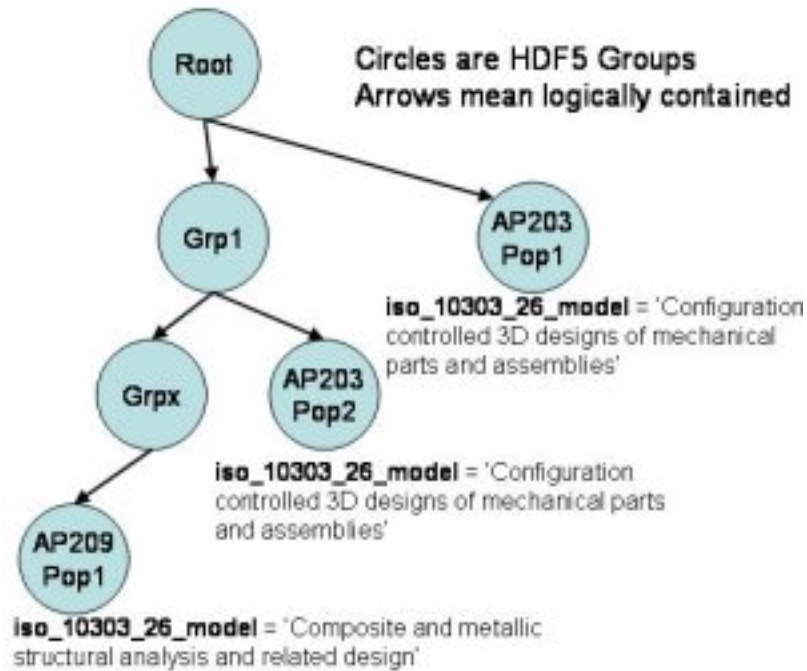
```
SCHEMA pets; ENTITY Dogs;
```

would result in part of the HDF5 Path name for the related HDF5 Dataset being "pets.Dogs". In the remainder of this document, this convention appears as <schema_id> + "." + <entity_id>.

4.3. The Population of EXPRESS Schemas as HDF5

For the purposes of this specification, EXPRESS schemas are considered to be defined for the purpose of constraining the validity of data populations. Multiple populations based on the same EXPRESS schema are possible and may be included in the same HDF5 file. As more than one way to represent EXPRESS-driven data values based on the same schema in HDF5, no single representation is required for any EXPRESS schema. The same HDF5 file may

contain populations of an EXPRESS schema that use different encodings for the data based on that EXPRESS schema. Information about the HDF5 constructs used to encoding the EXPRESS-driven data is stored in the HDF5 file along with the data itself. The following figure shows three HDF5 Groups containing data based on this specification.



Populations in HDF5 Groups

Each population of an EXPRESS schema is represented as an HDF5 Group named any `<user_defined_population_name>`. That HDF5 Group shall have the following HDF5 Attributes associated with it:

- `iso_10303_26_model` : which has a data value of the `<schema_id>`

The `iso_10303_26_model` HDF5 Attribute is the indicator to any software application that the HDF5 Group contains data encoded based on this specification.

That HDF5 Group may also optionally have the following standardized HDF5 Attributes associated with it:

- `iso_10303_26_description` : which has a data value of the `<user_defined_population_description>`
- `iso_10303_26_timestamp` : which has a data value of the corresponding to the extended format for the complete calendar date as specified in 4.2.1.1 of ISO 8601 concatenated to the extended format for the time of the day as specified either in 4.3.1.1 or in 4.3.3 of ISO

8601. The date and time shall be separated by the capital letter T as specified in 4.4.1 of ISO 8601. The alternate formats of 4.3.1.1 and 4.3.3 permit the optional inclusion of a time zone specifier (e.g. 2005-04-12T15:27:46-05:00)

- iso_10303_26_author : which has a data value of the <user>
- iso_10303_26_organization : which has a data value of the <user_organization>
- iso_10303_26_originating_system : which has a data value of the <software_system_name>
- iso_10303_26_preprocessor_version : which has a data value of the <software_application_and_version>

Warning:

The iso_10303_26 prefix assumes that this specification will eventually be standardized as ISO 10303-26. If that is not the case, then the prefix will change.

The contents of that HDF5 Group are specified in the remainder of this specification.

4.4. EXPRESS Schema Information as HDF5

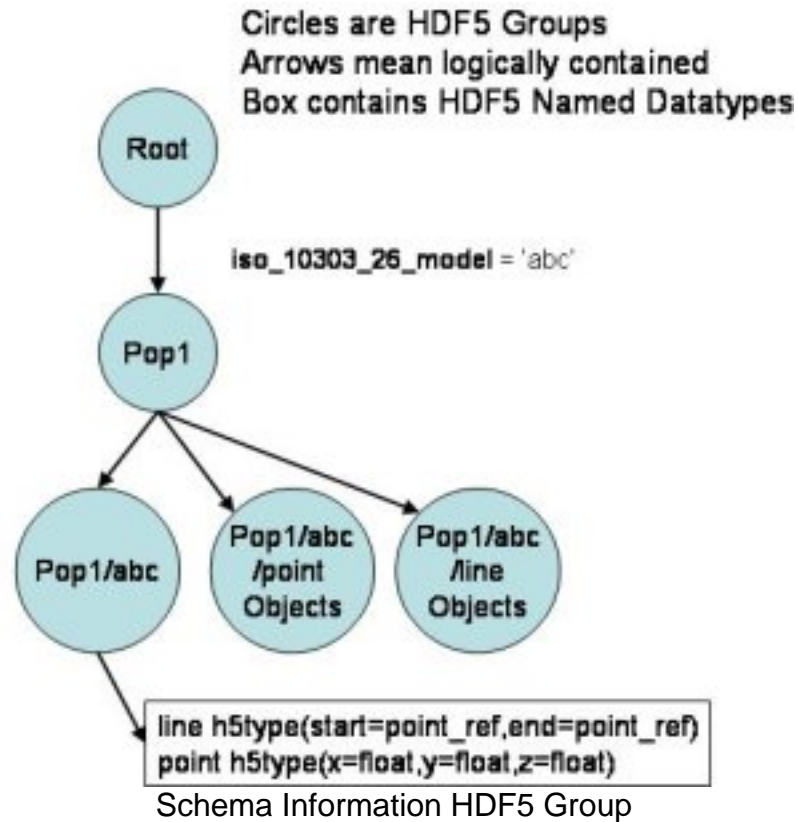
For any encoding of EXPRESS-driven data as HDF5 a single EXPRESS schema must be chosen to provide the context for the data. That EXPRESS schema may use other EXPRESS schemas in its definition. For the purposes of using HDF5 to represent EXPRESS-driven data, those used schemas only appear in the encoded data in cases where their name is required to be concatenated with an EXPRESS entity type name in order to make the name unique.

An HDF5 Group shall be used to represent the EXPRESS schema for a population. The local name of the HDF5 Group shall be as follows

<user_defined_population_name> + "." + <schema_id>.

The parent of that HDF5 Group shall be the HDF5 Group representing the population as described in Population of EXPRESS Schemas as HDF5

The following figure shows that one HDF5 Group contains the HDF5 Named Data Types related to the EXPRESS Schema for the population.



4.4.1. EXPRESS Defined Type Information as HDF5

This section specifies the mapping of most EXPRESS defined types into HDF5.

For EXPRESS defined types that are specializations of EXPRESS simple datatypes, an HDF5 Named Data Type shall appear within the HDF5 Group representing the EXPRESS schema for the particular population being represented. The local name of the Named Data Type shall be as follows:

```
<user_defined_population_name> + "." + <schema_id> + "." + <type_id>
```

.

The Datatype of the Named Data Type is the HDF5 representation of the base type of the defined type as mapped Datatypes.

For example, the following defined type in a schema named "s" for a population named "p"

```
TYPE x : REAL;
```

would map to an HDF5 Named Data Type with name = "p/s/x" with a base of HDF5 IEEE Floating Point 32 or 64 bits, Big- or Little-Endian.

The mapping for EXPRESS defined types that specialize other defined types works in a similar way. A new HDF5 Named Data Type is defined with a base datatype of the other HDF5 Named Data Type.

For example,

```
TYPE y : x; TYPE x : REAL;
```

would map to an HDF5 Named Data Type with name = "p/s/x" with a base of HDF5 IEEE Floating Point 32 or 64 bits, Big- or Little-Endian and an HDF5 Named Data Type with name = "p/s/y" with a base datatype the Named Data Type with name = "p/s/x".

The HDF5 Named Data Types that represent EXPRESS defined types are defined so that they have a parent which is the representation of the EXPRESS schema as specified in Schemas.

4.4.2. EXPRESS Enumeration Types as HDF5

EXPRESS enumeration types are mapped in the same way as other defined types in that they are represented as an HDF5 Named Data Type. However, the Datatype is specified directly as an HDF5 ENUM which is <string>:<integer> pairs where string is <enum_name> + "-" + <enum_literal>.

EXPRESS Enumerations values for attributes are mapped in the section on Datatypes.

4.4.3. EXPRESS Entity Type Information as HDF5

This section specifies the mapping of EXPRESS entity type information into HDF5.

For EXPRESS each entity type, an HDF5 Named Data Type, that is an HDF5 Compound Type, shall appear within the HDF5 Group representing the EXPRESS schema for the particular population being represented. The local name of the HDF5 Named Data Type shall be as follows:

```
<user_defined_population_name> + "." + <schema_id> + "." +  
<entity_id>
```

.

The EXPRESS entity type information being represented shall include information about all explicit attribute, including all inherited attributes. The details of this representation are as follows.

- The type of each attribute of the EXPRESS entity type, including inherited attributes, is

represented as a member (or Field) of the HDF5 Compound Type.

- The name of each member of the HDF5 Compound Type is the name of the EXPRESS explicit attribute, preserving the case.
- The type of each member of the HDF5 Compound Type is an HDF5 Datatype corresponding to the EXPRESS type of the EXPRESS attribute as specified in the Datatypes, Arrays and Aggregates sections of this specification.

4.5. EXPRESS Entity Instances as HDF5

The set of EXPRESS entity instances of the same EXPRESS entity type(s) are represented using an HDF5 Group. That HDF5 Group shall be a child of the HDF5 Group containing the population of the EXPRESS schema. That HDF5 Group contains one HDF5 Dataset which contains the data.

The local name the HDF5 Group containing the entity instances shall be as follows.

```
<user_defined_population_name> + "." + <schema_id> + "." +  
<entity_id> + "." + "Objects"
```

The HDF5 Group contains an HDF5 Dataset containing all the entity instances of the entity type. It also contains an HDF5 Dataset for each EXPRESS aggregate-valued attribute of that entity type (see Array Datatype Values).

The local name the HDF5 Dataset containing the entity instances shall be as follows.

```
<user_defined_population_name> + "." + <schema_id> + "." +  
<entity_id> + "." + "Instances"
```

The HDF5 Dataset that contains the data is based on the HDF5 Named Data Type representing the EXPRESS entity type as specified in Entity Type Information as HDF5. That HDF5 Dataset must also have an associated HDF5 Dataspace to define its rank and dimension. The HDF5 rank is be one, a single dimensional array contains the entity instances. The HDF5 dimension depends on the number of entity instances and on how the HDF Dataset is stored and so its value is not specified here.

The EXPRESS entity instance identifier used to refer to an entity instance as a data value is represented as an HDF5 Region Reference that includes the index of the entity instance in the containing HDF5 Array.

4.6. EXPRESS Datatype Values as HDF5

The following table specifies the representation of data values for EXPRESS simple and enumeration datatype values using HDF5 (see HDF5 Datasets).

EXPRESS Datatype Value	HDF5 Representation
INTEGER	HDF5 Standard 8, 16, 32 or 64 bits, Signed or Unsigned, Big- or Little-Endian
REAL	HDF5 IEEE Floating Point 32 or 64 bits, Big- or Little-Endian
BOOLEAN	HDF5 ENUM of the <string>:<integer> pairs BOOLEAN-TRUE:1, BOOLEAN-FALSE:0
LOGICAL	HDF5 ENUM of the <string>:<integer> pairs LOGICAL-TRUE:1, LOGICAL-FALSE:0, and LOGICAL-UNKNOWN:-1
STRING	Variable-length datatype with base type H5T_NATIVE_UNICODE (see Issues)
BINARY	(see Issues)
NUMBER	same as EXPRESS REAL representation
ENUMERATION	HDF5 ENUM of <string>:<integer> pairs where string is <enum_name>-"<enum_literal>

Table 1: Summary of Mapping of EXPRESS Datatype to HDF5

4.7. EXPRESS Array Datatype Values as HDF5

A array that is the value of an EXPRESS attribute is represented as an HDF5 Dataset containing an HDF5 Array that contains the actual data. In HDF5 all elements of an Array must be of the same HDF5 type. The number of dimensions of the Array must also be specified (HDF5 calls this the "rank") and the dimensions themselves must be specified.

The HDF5 rank shall be an integer specifying the the nested-ness of the Array and the HDF5 dimensions are the number of elements of the array.

The EXPRESS array is referenced using an HDF5 Dataset Reference. This HDF5 Dataset Reference is the value stored for the EXPRESS attribute value, it must be dereferenced in order to read the elements of the array.

Warning:

It is left to the application reading and writing HDF5 encoded data to handle the differences between the HDF5 Array index and the EXPRESS ARRAY index.

For example,

```
ARRAY[1:2] OF ARRAY[1:3] OF INTEGER
```

would have an HDF5 rank of 2 and dimensions 2 and 3.

The following table specified the representation of N-dimensional array values of the same EXPRESS datatype.

EXPRESS N-Dimensional Array Value Datatype	HDF5 Representation
INTEGER	H5T_ARRAY of base type for EXPRESS INTEGER
REAL	H5T_ARRAY of base type for EXPRESS REAL
BOOLEAN	same as for INTEGER
LOGICAL	same as for INTEGER
STRING	to be done
BINARY	to be done
NUMBER	same as for REAL
ENUMERATION	same as for INTEGER

Table 1: Summary of Mapping of EXPRESS Array to HDF5

4.8. EXPRESS Aggregate Datatype Values as HDF5

Single-dimensional EXPRESS LIST, BAG and SET Datatypes that have a basic type that maps to the same HDF5 Datatype are mapped identically to single-dimensional EXPRESS ARRAYS.

Warning:

It is left to the application reading and writing HDF5 encoded data to handle the differences between the HDF5 Array index and the EXPRESS LIST, SET and BAG list position.

As the dimensions of many SETs, LISTs and BAGs are defined by the content of the population of data, not the schema, the dimensions of the HDF5 Array cannot be specified here. Instead, those dimensions must be set based on the data itself.

For example,

```
LIST[2:?] OF LIST[2:?] OF INTEGER
```

would have an HDF5 rank of 2 but its dimensions cannot be set from the information in the schema. If the data included two lists each containing three integers then the dimensions would be set to 2 and 3.

5. Examples

Providing test data for a binary data representation is difficult. For that purpose, this specification provides several simple test datasets using an XML format and related HDF5 DTD.

The details of the DataFromFile XML element are not specified in any detail in the DTD. In the sample datasets provided with this specification the following rules are followed.

1. CompoundType data is enclosed in "{" and "}".
2. ArrayType data values are enclosed in "[" and "]"

Example EXPRESS schemas and schema fragments, EXPRESS-driven data samples, and data samples encoded as HDF5 XML using the HDF5 DTD[2] follow.

5.1. test1_line example

test1_line EXPRESS

test1_line sample HDF5 XML data

6. EXPRESS-HDF5 Mapping Issues

This section describes the issues in mapping between ISO EXPRESS and HDF5.

6.1. Initial Issues from V 0.1 development

1. The handling of STRING need further study. NATIVE types may make data not interoperable across platforms.
2. What exactly does HDF5 BITFIELD mean and is it OK for EXPRESS binary?
3. We need to allow including raster files, etc.
4. In STEP there are references to external files so how would we make that reference if the external file was another dataset inside the same HDF5 file?
5. Do we want to base this work on SDAI or be based on native HDF5?

6.2. Issues Against V 0.1

An issues log against this V 0.1 mapping will be created and documented here so if you have issues, please send them to the team.

1. How do I know which HDF5 Groups in the file contain EXPRESS-driven data? Proposed Resolution : added iso_10303_26_model HDF5 attribute.
2. How are EXPRESS entity instance identifiers represented? Proposed Resolution : use HDF5 Region References. Note that a second approach of using a compound identifier

containing the HDF5 Dataset name and an integer index into that Dataset is worth investigation.

3. How is optionality addressed? Open Issue

7. Relationship to ISO 10303-21 Part 21 Clear text encoding

This section briefly explains how Part 21 and this specification are related.

Part 21 Concept	HDF5 Concept
Data Section	The HDF5 Group containing a data population based on a single schema (see the section on populations).
Header Section Attributes	HDF5 Attributes associated with the HDF5 Group containing a data population based on a single schema (see the section on populations).
Part 21 object Identifier #<n>	HDF5 Region Reference (see the section on entity instances).
Part 21 aggregate groups (...) and ((...) , (...))	Aggregates have identifiers in HDF5. An HDF5 Dataset contains the N-dimensional aggregate values (see the section on array values).

Table 1: Part 21 Concepts Mapped to HDF5 Concepts

8. Bibliography

[1] Introduction to HDF5, March 2005 [cited 2005-04-09]. Available from World Wide Web: <http://hdf.ncsa.uiuc.edu/HDF5/doc/H5.intro.html>.

[2] The HDF5 XML Information Page, March 2005 [cited 2005-04-09]. Available from World Wide Web: <http://hdf.ncsa.uiuc.edu/HDF5/XML/>.