**Title: HDF5 File families**

# S-100 Maintenance - Change Proposal Form (Draft)

| | | | |
|---|---|---|---|
| **Organisation** | Raphael Malyankar | **Date** | 06-Feb-2019 |
| **Contact** | Raphael Malyankar | **Email** | raphaelm@portolansciences.com |

## Change Proposal Type *(Select only one option)*

| 1.Clarification | 2.Correction | 3.Extension |
|---|---|---|
| | | X |

## Location (*Identify all change proposal locations)*

| S-100 Version No. | Part No. | Section No. | Proposal Summary |
|---|---|---|---|
| 4.0.0 | 10c | 17.5 (new) | Add rules and description of metadata for HDF5 “file families”. |
| | 10c | 18.1, 18.2 (new) | Add implementation guidance related to HDF5 “file families”. |
| | | | |
| | | | |

## Change Proposal

*This change proposal provides for the splitting up of a logical HDF5 data file into multiple physical files. Some product specifications (e.g., S-102) may use HDF5 “file families” to break up a logical data file into several physical data files. This might be done to break up datasets into pieces for easier distribution. If this is done, there needs to be a common way for discovery metadata to indicate that a logical file has been split up and to describe digital signatures.*

*Addition of a second metadata attribute providing the size of individual physical files was discussed:*
*fileMemberSize: PositiveInteger [0..1]: the size in bytes of each file member of a an HDF5 family.*
*However, the discussion concluded that it was not needed.*

*The dataset discovery metadata in the exchange set catalogue tentatively corresponds to the logical file, not to each physical file. The related considerations are:*

*(1) Validation of an exchange set that contains a file family requires metadata indicating the total number of physical members. Since an exchange set can in theory contain multiple datasets each of which is broken up into a file family, this indicator needs to be in the dataset discovery metadata block.*
*(2) If there is a separate discovery metadata block for each member, the number of files would have to be repeated in the discovery metadata block for each member of the file family.*
*(3) The physical data files will have the same discovery metadata except for file name (which will differ only in the _x suffix) and digital signature.*
*(4) Making discovery metadata correspond to a logical file instead of a physical file means validators must check if a discovery metadata block is for a single file (in product specifications*

*which do not use file families) or for a file family, by checking for the presence of a metadata attribute indicating the number of members.*

*Digitally signing the family as a collection is based on the assessment (derived from the openssl 1.1.1 manual page at www.openssl.org) that the openssl commands for signing and verifying a data file allow signing (or verifying) a collection of files in the same command.*

*Reasoning that such product specifications are likely to be a minority of S-100-based product specifications, the prescribed extensions are framed in terms of content and metadata attributes that those product specifications which use file families must add to their documentation and schemas, and not as common S-100 material and metadata which all product specifications include by default. This is so that product specifications which do not use file families (presumably the majority) are not forced to specify that the "file family" extensions do not apply.*

*The S-100 WG is invited to discuss this argument and endorse limiting the additional "file family" material to only those product specifications which need it.*

## 10c-17.5 File families

### 10c-17.5.1 Use of file families

Product specifications may use HDF5 "file families" to break up a logical data file into several physical data files. This might be done to break up datasets into pieces for easier distribution. The names of files in the file family are derived from the base name of the logical HDF5 file. Given the logical file "myfile.hdf5" the first file in the family "myfile_0.hdf5" contains the index for the logical file as well as the first of the data from the dataset.  The other files, named "myfile_1.hdf5, myfile_2.hdf5, etc." contain data.

Product specification developers should note that since a common reason for file families is to break up very large datasets into more manageable pieces, the product specification may need to manage other aspects so as to permit such a break-up of large datasets.

### 10c-17.5.2 Metadata for file families

If file families are allowed, the corresponding dataset discovery element in the exchange catalogue describes the logical HDF5 file, not the physical files – that is, even though the exchange set may contain more than one physical file for the logical dataset, there is only one dataset discovery element for the whole collection of file family members for that logical dataset.

Product specifications which allow exchange sets to include HDF5 file families must add a metadata attribute in dataset discovery metadata to indicate the number of file family members in an exchange set. This attribute serves a dual purpose – indicating that the HDF5 "file family" is used, and indicating the number of physical files for the logical dataset. The metadata attribute must be defined as specified in Table 10c-X.X below:

**Table 10c-X.X Additional discovery metadata for HDF5 file families**

| Role name | Name | Description | Mult. | Type | Remarks |
|---|---|---|---|---|---|
| numFamilyMembers | Number of file family members | The number of HDF5 file family members in which the logical HDF5 file is divided in this exchange set. | 0..1 | PositiveInteger | If numbering starts with 0, the value will be 1 more than the highest suffix for this file family. |

Note that this attribute must be added in the metadata clauses of each product specification that wishes to use HDF5 file families – it is not included in common metadata for all S-100 product specifications described in Parts 4a and 4b. This means, for example, that neither S-101 nor S-111 (assuming S-111 does not permit file families) will include this attribute in their discovery metadata.

The digital signature of a file family must be generated from the entire collection of files in their natural sequence (that is, the command to generate the signature must list the members of the file family in the order 0, 1, 2, and so on). Product specification authors should note that the signature in an exchange set will therefore depend on the number of physical file family members into which the logical data file is broken up and will change if the logical file is split into a different number of physical files.

The *fileName* metadata attribute of dataset discovery metadata must name the logical file (for example, “myfile.hdf5”, not “myfile_0.hdf5”).

### 10c-18 Implementation guidance

*[Add the following clauses.]*

### 10c-18.1 Processing of file families

To open the "file family", with h5dump or other tools that come from the HDF Group pass the filename "myfile_%d.hdf5".

For application developers, the suggested way to open an HDF5 file that uses the file family property is described below:

1) Create a file access property list.
2) Modify it to use the file family feature.
3) Pass the modified property list to H5Fopen.
4) Close the property list.
5) Continue working in HDF5.

### 10c-18.2 Validation of exchange sets which include file families

Validators must check if a discovery metadata block is for a single file (e.g., in product specifications which do not use file families) or for a file family, by checking for the presence of the *numFamilyMembers* metadata attribute described in clause 10c-17.5.2. Note that if a logical file is split into an HDF5 file family, there will be one dataset discovery metadata block in the exchange catalogue XML for each **logical** file in the exchange set, not one for each **physical** file.

EXAMPLE: An exchange set containing a logical file split into five physical files “myfile_0.hdf5” … “myfile_4.hdf5”, will have only a single dataset discovery metadata block that names the logical file “myfile.hdf5”. It will have attribute *numFamilyMembers*=5.

## Change Proposal Justification

Some application areas may need to utilize "file families" when using HDF in order to be able break a logic file into several physical files. The amount of data ranges from several Gb to several Tb.  In order to support a wide range of customers, these applications use the "file family" concept to break the data (HDF) into pieces for easier distribution.

This proposal provides a common specification for application areas and product specifications that need to break up data into file families.

The prescribed extensions are framed so that they are used only in product specifications which use file families, so as not to compel other S-100 product specifications to specifically exclude the “file family” extensions.

## What parts of the S-100 Infrastructure will this proposal affect?

☐ S-100 Feature Concept Dictionary Interface or Database
☐ S-100 Portrayal Register
☐ S-100 Feature Catalogue Builder
☐ S-100 Portrayal Catalogue Builder
☐ S-100 UML Models

**Please send completed forms and supporting documentation to the secretary S-100WG.**