

Specification version 0.1

2018-03-13

Revision 17 (changes since 15 are marked blue)

revision	on behalf of	description
17	Dolphin / xSimulation	added attribute "branceNumber" to signal set. This speeds up reconstruction of simulation branch data.
16	signal values in video replay	added attributes "maximum", "minimum", "mean" and "standardDeviation" to signal. This speeds up loading of signals corresponding to video.

- 1 Introduction
 - 1.1 Intended usage
 - 1.2 Supported features
- 2 Glossary
- 2.1 HDF5 Dependencies
- 3 H5M Structure
 - 3.1 Group "/" (root)
 - 3.2 Group Signal Set
 - 3.3 Signal : Dataset
- 4 H5M Data Descriptions
 - 4.1 Naming conventions
 - 4.2 Attributes
 - 4.3 Data types
 - 4.3.1 Default value
 - 4.3.2 Strings: favour unicode (UTF-8) over ascii
 - 4.3.3 iso_fmt
 - 4.3.4 unit_str
 - 4.3.5 prop_str
 - 4.3.6 obj_ref
 - 4.3.7 ds_type

1 Introduction

H5M is an acronym for **HDF5 MARIN Datasets File**.

The data which it contains can have any available source at MARIN: Basin measurements, simulator runs, CFD calculations, full scale measurements and post-processing cq analysis of this data.

Important: Reading or writing H5M requires use of HDF5 libraries. Please check [HDF5 Dependencies](#) for version compliance.

This convention specifies how the content of a HDF5 file is organised for MARIN. For more information about HDF5 itself and for generic HDF5 viewers please visit <https://support.hdfgroup.org/>.

The glossary of terms is described in the next chapter. The data structure and their properties are specified in chapter 3. Details of the data types used and the followed naming convention are specified in chapter 4.

1.1 Intended usage

MARIN Internal use:

- Exchange and access results of the measurements in the basins to data analysts for further processing. This contains the raw data from MMS2 and supersedes the MMS format.
- Exchange the results of simulations and calculations.
- Exchange the results of analysis and other post-processing of raw data between data analysts and between departments.

MARIN External use:

- Exchange processed results of measurements or calculations or simulations or experiments or analysis with clients. (Portal, e-mail, ftp, ...)

1.2 Supported features

The specified format supports the following features:

1. Identification of the software and system configuration used to write the data set.
2. Identification of the process step(s) that generated the data set. (i.e. experiment number, project number, etc)
3. Identification of the experiment variables (a.k.a. signals) and – if applicable - the sensors used.
4. Grouping of signals with shared origin or properties. This is called a signal set.
5. Link a signal set to another signal set to create a dependency of one to another.
6. Time branching: The available Signal Sets can form a graph from which a Signals can be composed by following a specific path.

2 Glossary

Experiment	In the context of this convention, any action that results in one or more signals that need to be stored. E.g. MMS2 Measurement, ReFresco Calculation, XMF Simulation, SHARK data analysis.
H5M	MARIN convention of structuring the content of a HDF5 container for Experiment Datasets.
Base Signal Dimension Signal Master Signal	Independent value range. E.g. the timestamp values for a time series signal that correspond with each sample value.
Signal	Measured, calculated or simulated variable; range of values.
Signal Set	Set of signals grouped by one or more common properties. E.g. origin, experiment run, sample rate, sample count, timestamp.
Dependant Signal	Dependent value range. E.g. the sample values at each corresponding timestamp for a time series signal or the RAO value at certain values on several frequency axis.
A / O / NS (see attributes)	A: always (must be present) and always with a valid value – critical attribute NS: must be present, but may be “not specified” – not critical, but users should be aware if missing O: may be omitted, only specify if value is meaningful – optional
<data_type> []	Represents an array of the type <data_type> . A scalar is represented only by <data_type>

2.1 HDF5 Dependencies

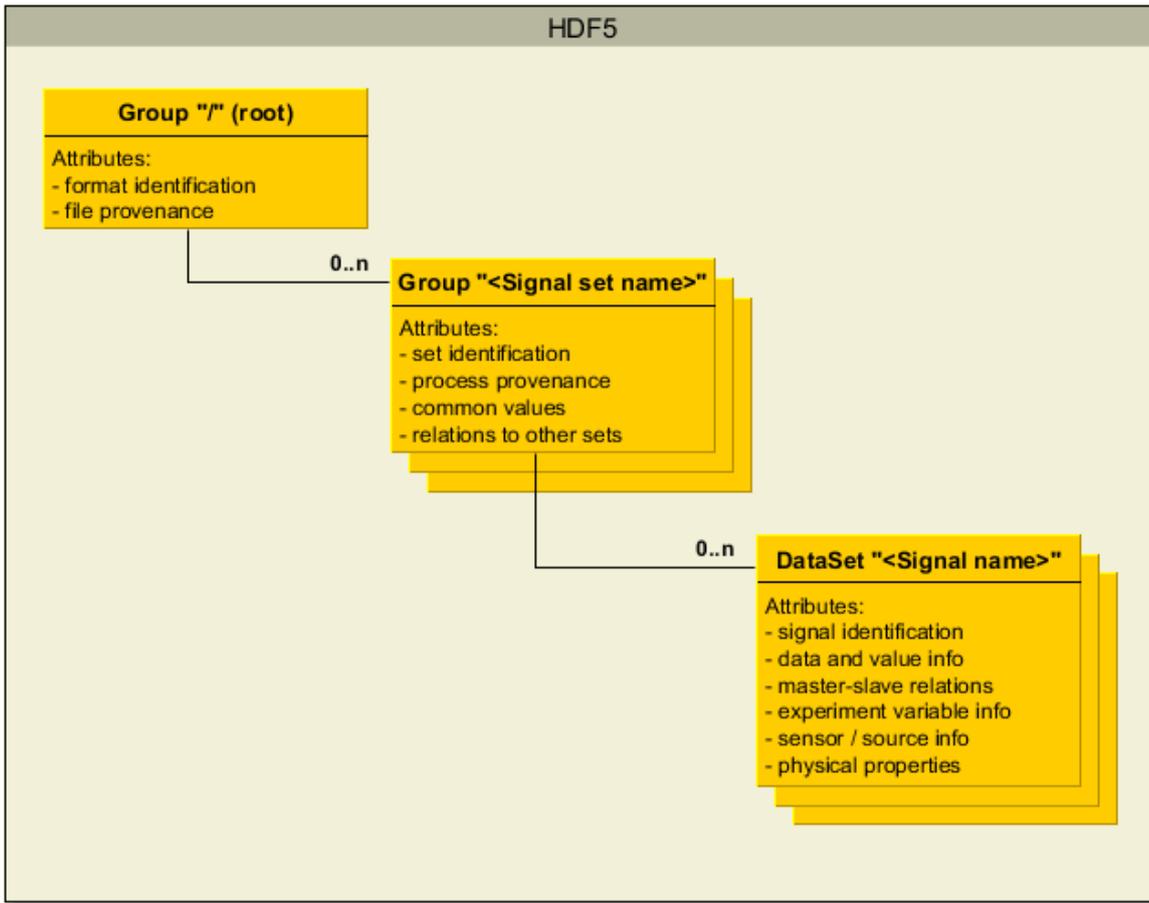
The current version of the HDF5 libraries to be used is **1.8.18**. This might change to 1.8.20 in the near future.

The upgrade to 1.10 has been postponed because the hdf5 object ID datatype is incompatible between v1.10 and v1.8. Files of version 1.8 can be read by a library of version 1.10. On the contrary a file version 1.10 **cannot** be correctly read by a library of version 1.8. The upgrade is pending until all h5m readers (eg, ReFRESKO-XMF on linux cluster) made the step to 1.10 (or the new 1.12).

3 H5M Structure

The H5M content is structured as depicted in the diagram below.

Each tree node item will be explained in the next paragraphs in this chapter. The location of the item is defined by it's address in the HDF5 file, called path. Also all available attributes are specified per item.



3.1 Group "/" (root)

The root group contains:

- attributes that identify the used H5M format.
- zero or more HDF Groups for each set of signals.
- attributes that identify the provenance of the file.

Name: root

Path: "/"

Attributes:

Name	Role	Data type	Exists? A / O / NS	Description	Convention / example
name	i	utf8	A	Name of the convention; constant	"H5M"
description	i	utf8	A	Description of the format; constant.	"HDF5 MARIN Datasets File"
version	a	utf8	A	H5M Format version number	"0.1"
documentation	i	utf8	A	Where to find this convention	" https://mods.marin.nl/display/H5M "
hdf5Version	i	utf8	A	Version of HDF5	"1.8.19"
libraryName	i	utf8	A	Name of the library that performed the actual writing. In case of appending signal sets only the last editor is tracked.	"pymarin", "Marin. Experiments.IO.H5M"
libraryVersion	i	utf8	A	Version of this library	e.g. "7.0.1"

applicationName	i	utf8	NS	Name of the application that wrote this file	"SHARK", "SHARK: some vistrail.vt"
applicationVersion	i	utf8	NS	Version of this application used.	"0.0.20"
systemName	i	utf8	O	Name of the system running the application. This not necessarily the system that performed the experiment (e.g. measurement).	"LP3138", "MMS2"
systemVersion	i	utf8	O	Version or configuration of this system	(tbd)
dateTimeOfCreation	i	iso_fmt	A	Date and time of the moment the file was created.	ISO 8601 "2017-09-27 T21:13:00.012345"
userName	i	utf8	NS	Name of user / author	"user123"
notes	i	utf8	NS	Any additional notes at file level.	(free text)
writeErrors	i	utf8[]	O	Specifies any errors that occurred while writing the file..	

3.2 Group Signal Set

The signal set group contains a set of signals that are logically a single group. Such signals share common properties like number of samples and sample rate. Which properties are common has to be determined by the consumer of the data; either a human or an application.

Name: HDF Group name of the set of signals.

Path: "/<signal set name>"

Attributes:

Name	Role	Data type	Exists? A / O / NS	Description	Convention / example
type	a	utf8	NS	type of signals in the set. Value NS maps to "General".	"General", "Frequency", "Time"
rawName	i	utf8	O	original signal set name if this had to be renamed to be used as valid hdf5 name.	
description	i	utf8	NS	additional description of the set	(tbd)
parent	s	obj_ref	O	Link to parent signal set.	(tbd)
parentName	s	utf8	O	For information only: the name of the parent signal set.	
dataScale	a	float 64	A	Scale of the signal values in the set with respect to full scale	1.0 (at full scale) 23.456 (model scale, smaller than full scale) 0.023456 (model scale, larger than full scale)
waterDensityFactor	a	float 64	NS	water density factor to be used for scaling data values to full scale. Preferred 'not specified' over a default value.	1.432
stepSize	i	float 64	A	if one and only one master signal for all signals and it is equidistant a value; otherwise, NaN In case of a time master this is the inverse of the sample rate.	
dateTimeRecordingStart	a	iso_fmt	A	date and time of first sample of the time signal of measurement or simulation. Required for time series only..	(tbd)
dateTimeRecordingEnd	a	iso_fmt	O	date and time of last sample of the time signal of the measurement or simulation. (May disappear in future versions.)	(tbd)
projectNo	a	int32	A	projectnumber	"80220"
projectSubNo	i	int32	NS	subnumber	"368"
programNo	a	int32	A	test programme number	1
source	a	utf8	A	name of the source application or facility Within one project each 'programmeNo' is a fixed combination with the 'source'.	"SMB", "ReFresco"

categoryNo	a	int32	A	Number of the test category used.	2
testNo	a	int32	A	Number of the test setup used.	3
experimentNo	a	int32	A	Number of the experiment settings used.	4
measurementNo	a	int32	A	Number of the actual measurement c.q. experiment execution	2
branchNo	a	int32	O	Number of the acquisition branch. E.g. time traces from simulations where jumps to earlier points in time are possible.	1
sequenceNo	a	int32	O	Consecutive number of the signal set wrt to other signal sets. E.g. time traces from simulations in a given branch result in consecutive sets each time the set of logged properties changes.	13
modelScale	a	float64	A	Scale of the model in this project. (Unrelated to the signal values in the set.)	23.456
notes	i	utf8	NS	Any additional information about this signal set.	(free text)
writeErrors	i	utf8 []	O	Specifies any errors that occurred while writing the signal set (group).	

3.3 Signal : Dataset

The signal is a dataset containing the samples of the measurement or simulation or calculation or postprocessing step. A signal maps to an experiment variable.

Name: HDF Dataset name; must be unique in the set. It is the HDF-safe name of the signal. For reference the original potentially HDF-unsafe name is provided with the data in the 'signalSource' attribute.

Path: "/<signal set name>/<signal name>"

Datatype: float32, float64

Dimensions: 1 (time domain) to 7 (frequency domain)

Additional properties of the signal are added as attributes. Below are the common attributes. In sub sections domain or source specific attributes can be found.

Attributes:

Name	Data type	Exists? A / O / NS	Description	Convention / example	Schema
rawName	utf8	O	original signal name if this had to be reformatted to be used as hdf5 name.		
unit	unit_str	A	name of the unit of the signal. This defines the quantity of the experiment variable.	"m/s", "-", "rad"	sp1, da1
signalType	prop_str	NS	defines the experiment variable or simulation property or the kind of quantity, see Signal Type .	"velocity", etc "angular velocity"	sp1, da1
description	utf8	A	description of the signal. More detail than name. E.g. the kind of quantity or experiment variable; whether it is an absolute value or a delta.	(tbd)	
timeOffset	float64	NS	the time offset in seconds between real start time and the moment the timestamp is created.		da1
order	int32	O	In case of frequency data specifies whether it is a first or second order effect or otherwise.	1: first order. 2: second order. -1: Not applicable.	da1
minimum	ds_type	O	lowest value in the dataset		sp1
maximum	ds_type	O	highest value in the dataset		sp1
mean	ds_type	O	mean or average value of the dataset (tbd : weighted arithmetic mean?)		sp1
standardDeviation	ds_type	O	standard deviation of the values in the dataset		sp1
position	float64 [3]	NS	location of the variable in the specified Coordinate System.	[0.0, 0.0, 0.0]	da1

direction	float64 [3]	NS	direction of the signal c.q. experiment variable in the specified CS. In case of a rotation it is the direction of the axis of rotation.	Roll in ACK: [1.0,0.0,0.0] Sway in ACK: [0.0, 1.0, 0.0]	da1
referenceSystem	utf8	NS	specifies in which Coordinate System position and direction are given.	"ACK", (todo: complete list)	da1
signalSource	utf8	O	holds information about the sensor. Depends on system used. Value might be interpretable by a tool corresponding to the data source. Also contains the original signal name from the measurement system.	(tbd)	
channelNo	int32	O	if no sensor information is provided holds the channel number in case of measured data.	123	
bases	obj_ref []	O	List of object references to all datasets that are a master signal to this signal.	[objRef("time")]	h5m
baseNames	utf8[]	O	Names of the base signals (informational only. Not intended for rebuilding the datamodel)		h5m
notes	utf8	A	Any additional information about this signal.	(free text)	
writeErrors	utf8[]	O	Specifies any errors that occurred while writing the signal (dataset). Mostly value errors.		

Note: There is no explicit sample rate property. Sample rate is a specific attribute of equidistant time based signals. If sample rate needs to be visible it can be specified in the attribute 'description' of the signal set group.

4 H5M Data Descriptions

This chapter specifies some conventions with respect to the data and HDF attributes. Client code – a.k.a. applications – and h5m read/write implementations are allowed to fail if these conventions are not followed.

4.1 Naming conventions

Names of HDF5 attributes follow the camelCase naming convention, the first character of every word is capitalised except for the first word. This applies only to the attribute names and not the actual values. HDF5 Datasets and Groups get their name from the content (e.g. signal name, signal set name) and thus follow the convention of their own context. Items in the hdf5 file are accessible via their path(s) through the tree and the path consists of the names of the tree items. (This implies that names must be unique at each level.)

4.2 Attributes

Any HDF tree node item can have attributes. Attributes can have optional values or be optional themselves. The list of attributes per HDF tree node item specifies these cases as described in the sections below.

Attributes can be mandatory or optional. This is specified in the column 'Exists?' as follows:

A	Always	Attribute is mandatory and must have a valid value.
O	Optional	Attribute is only present if a valid value is available.
NS	Not Specified	Attribute is mandatory but may contain the string value 'not specified' if no valid value is available. Note: Although HDF supports a <i>null</i> dataspace for scalars and vectors it is not used because it lacks the ability of specifying <i>why</i> it has no value.

Attributes play a certain role. Currently the following roles are recognized in column 'Role':

i	Informational	Attribute is intended to be human interpretable.
a	Automation	One or more applications depend on the value of this attribute.
s	Structural	The value defines the relations in the data-model. Both human and computer interpretable.

4.3 Data types

Each entry in this section defines a specific type notation that is used in the chapter 'H5M Structure'.

4.3.1 Default value

The default value for attributes is the UTF-8 string value 'not specified'. It is important to notice that the data type of any attribute changes to UTF-8 string if there is no value specified and this default is set.

4.3.2 Strings: favour unicode (UTF-8) over ascii

All strings (names, datasets and attributes) in this convention are UTF-8 unless specified otherwise. The use of UTF-8 instead of ascii is encouraged by the HDF Group.

4.3.3 iso_fmt

Values of datetime attributes like 'dateTimeOfCreation' are added to the hdf item as a **string** attribute with formatting in accordance with ISO 8601. These values can optionally include timezone information. The HDF string is fixed length utf-8.

Examples:

- "2017-09-19T08:26:30.5606524" – no offset specified, thus local and could be anywhere.
- "2017-09-19T06:26:30.5606524Z" – Zulu time, UTC offset 0: at Greenwich UK without daylight saving time correction.
- "2017-09-19T08:26:30.5606524+02:00" – including offset with respect to UTC.

Including thus the local date and time of creation and optionally including time zone offset with respect to UTC (as opposed to GMT which includes daylight saving time correction). This offers readability for the users and can be easily converted to a datetime in loader libraries in python or dotNet or otherwise.

To get the UTC time (i.e. UTC Zulu time) one has to subtract the offset from the specified date-and-time. For example:

- "2017-09-19T08:26:30+02:00" is equal to "2017-09-19T06:26:30Z".
- "2017-09-19T08:26:30-04:00" is equal to "2017-09-19T12:26:30Z".

UTC differs from GMT in that it has no daylight saving time correction. To get the GMT time one has to add the daylight saving time correction. Using an API that supports iso 8601 format can do this automatically.

4.3.4 unit_str

The unit_str is a variable length HDF string. The content is human and computer readable. Therefore any unit is limited to a predefined list of acceptable units and formatted using a specific unit definition notation.

Examples of unit definition strings: "N.m", "kg.m²/s³", "kg.m.s⁻²"

A list of available or possible values is available at MARIN [here](#), and is not part of this convention (and probably never will be).

4.3.5 prop_str

The prop_str is a HDF string value which content is limited to a list of acceptable names for quantities, properties or experiment variables.

A list of predefined property / measurand names is not available yet.

4.3.6 obj_ref

The obj_ref is an HDF Object Reference value. It is used to define relations between individual entries in the file.

Note: Most common HDF viewers will only render this type as "object reference" and not specify to which item it refers. If applicable such an attribute is accompanied by an attribute that specifies the name(s) of the referred object(s).

4.3.7 ds_type

Used for the datatype of attributes that are the same as the *element type* of the dataset. Denoted "ds_type" for scalar, "ds_type[1,2]" for a 1 x 2 array, "ds_type[,,]" for a 3D array of any size and so on.