CEDAR Madrigal Hdf5 Format Specification

Oct 20, 2015

Bill Rideout

MIT Haystack Observatory

Introduction

This document is intended to describe the format of the new CEDAR Madrigal Hdf5 Format, which is a replacement for the old CEDAR database format. This new format is Hdf5, but beyond that must follow the layout rules described in this document. This Hdf5 format is designed to be entirely self-describing, so that any user who understands the Hdf5 format can fully understand the data in a CEDAR Madrigal Hdf5 format file without any reference to this document or any other. However, this document is designed for those who want to **create** CEDAR Madrigal Hdf5 format files.

The rules for creating CEDAR Madrigal Hdf5 format files are rigorously defined for two reasons:

1. To make sure all CEDAR Madrigal Hdf5 format files are fully self-describing, and
2. To make sure Madrigal can parse the files, allowing derived parameters, filtering, and searching to be built into Madrigal's capabilities.

CEDAR data model

The new CEDAR Madrigal Hdf5 Format retains the original CEDAR database data model and adds to it, and the metadata behind that format is unchanged. That metadata is the instrument metadata, the parameter metadata, and the kind of data metadata. However, all aspects of the model tied to 16 bit integers have changed. This section gives a complete description of that data model, with special emphasis on things that have changed with the elimination of 16 bit integers.

The CEDAR data model defines a CEDAR data file as a collection of CEDAR records. Each record has associated with it an instrument, a kind of data, a start time in UT, and end time in UT, and scalar and/or vector parameters with associated data.

Scalar parameters have a single value per record, whereas vector parameters have multiple values per record. All vector parameters have the same length in a given record. Note that the vector parameters often represent spatial dimensions, but can also represent other physical parameters such as energy. Time, however, cannot be represented in a vector parameter, because time always varies record to record. Having time vary as a vector parameter would make the time of the record ambiguous, and as such, is not allowed.

New with Madrigal 3.0 is the requirement that the independent parameters for vector information must be specified. Any set of vector data must have at least one parameter specified as an independent parameter. This information was not included in the CEDAR data format, so it must come from elsewhere when importing data from Madrigal 2 to Madrigal 3. This will be discussed in detail in the specification of the Hdf5 CEDAR format below.

Instruments are defined in the file instTab.txt. See the standard Madrigal administrators document for a full description. Each instrument must have a unique code. With Madrigal 3.0, instrument codes can be any integer, and are not limited to 16 bit integers. Madrigal administrators can add their own instrument, but should send any modifications to the OpenMadrigal administrator.

Kinds of data are defined in the file typeTab.txt. See the standard Madrigal administrators document for a full description. Each kindat of data must have a unique code. With Madrigal 3.0, kindat codes can be any integer, and are not limited to 16 bit integers. Madrigal administrators can add their own instrument, and do not need to send any modifications to the OpenMadrigal administrator. This is because kinds of data with Madrigal 3.0 will be set by both the instrument code and the kindat code, so kindat codes do not need to be unique across all Madrigal sites.

All parameters used in a Hdf5 CEDAR should be defined in the file $MADROOT/metadata/parmCodes.txt. This file takes the place of the old 16-bit oriented file parcods.tab, which limited the range of the possible parameter codes. Madrigal administrators may add parameters to this file, but they should notify the OpenMadrigal administrator if they do.

Every parameter in parmCodes.txt must have a unique positive integer and a unique mnemonic. Also, the string formed by placing a “D” in front of the mnemonic must not exist in parmCodes.txt. This is because the mnemonic for the error parameter associated with any standard parameter is “D” plus the standard mnemonic. A Madrigal administrator will be able to verify any additions to parmCodes.txt are valid by running the check script: *$MADROOT/source/madpy/scripts/bin/checkParmsCodes.py <new_parmCodes.txt_version>*.

Values of standard parameters must all be floats. If no data is available, NaN should be used to indicate missing values. This takes the place of the old CEDAR data format 16 bit integer value -32767. Values of error may be positive numbers, NaN if no error value is available, or two special negative values: -1.0 and -2.0. The value -1.0 indicates that the associated standard parameter value is assumed, not measured. This replaces the old CEDAR data format 16 bit integer value -32766. The value -2.0 indicates that the associated standard parameter value is known to be incorrect. This replaces the old CEDAR data format 16 bit integer value +32767.

The latest version of all metadata can now also be easily accessed from any Madrigal 3 site by choosing the top level *Access Metadata* menu.

Specification of Hdf5 CEDAR format

The Hdf5 CEDAR format will be based on the export Hdf5 format included in Madrigal 2.6, with a few additions, as noted below. This section includes a full description of the requirements for any Hdf5 file to be considered an Hdf5 CEDAR file.

At the top level, an Hdf5 CEDAR file must have two groups, Data and Metadata. Each group is described separately below.

Data group

The Data group has one required dataset, called “Table Layout”. It also has an optional group called “Array Layout” which contain data arranged in multiple dimensions – time, measured parameter, and one of more independent spatial parameters. It is used for data that has consistent independent spatial parameter values for all records.

Table Layout dataset

The data group has one required dataset, called “Table Layout”. This layout is a single table that contains all the data in the file, and is meant to be easy for the user to understand at a glance. Each parameter in the file has a column. The number of rows is the sum of the lengths of the vector parameters in each record, where records without vector parameters add only one row. This dataset is is of data type compound, where the column names are the lower case parameter mnemonics, and the values are stored as 64 bit floats.

In order to make sure the required information in each record is also included in “Table Layout”, a number of parameters are required, and they must be listed first in the record and in this order:

1. year – the UT year at the average time in the record
2. month - the UT month at the average time in the record
3. day - the UT day at the average time in the record
4. hour - the UT hour at the average time in the record
5. min - the UT minute at the average time in the record
6. sec - the UT second at the average time in the record
7. recno – the record number of this record (starts at 0)
8. kindat – the kind of data code of this record.
9. kinst – the instrument code of this record.
10. ut1_unix – the unix time at the start of the record (seconds as float since Jan. 1, 1970 UT). Times before 1970 are negative.
11. ut2_unix – the unix time at the end of the record (seconds as float since Jan. 1, 1970 UT). Times before 1970 are negative.

The times given as year, month, day, hour, min, and sec are meant for ease of use, wheres ut1_unix and ut2_unix can be as precise as desired and determine the exact time the record began and ended. These required parameters are the equivalent of the old CEDAR data format prolog information.

All other parameters in this table follow the first required eleven. Note that a value may appear as Nan in this table representing the old CEDAR missing value. The CEDAR special values for error parameters, assumed and known bad, are described below.

From "Table Layout" dataset alone, it is not always obvious which parameters are 1D, which are independent 2D spatial parameters, and which are dependent 2D parameters. However, the Metadata._record_layout dataset described below gives that information.

Array Layout group (optional)

The data group has one optional group, called "Array Layout". This layout exists if and only if the metadata dataset "Independent spatial parameters" exists. The idea is that users would like to access data with the number of dimensions equal to the number of independent parameters, time and the independent spatial parameters listed in "Independent spatial parameters", as opposed to the flattened data in "Table Layout". The Cedar data model does not required records to have 2D data (that is, measurements at different points in space), and so this group is not required.

There are also cases where the data in the file would be more usable if separated into more than one array layout. For example, a phased array incoherent scatter radar may make measurements on several beams at once. For the user, the array data would be more helpful if it were separated by individual beams. It is possible that more than one parameter may be used to split data into separate array layouts. For example, at Millstone Hill, data is split by both pulse length and mode type (single-pulse or alternating code). The number of different array layouts in the Hdf5 file will be the number of unique values of the combination of all parameters used to split data. These parameters used to split array data are listed in the optional metadata dataset called "Parameters Used to Split Array Data".

If there is no "Parameters Used to Split Array Data" metadata dataset, then the array layout is directly beneath the "Array Layout" goup. If there is a "Parameters Used to Split Array Data" metadata dataset, then beneath the "Array Layout" group, there will be subgroups with the names "Array with <parm>=value [and <parm>=value …] where the number of <parm>=value sections is equal to the length of the "Parameters Used to Split Array Data" metadata dataset.

The array layout itself, whether it is directly under the "Array Layout" group or a subgroup described above, consists of:

1. "1D Parameters" group
2. "2D Parameters" group
3. Layout Description – an array of strings describing this layout
4. timestamps – average time for record in seconds since 1970/01/01 UT midnight (int)
5. One of more arrays of the values of independent spatial parameters. Name is mnemonic of parameter.

The "1D Parameters" group consists of "Data Parameters" dataset exactly like the "Data Parameters" dataset under / Metadata, except that only 1D parameters (scalar values) are listed. In addition, there is a dataset for each 1D parameter with the mnemonic as a name, and a length equal to the number of records (or timestamps).

The "2D Parameters" group consists of "Data Parameters" dataset exactly like the "Data Parameters" dataset under / Metadata, except that only 2D parameters (vector values) are listed. In addition, there is a dataset for each 2D parameter with the mnemonic as a name, and with number of dimensions = 1 + the number of independent spatial parameters. The length of the dimensions are (number of unique values of independent spatial parm 1, [number of unique values of independent spatial parm 2, …] number of records or timestamps.

Metadata group

The Metadata group has four required datasets: "Data Parameters", "Experiment Notes", "Experiment Parameters" "and "_record_layout". There are also two optional datasets, "Independent spatial parameters" and "Parameters Used to Split Array Data". There two optional datasets will only be present when there are parameters to list in each list. Each dataset is described below.

Data Parameters dataset

The Data Parameters dataset is a table-formatted dataset with the following columns:

Name	***Description***	***Type***
mnemonic	CEDAR mnemonic (no spaces)	String
description	Short description of parameter	String
isError	1 if describing error, 0 otherwise	Integer
units	Parameter units (N/A in no units)	String
category	CEDAR parameter category	String

Experiment Notes dataset

The Experiment Notes dataset stores all the text formally stored in the old CEDAR catalog and header records. It is the section used for any text notes on the data. This dataset is a table-formatted dataset with the following columns:

Name	***Description***	***Type***
File Notes	Text description of file	String – maximum of 80 characters per entry

Experiment Parameters dataset

The Experiment Parameters dataset describes overall parameters associated with this experiment. This dataset is a table-formatted dataset with the following columns:

Name	***Description***	***Type***
name	Experiment parameter being given	String
value	Value of experiment parameter	String

Recommended name/values to be included are:

name	***value***
instrument	Instrument name

Instrument codes(s)	Instrument code (unique CEDAR integer). Comma-separated list if more than one.
kind of data file	Description of kind of data. More than one description if more than one kind of data.
kindat codes(s)	Kind of data code (identifier unique to a given Madrigal site). Comma-separated list if more than one.
start time	Start time of experiment in format YYYY-MM-DD HH:MM:SS UT
end time	End time of experiment in format YYYY-MM-DD HH:MM:SS UT
Cedar file name	Full path to Cedar file when loaded on Madrigal site
status description	Status description of file when this file created
instrument latitude	Instrument latitude (-90 to 90)
instrument longitude	Instrument longitude (-180 to 180)
instrument altitude	Instrument altitude in km

Independent Spatial Parameters dataset

The Independent Spatial Parameters dataset is an optional table-formatted dataset. It must exist if any parameter in the file is 2D. It has the following columns:

Name	***Description***	***Type***
mnemonic	CEDAR mnemonic (no spaces)	String
description	Short description of parameter	String

Parameters Used to Split Array Data dataset

The Parameters Used to Split Array Data is an optional table-formatted dataset. It must exist if Array Layout exists and has multiple datasets below it. It has the following columns:

Name	***Description***	***Type***
mnemonic	CEDAR mnemonic (no spaces)	String
description	Short description of parameter	String

_record_layout dataset

The _record_layout dataset is meant to be an internal dataset that contains metadata to speed up programs that interact with this file. This dataset is a table-formatted dataset with a column for ever parameter in the file (in the same order as listed in

Data/"Table Layout"). It has just one row, since all records in that file must now be consistent. The value for each parameter/row is 3 if this is an independent vector parameter, 2 for if this is a non-independent vector parameter, and 1 for a scalar parameter.

Whether a vector parameter is independent or not is not defined by the old CEDAR database format. When a old CEDAR database format file is upconverted to HDF5 CEDAR, an ini file in the form of cachedFile.ini can be used to determine which vector parameters are independent. If no such file is appropriate, then the following parameters list will be searched, and the first one found will be used:

6. range
7. gdalt
8. altv (virtual height)
9. paclat
10. cgm_lat

If none of these parameters are found, an exception is thrown.

Independent Spatial Parameters dataset (optional)

The Independent Spatial Parameters dataset is metadata that lists all the independent spatial parameters used in this file. Independent spatial parameters are those used to specify the location of multiple spatial measurements that occur in the same record. The CEDAR database model does not require that there be multiple spatial measurments in each record, so it is possible that there are no independent spatial parameters in a file. In this case this entire dataset will be skipped. This metadata is used to create the Array Layout described above. The number of dimesions in the array layout 2D datasets will be one greater than the number of independent spatial parameters (the one extra dimension always represents time). The Independent Spatial Parameters dataset is a table-formatted dataset with the following columns:

Name	***Description***	***Type***
mnemonic	CEDAR mnemonic (no spaces)	String
description	Short description of parameter	String

Parameters Used to Split Array Data dataset (optional)

The Parameters Used to Split Array Data dataset is metadata that lists all the parameters used to divide array data into different sets. For example, a phased array incoherent scatter radar may make measurements on several beams at once. For the user, the array data would be more helpful if it were separated by individual beams. It is possible that more than one parameter may be used to split data into separate array layouts. For example, at Millstone Hill, data is split by both pulse length and mode type (single-pulse or alternating code). The number of different array layout in the Hdf5 file will be the number of unique values of the combination of all parameters in this dataset. The Parameters Used to Split Array Data dataset is a table-formatted dataset with the following columns:

Name	***Description***	***Type***
mnemonic	CEDAR mnemonic (no spaces)	String
description	Short description of parameter	String