National Critical Zone Observatory Program

CZO Display File Specification

…

Developed by: Ilya Zaslavski, David Tarboton and Mark Williams

Maintained by: Tom Whitenack

1/31/2011

Table of Contents

Introduction

The file format for delivering CZO data in machine/human readable form is referred as a "Display file format" – This format consist of a main header file, with ancillary information such as site and method metadata, is stored in separate files.

Header File Format Specification

Overall structure

- - A header block
- - A data block (optionally a separate file).

The data block contains columns that specify individual data values. There is no limit on number of columns or rows. Data values may be text or numeric (depending upon the data type dictated by the column information in the header block). Columns are comma separated (text that contains commas should be included in quotes).

The header block specifies first information that pertains to all columns (as default parameters) and then column specific information.

Header block syntax

Prefix

\header

Three groups

- Doc group
- Default parameter group
- Column header group

Doc group.

- - Title
- - Abstract
- - Investigator (Multiple investigators may be listed where the investigator has changed during the course of time)
- - Variable names
- - Keywords
- - Citation
- - Publications
- - Comments

The title is used as a unique identifier for data series, so changes to the title will result in breaks in series that need to be managed and communicated to users and tools that work with the data to consolidate time series where the title may have changed.

Doc Attributes	**Description**
Title	A title for the set of data series in the file
Abstract	Description of the data
Investigator contact Information	Name and contact information for investigator responsible for the data
Keywords	Keywords useful for discovery of the data series
Variable names	Names for variables for the data series
Citation	Text string that give the citation to be used when the data are referenced.
Publications	Publications related to this data
Comments	Additional comments related to interpretation and use of this data

Default parameter group

The specification of a default parameter means that it pertains to all data in the file except when overridden by a specific column specifier. This capability is provided to encourage the strategy to put as much content in default parameters as possible and specify them only once, rather than repeating information with column headers.

Syntax

DEFAULT_PARAMETER. <SeriesAttribute>=<SeriesAttributeValue>,
DEFAULT_PARAMETER. <SeriesAttribute>=<SeriesAttributeValue>, ...

When part of the default parameter group series attribute specifications set default series attributes that pertain to the entire file

<SeriesAttribute> specifies the series level attributes (controlled properties) that pertain as defaults to all the series in the file. This should be from the seriesAttributes table.

<SeriesAttributeValue> the specific value of the series level attribute.

There may be multiple <SeriesAttribute>=<SeriesAttributeValue> specifications for each default parameter record. There may also be multiple default parameter records. Specifications that are part of the same default parameter specification are linked and may be used to specify linked series attribute defaults, such as the units associated with a default offset or time support value.

Examples

DEFAULT_PARAMETER. site ="GREEN LAKE 4"

DEFAULT_PARAMETER. offset_value ="2"

DEFAULT_PARAMETER. quality_control_level ="0"

DEFAULT_PARAMETER. missing_value_indicator value ="-9999"

DEFAULT_PARAMETER.

Column header group

COL<n>. label=<column type>, value=<Value>,
<SeriesAttribute>=<SeriesAttributeValue>

<n> indicates column number.

<column type> may be either ValueAttribute or VariableName depending on whether the column contains a value level attribute (such as DateTime, Offset) or data values. Value attributes are measured properties associated with the data value while VariableName refers to a controlled property, i.e. the quantity selected to measure.

<Value> specifies the content of the column and should be the attribute (from valueAttributes table) in the case where label is ValueAttribute, or a VariableName value in the case where label is VariableName.

<SeriesAttribute> specifies the series level attributes (controlled properties) that pertain to the series whose data values are in the column. This should be from the seriesAttributes table.

<SeriesAttributeValue> the specific value of the series level attribute for the series held by the column.

There may be multiple <SeriesAttribute>=>=<SeriesAttributeValue> specifications for each column to specify required series parameters.

SiteCode

Some CZO's are collecting data in a way that multiple values are being collected by the same logger device at the same time, and represent slightly different sampling locations clustered around the logger. As a result, a data logger will have columns of data representing measurements from different sampling locations. To accomodate this senario, each sampling location will be added to the sites.csv configurtation as a new site and will then be referenced by siteCode for each column.

For example, "UpMetN" is the "parent" site which represents the logging device. The sample locations are cdde and cduc would be represented as follows in the column specification:

COL2. label=VariableName, value=StreamFlow, units=m3/s, TimeSupport= 3, TimeSupportUnits=hr,**siteCode=UpMetN_cdde**

COL3 label=VariableName, value=StreamFlow, units=m3/s, TimeSupport= 3, TimeSupportUnits=hr,**siteCode=UpMetN_cduc**

In other words, if all measurements in a data file are made at a single sampling location, this location (site) can be specified in DEFAULT_PARAMETER. If the data file contains data for multiple sampling locations (sites) and each row contains measurements taken at a specific individual site, then we can use a column defined as the siteCode. If each row contains measurements from several sampling locations then we add siteCode designation to column description.

Therefore, siteCode in the column specification should not be used if a DEFAULT_PARAMETER is used to set the siteCode or if a column is defined as the siteCode. In case there are, the rule will be that the siteCode included in the column specification will override the other two.

The second information model issue we ran into was how to deal with multiple offsets (e.g. top of the canopy, bottom of the canopy). Trees grow, so we shall try to use relative vertical offsets. One way to do it within the current schema is to create a numeric coding scheme for different types of offsets (e.g. 1=upper crown, 2=lower crown, etc.). Another way is to add new sample medium entries (as in http://his.cuahsi.org/mastercvreg/edit_cv11.aspx?tbl=SampleMediumCV&id=533576939) and add the medium designation to column specification in the header file. Any ideas, examples, clarifications are welcome.

Examples

COL1. label=ValueAttribute, value=DateTime, UTCOffset=-7, Timezone=MST, format="YYYYMMDD hh:mm"

COL2. label=VariableName, value=StreamFlow, units=m3/s, TimeSupport= 3, TimeSupportUnits=hr,

NoDataValue=-9999, SampleMedium=water, method=method1, Offsetvalue = 3, OffsetValueUnits=m , offsetDescription = "Depth below surface",

COL3. label=VariableName, value=pH, units=pH units, missing value indicator=-9999

COL4. label=VariableName, value=conductance, units=uS/cm @ 25 degrees C,

Splitting the CZO Data Display File into .hdr and .dat files

Splitting the data display files into a header (.hdr) and a data file (.dat). This change is requested by several CZO sites. The reason is that frequently updated loggernet files don't contain metadata. It would be more convenient to keep metadata in a separate file (which is created once, and usually not updated as the data file changes).

An example is provided by Brian Bills, Shale Hills CZO.

The .hdr files include the \doc, \header (with variable metadata) and possibly other sections such as \log.

The .dat file contains the measured values. In the sample provided by Brian (which appears typical for Loggernet files), it also contains variable and unit definitions that occupy the first 4 rows. For generality, we suggest that the header file contains a directive specifying how many rows in the data file need to be skipped, in the following format:

```
DEFAULT_PARAMETER.DataBeginsOnRow = "5"

DEFAULT_PARAMETER.DataFile = "mydata.dat"
```

Data block syntax Example

\data

GREEN LAKE
4,820311,,6.4,18,88.51,0.40,,114.77,24.68,21.75,10.23,25.389,,58.296,83.200,,,,,,,,,,,,,,,,,,

GREEN LAKE
4,820422,,5.7,18,90.15,2.00,,99.80,24.68,17.40,12.79,9.591,,72.870,44.928,,,,,,,,,,,,,,,,,,

GREEN LAKE
4,820506,,6,17,81.95,1.00,,99.80,16.45,26.10,12.79,16.926,,85.362,39.936,,,,,,,,,,,,,,,,,,

GREEN LAKE
4,820527,,6.3,17,99.98,0.50,,139.72,25.50,23.93,8.95,24.543,,108.264,34.944,,,,,,,,,,,,,,,,,,

Series level attributes

The header and data blocks need to be constructed so that the following attributes are specified for each data value in a **CZO time series display file.** Attributes are to be interpreted as required, meaning that they must be specified in either a default parameter or column header, unless designated as optional below.

Attributes	**Description**
SiteCode	Code used to identify the site (refers to sites file)
Units	The units associated with a data value
Method	Identifier to point to a record in the methods file
OffsetValue	The value of a measurement offset. (Optional)
OffsetDescription	Full text description of the offset value. (Optional, but required if OffsetValue is given)
SampleType	Type of sample, e.g. grab, from groundwater, from leaf. From sample type preferred value table

Attributes	Description
VariableName	Name of the variable from the variables preferred value table.
SampleMedium	The medium of the sample. This should be from the SampleMediumPV preferred vocabulary table.
ValueType	Text value indicating what type of data value is being recorded. This should be from the ValueTypeCV controlled vocabulary table. (e.g. Field measurement, modeled, derived)
TimeSupport	Numerical value that indicates the temporal footprint of the data values. 0 is used to indicate data values that are instantaneous. Other values indicate the time over which the data values are implicitly or explicitly averaged or aggregated.
TimeSupportUnits	Units of time support value from Units PV table.
DataType	Text value that identifies the data as one of several types (e.g. min, max, average). PV
DataLevel	Level used to identify the level of quality control to which data values have been subjected.
NoDataValue	The value used to encode no data
UTCOffset	Offset in hours from UTC time of the corresponding LocalDateTime value.
TimeZone	Time zone where observation site is located (e.g. Mountain time)
OffsetValue	Distance from a datum or control point to the point at which a data value was observed. If not given the OffsetValue is inferred to be 0, or not relevant/necessary.
OffsetDescription	Description of how offet values are defined to include point of reference and direction
OffSetUnits	Units with which the offset value is measured (Units PV)
CensorCode	Text indication of whether the data value is censored from the CensorCodeCV controlled vocabulary.

Value level attributes

The following value level attributes may be specified as columns in a **CZO time series display file.** Only one DateTime column is permitted in any single file. OffsetValue columns are interpreted as pertaining to the column immediately to the left (and as such they cannot be the leftmost, nor follow a DateTime column).

Attributes	Description
DateTime	The date and time at which the value was observed
OffsetValue	The value of a measurement offset. (Optional). [Note that OffsetValue may be either a series level, or value level attribute for any data series, depending upon whether it is a controlled or measured property.]

Censor indicators and data qualifiers

There is a need to designate specific data values as either censored or qualified (or both). Logically these are value level attributes however for display and readability purposes the CZO requirement was set that these not be stand alone columns in the time series display file. Rather individual data values in any column may be indicated as censored or qualified using the following syntax:

<datavalue>:c=<censorCode>,q=<qualifierCode>

censorCode and qualifierCode should be from the associated preferred value tables. Multiple qualifiers are permitted.

Examples

45.3:c=LT,q=E,q=A

Sites File Format Specification

Sites are locations at which measurements are recorded. The time series display file should identify sites using a site code unique to each CZO (When this identifier is presented by CZO central it will be preceded by a specific CZO identifier to avoid site code conflicts across CZO's). The CZO site file shall contain the following attributes for each site and be presented as an ASCII comma separated value file, with a one row header of attribute labels.

Site File Attribute labels	**Description**
SiteCode	Code used by organization that collects the data to identify the site
SiteName	Full name of the sampling site.
Latitude	Latitude in decimal degrees.
Longitude	Longitude in decimal degrees. East positive, West negative.
LatLongDatum	The Spatial Reference System of the latitude and longitude coordinates in the SpatialReferences table.
Elevation	Elevation of site.
VerticalDatum	Vertical datum of the elevation. Controlled Vocabulary from VerticalDatumCV.
LocalX	Local Projection X coordinate. (Optional)
LocalY	Local Projection Y Coordinate. (Optional)
LocalProjection	Identifier that references the Spatial Reference System of the local coordinates. (Optional)
PosAccuracy	Value giving the accuracy with which the positional information is specified. (Optional)
Comments	Comments related to the site. (Optional)

Methods File Format Specification

Methods used to make measurements are series level attributes. The time series display file should identify methods using a method code unique to each CZO (When this identifier is presented by CZO central it will be preceded by a specific CZO identifier to avoid method code conflicts across CZO's). The CZO method file shall contain the following attributes for each method and be presented as an ASCII comma separated value file, with a one row header of attribute labels.

Attributes	Description	Link
Method	Description of each method.	Hyperlink to external reference on the method (Optional)

Preferred Vocabularies

CZO central to host the following shared vocabularies. Moderators to be designated by CZO PI's based on expertise in each category.

- Variable names. The specific name of the quantity being measured (grouped into categories with a keyword list associated with each name. Capability to display and moderate within categories. Need a field for keywords and categories to be added to present CUAHSI HIS system) (e.g. Precipitation, Streamflow, Nitrogen, Soil moisture)

  - - Units (extended from CUAHSI HIS) (e.g. m, g/L)
  - - Value type (from CUAHSI HIS) (e.g. Field observation, derived value, model output)
  - - Sample type (from CUAHSI HIS) (e.g. stream water, ground water, rock, soil)
  - - Data type (from CUAHSI HIS) (e.g. average over interval, cumulative, continuous, sporadic)
  - - Data level (based on Ameriflux list) (e.g. level 0=raw data, level 4 = fully infilled and quality controlled)
  - - Spatial references (extensible based on EPSG) (e.g. NAD 1983, WGS84, UTM zone 11)
  - - Censor code (from CUAHSI HIS) (e.g. less than, not-censored, non detect)
  - - Qualifier code (in CUAHSI HIS qualifiers are not a PV. A CZO specific set of qualifiers will need to be developed)
  - - Vertical datum (from CUAHSI HIS) (e.g. Mean Sea Level, NGVD29)
  -

CZO Publication Configuration file

This file specifies the location of all header, method, and sites files being served by a CZO.

Two main "nodes" are the "root_url" and "category":

Root_URL: Url pointing to the root directory, against which all subdirectories are resolved

Category: thematic grouping of files. Below we provide examples for category "Hydrologic Time Series." Other categories might be named "Geophysical" , "Geochemical", etc., and will point to respective data files.

Within the Category section are three sub nodes:

Sites: contains a relative path to the sites file (sites.csv) (one file)

Methods: contains a relative path to the methods file (methods.csv)(one file)

Header: contains a relative path to header files (one or more header files. For example, with a standard Loggernet setup we would expect a single header file per site).

It is expected that .hdr and .dat files will have the same root name.

In this design, we assume a single sites file, a single methods file, and multiple header-data file pairs. The "single file" restriction may be relaxed in the future if needed.

Example Config File:

\Root_URL

http://spatial.sdsc.edu/lab/czo/

\Category

Hydrologic Time Series

\Sites

hydrodata/sites.csv

\Methods

hydrodata /methods.csv

\Headers

hydrodata /2009/site1.hdr

hydrodata /2009/site2.hdr

hydrodata /2009/site3.hdr

hydrodata /2009/site4.hdr