GCE Data Toolbox: Metadata-driven Software for Data Acquisition, Quality Control and Synthesis

Wade Sheldon, Georgia Coastal Ecosystems LTER, University of Georgia (*email:* sheldon@uga.edu)

Introduction

The effort required to process, document, and quality control raw data from sensors is often a limiting step in bringing environmental data online. Similarly, the effort required to find, download and refactor data collected by others can prove limiting in large-scale synthesis efforts.

However, the **GCE Data Toolbox** (MATLAB-based software developed at GCE-LTER) has proven effective in overcoming both of these barriers. This software can automate processing of data collected by a wide variety of data logger systems, from initial acquisition through quality control and distribution of documented data sets and plots. It is equally adept at harvesting and integrating existing data from national monitoring programs and environmental databases (e.g. LTER ClimDB/HydroDB, USGS NWIS, NOAA NCDC, NOAA NERR).

This poster provides a brief overview of the toolbox, which is freely available for use by other LTER sites.

Managing Data and Metadata

Underlying the software is a robust data model that combines fine-grained metadata, data columns, Q/C rules and Q/C qualifier flags into a well defined data structure. Data table size is only limited by computer memory, and any number of Q/C rules can be pre-defined for each column to assign flags automatically upon loading. Data values are intrinsically "shadowed" by metadata and Q/C flags throughout all operations, and flags are updated whenever data values change.

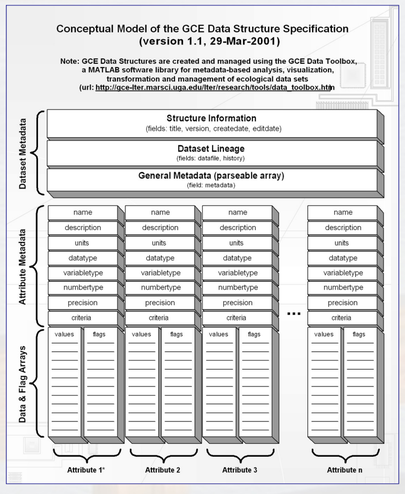

Figure 1. Conceptual model of the GCE Data Structure specification, the underlying data model used by the GCE Data Toolbox software.

Generating Metadata

Basic data column (attribute) metadata are generated automatically when data are loaded, but comprehensive metadata can also be pre-defined and saved as reusable metadata templates. Templates contain boilerplate documentation, detailed data column descriptors (e.g. name, units, description, data type, precision), and Q/C rules for each column. When templates are applied, data are validated and Q/C rules are applied automatically. Multiple templates can be defined for each data source to support different reporting standards or vocabularies.

Acquiring Sensor Data

Raw data can be imported from delimited text or MATLAB files, as well as proprietary export formats used by common environmental data loggers, e.g:

- Campbell Scientific CR10x and CR1000-3000
- SeaBird Electronics CTDs and MicroCATs
- In-Situ, Schlumberger, Hach-Hydrolabs well loggers
- Hobo TidbiT temperature loggers

Data files can be loaded from a file system or via network protocols (SMB/CIFS, HTTP, FTP, SOAP) for remote deployments. Batch processing all files in a directory and executing imports on a timed basis are both supported.

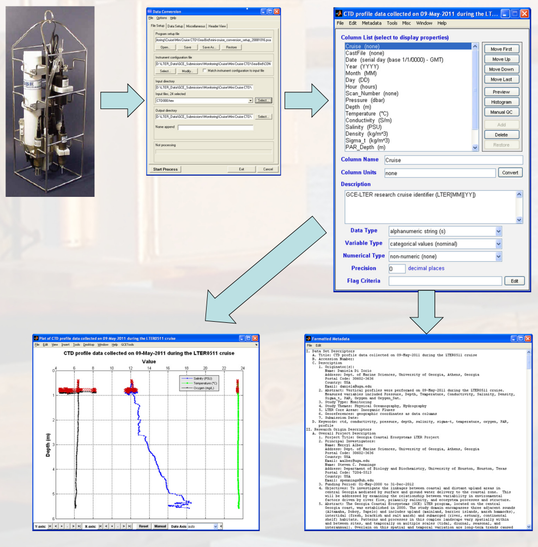

Figure 2. Illustration of the workflow for importing a SeaBird CTD data file. Raw data are batch-exported by SBE software, then imported into the GCE Data Toolbox to create documented, Q/C'd data in one step. Note the automatic Q/C flags (in red) assigned to values collected during the initial soaking period. Batch scripting is also supported.

Transforming and Analyzing Data

After data are imported a wide variety of tools are available for transforming and analyzing them. Metadata are used to configure dialogs automatically and verify suitability of data selections for each transformation. Transformation steps, calculations and data changes are logged to the metadata automatically to document the complete processing lineage of the data set.

Examples of transformations include:

- Unit inter-conversions (including English <-> Metric)
- Filtering records by value or mathematical expression
- Sub-setting data by column or row indices
- Statistical data reduction by aggregation, binning
- Temporal scaling/re-sampling
- Gap-filling via gated interpolation (various algorithms)
- Splitting compound data series into individual columns
- Integrating multiple data sets via relational joins or unions

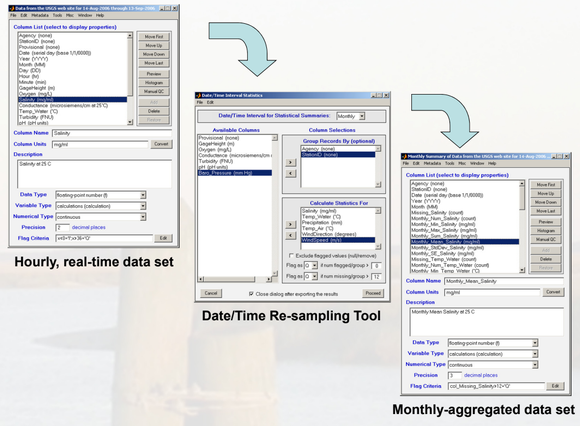

Figure 3. Data transformation using the GCE Data Toolbox date/time re-sampling tool. Numbers of flagged and missing values in the source data are automatically tallied in the derived data set, and these tallies can be used to flag statistical results automatically when user-set thresholds for flagged and/or missing values are exceeded.

Harvesting Data for Synthesis

A powerful feature of the GCE Data Toolbox is support for "harvesting" data from the LTER ClimDB/HydroDB, USGS and NCDC databases directly over the Internet, using either interactive graphical dialogs (fig.4) or scriptable command-line functions. Data can be retrieved for any ClimDB, USGS NWIS or NOAA NWS station across the country, and browsable lists of stations grouped by site (ClimDB) or state/territory (USGS, NCDC) are displayed in data request dialogs for selecting stations.

This capability, combined with the metadata templating and Q/C flagging features already described, allows users to simultaneously acquire and standardize large amounts of long-term climate and hydrologic data from across the US with just a few mouse clicks or commands.

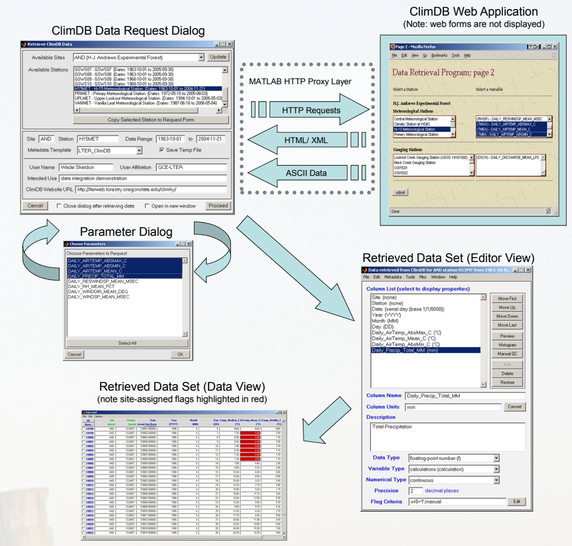

Figure 4. Retrieving data from ClimDB/HydroDB over the Internet and viewing the data using the GCE Data Toolbox. Note that ClimDB web pages are not displayed to the user (i.e. network operations are handled transparently). A similar dialog is available for retrieving data from the USGS NWIS and NCDC databases.

Conclusion

The GCE Data Toolbox was developed to meet the specific needs of LTER data management and analysis, and has been refined through constant use over 10 years. This software is a key component of the GCE Information System, and is also used at CWT, SEV and other sites.

This software currently runs the **USGS Harvesting Service** for HydroDB, automatically downloading and transforming long-term streamflow data for 81 USGS stations near 13 LTER sites on a weekly basis. Work is also underway to provide native EML support, so that the software can be used as a testbed for executing LTER NIS workflows. Once complete, these additions will make the software even more useful for LTER synthesis activities.

More Information

The GCE Data Toolbox is freely available as open source software under a GPLv3 license. Additional information, documentation and download links are available at:

https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox

GCE – Information Management