**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 open source software.

---

**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 predefined 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.

---

**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 (SMBCIFS, HTTP, FTP, SOAP) for remote deployments. Batch processing all files in a directory and executing imports on a timed basis are both 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/sampling
- Gap-filling via gated interpolation (various algorithms)
- Splitting compound data series into individual columns
- Integrating multiple data sets via relational joins or unions

---

**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 predefined 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.

---

**Concluding Remarks**

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 other LTER sites. This software also runs the USGS Harvesting Service for LTER HydroDB, automatically downloading and transforming long-term streamflow data for 85 USGS stations near 13 LTER sites on a weekly basis. Work is also underway to provide native EMQ 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 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