Georgia Coastal Ecosystems LTER
Information Management

Data Management

Adam Sapp - Information Manager
Wade Sheldon - Information Management System Administrator
University of Georgia
Athens, Georgia
GCE Data Toolbox

- © 2002-2021 Wade M. Sheldon and the GCE-LTER
- Generalized MATLAB package for metadata-based data management
- Any number of numeric and text variables
- Attribute metadata for each variable (name, units, description, data type, variable type, precision, ...)
- Structured documentation metadata for dynamic updating, formatting
- Versioning and processing history info (lineage)
- Quality control rules for every variable, qualifier flags for every value
Data Management Cycle

1. Data Source
2. Import Data
3. Add / Import Metadata
4. Q/C Analysis
5. Post-Process
6. Synthesis
7. Archive / Publish
Data Management Cycle

Data Source → Import Data → Add / Import Metadata → Q/C Analysis → Post-Process → Synthesis → Archive / Publish

Data Submission
Data Submission Pathways

- Automatically harvested data
  - These data are automatically harvested from data loggers or online systems via network telemetry
  - Data are directly imported into the GCE data toolbox for automatic processing
- Examples
  - Flux Tower
  - Weather Stations
  - Streamflow gauges
Data Submission Pathways

- **Sensor data**
  - These data require periodic manual downloading
  - Data are synchronized to the GCE servers for semi-automated processing
  - Examples
    - CTD cast
    - Hydrographic sondes
    - Well loggers
Data Submission Pathways

- Monitoring data
  - Infrequent collection or derived from laboratory analysis
  - Repeat submissions - same parameters are measured
  - Sample information is stored in a centralized database
- Examples
  - Fall vegetation monitoring
  - Nutrient concentrations from water samples
Monitoring Data

- Specialized web forms and spreadsheet templates developed to ease data entry and provide initial quality control.
Data Submission Pathways

- Investigator submitted data
  - Not amenable to automated processing
  - Submitted by web-based metadata forms and data submission spreadsheets
- Examples
  - Directed studies
  - Graduate student projects
  - Non-tabular data
Investigator Submitted Data

- Study metadata entered via web-based submission forms
- List boxes for selecting personnel, site locations, species info, instruments (details imported from metabase)
- Help text in pop-up text balloons for every field
- Information is saved in a centralized database then copied to metabase when approved by IM
Investigator Submitted Data

- Tabular data submitted in data submission template (MS Excel)
- Template guides investigator in describing each field in a data table
- Submitted files are saved on file server
- IM office is automatically notified when a data set is ready for review
Data Management Cycle

Data Source -> Import Data -> Add / Import Metadata -> Q/C Analysis -> Post-Process -> Synthesis -> Archive / Publish

Data Processing
Data Processing – Import Data

- **Generic Parsing**
  - Delimited text (CSV, space, tab)
  - MATLAB variables (arrays, matrices, structs)

- **Specialized Parsing**
  - Vendor-specific logger files
    - Campbell Scientific Instruments
    - Sea-Bird CTD, sondes
  - Network data sources
    - Federal databases (USGS, NWIS, NOAA,...)
    - EML repositories (EDI Data Portal, DataONE, KNB)
    - SQL database queries

- **Custom Parsing**
Data Processing – Add/Import Metadata

- Imported along with data
  - Logger file headers (Campbell, Sea-Bird)
  - Station, parameter information from USGS, NOAA
  - Tokenized headers from Data Submission Template
- Imported from other GCE data structures
- Imported from data repository (EML/XML)
- Added from user-defined ‘templates’
  - Column (attribute) metadata matched to ‘variables’
  - Boilerplate documentation
  - GUI tool for creating/managing templates
Data Processing – Add/Import Metadata

Metadata Templates

<table>
<thead>
<tr>
<th>Name</th>
<th>CD-GRS_Ford_Pulaski</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td></td>
</tr>
<tr>
<td>Field</td>
<td></td>
</tr>
<tr>
<td>CONTENTS</td>
<td></td>
</tr>
</tbody>
</table>

Metadata Contents (click on rows to view/edit contents)

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>FIELD</th>
<th>CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Note: use | characters to force line breaks in the formatted metadata)
Data Processing – Q/C Analysis

- Programmatic Q/C Analysis
  - “Rules” define conditions in which values should be flagged
  - Unlimited Q/C rules for each variable
  - Rules evaluated when data loaded and when data or rules change
  - Rules predefined in metadata templates to automate Q/C on import

- Interactive Q/C Analysis and Revision
  - Qualifiers can be assigned/cleared visually on data plots with the mouse
  - Qualifiers can be propagated to dependent columns
  - Qualifiers can be removed or edited (search/replace) if standards change
Data Processing – Q/C Analysis
Data Processing – Q/C Analysis

Data from Sea-Bird Electronics 37-SM MicroCAT sonde S/N 3746 deployed at the Altamaha River hydrographic datalogger deployment near Rockdedundy Island from 10-Jun-2011 to 08-Sep-2011

- Temperature (°C)
- Salinity (PSU)
- Depth (m)

Flag: no value change (col_Salinity,0.3,0.3,3) = 'F'

col_Depth < 0.2 = 'Q'

col_Depth < 0 = 'I'

Date (GMT)
Data Processing – Q/C Analysis

Data from Sea-Bird Electronics 37-SM MicroCAT sonde S/N 3746 deployed at the Altamaha River hydrographic datalogger deployment near Rockedundy Island from 10-Jun-2011 to 08-Sep-2011

- Temperature (°C)
- Salinity (PSU)
- Depth (m)

Flag no value: col_Depth<0.2 = 'Q'
Flag no value: col_Depth<0 = 'I'

Graph shows time series data from June 10, 2011 to September 8, 2011 with temperature, salinity, and depth measurements.
Data Processing – Post-Processing

- Calculated columns can be generated using mathematical formulas, functions
- Data can be gap-filled, drift-corrected
- Derived data sets can be created by filtering values or refactoring data table structure (e.g. combining or splitting columns)
- Data can be re-sampled or summarized by aggregation, binning and date/time scaling
Data Processing – Post Processing
Data Processing – Synthesis

- Multiple data sets can be combined by merging (union) and joining on key columns
- “Lumping” vs. “Splitting” debate
- We take a hybrid approach for long-term monitoring data
- Lumping provides a long-term time series in one data set
- Splitting provides a shorter time-frame of data but retains finer-scale metadata
Discoverability and Interoperability

- Metadata are encoded in EML 2.1 (LTER standard metadata format)
- All processing steps are automatically recorded in the metadata record
- When possible, dataset keywords are mapped to LTER controlled vocabulary and units are mapped to LTER unit dictionary
- Finalized data sets are archived in standard text and MATLAB formats
- Non-tabular data are processed and documented by the investigator and finalized data are archived in domain-specific formats (e.g. shapefiles, raster images, file geodatabase, ...
Data Management Cycle

Data Source → Import Data → Add / Import Metadata → Q/C Analysis → Post-Process → Synthesis → Archive / Publish

Data Availability
Data Availability – Archiving

- Data can be displayed on MATLAB-generated web pages and dashboards
Data Availability – Archiving

- Rapid processing and web posting of data and visualizations from field instruments
Finalized data sets are available through the GCE data catalog.

- Data summaries and metadata are made publicly available immediately.
- The accompanying data files are released to the public within 2 years in compliance with LTER and NSF data access policies.
- At the end of the two-year embargo, data are automatically released and made publicly available.
- Public data are synchronized to the EDI Data Portal monthly for distribution through EDI, DataONE, BCO-DMO and related repositories.
- All GCE data sets comply with LTER network standards and protocols.
Georgia Coastal Ecosystems LTER Information Management

Project Support

Adam Sapp - Information Manager
Wade Sheldon - Information Management System Administrator
University of Georgia
Athens, Georgia
Databases

- **Metabase**
  - Personnel, site geography (study area polygons, point locations), instrumentation
  - Data sets (studies, methods, entities, attributes, files)

- **Submissions**
  - Calendar items, announcements, research applications, file archive

- **Bibliography**
  - Authors, keywords, citations, DOIs, reprints

- **Taxonomic**
  - Taxonomic classification from ITIS, species images
Dynamic Webpage Generation

Web pages are dynamically cross-linked for discovery of related information.
Site Geography

- GIS connected to web site through web services and map interfaces
- Access to site info and research data via links in placemark balloons
File Archives

- We developed a generalized file archive for documents/imagery
- Archive cross-referenced to bibliography and taxonomic database photos
- All GCE members can contribute, update their entries
Research Applications

- Applications reviewed by field crew, IM, lead PIs to manage resources, conflicts
- Ensures compliance with Marsh Protection Act, facilitates CRD/NOAA permitting
- Compatible system developed for SINERR, UGAMI for info sharing
- Field tags generated for display at sites for contact info, ensuring clean up
Bibliography Management

- Researchers can enter their GCE related publications
- Cross-linked to data sets
- Query and export results to various formats (BibTeX, EndNote, Text)
Research Support

- Metadata creation
  - Personnel details, study site locations, taxonomic classifications, instrumentation details

- Site conditions
  - Tide predictions, near real-time weather conditions, hydrographic data plots

- Provisional data pages
  - GCE Members can access provisional data before they are published

- Software and support
  - Maintain an ArcGIS license server and offer support for GIS software
  - Maintain an SVN repository server for software development
Communication

- Web forms allow GCE members to contribute and update information
- Email lists and auto-generated newsletters keep everyone informed
- Host dedicated Education and Outreach WordPress site for Schoolyard
Project Webpages

- Field logs to show what work was done and by whom
- Group calendars to show upcoming field work
- This helps avoid conflicts and promotes collaboration
- Future work – to add group-specific file archives
Synology/Drone Imagery

- Large amounts of imagery created during GCE IV from drone flights and wildlife cameras
- Other remote sensing data (multi-year aerial imagery, multi-spectral flights, thermal IR surveys)
- Automatic synchronization of imagery and videos to Synology server at UGA
- Provide secure access to server for GCE working groups.
Automated Workflow Tracking System
(Steve’s Nag Machine)

- Tracks different parts of our fall monitoring study and
- Sends reminder emails when a task is coming due
- Ensures protocols are reviewed and updated regularly
- Ensures monitoring data are submitted and reviewed in a timely manner