Automating Sensor Data Processing and Q/C with the GCE Data Toolbox

Wade Sheldon (GCE-LTER)

Brian Herndon (CWT-LTER)

Background & Motivation

- Georgia Coastal Ecosystems LTER started in 2000
  - Needed to standardize data processing, quality control, documentation
- No ready-to-use software for LTER data management
  - Lots of great papers and reports, no tools to download
  - Most LTER sites were using “flat files” – limiting
  - Relational databases emerging – proprietary, complex, require network
- Developed custom data management framework in MATLAB
  - Experienced using MATLAB for automating data processing, GUIs
  - Better code-reuse potential than database/web solution
  - Best compromise: file-based but supports fully dynamic operations
- Shared software with other sites, scientists
  - 2002 – compiled version publicly released
  - 2009 – adopted by CWT
  - 2010 – established Wiki site (Trac), released as open source (GPL)
  - 2013 – ARRA funding for training, enhancement
  - 2015 – over 4000 downloads, in use at ~8 LTER sites

What is MATLAB?

- From Mathworks (www.mathworks.com):

  "MATLAB is a programming environment for algorithm development, data analysis, visualization, and numerical computation."

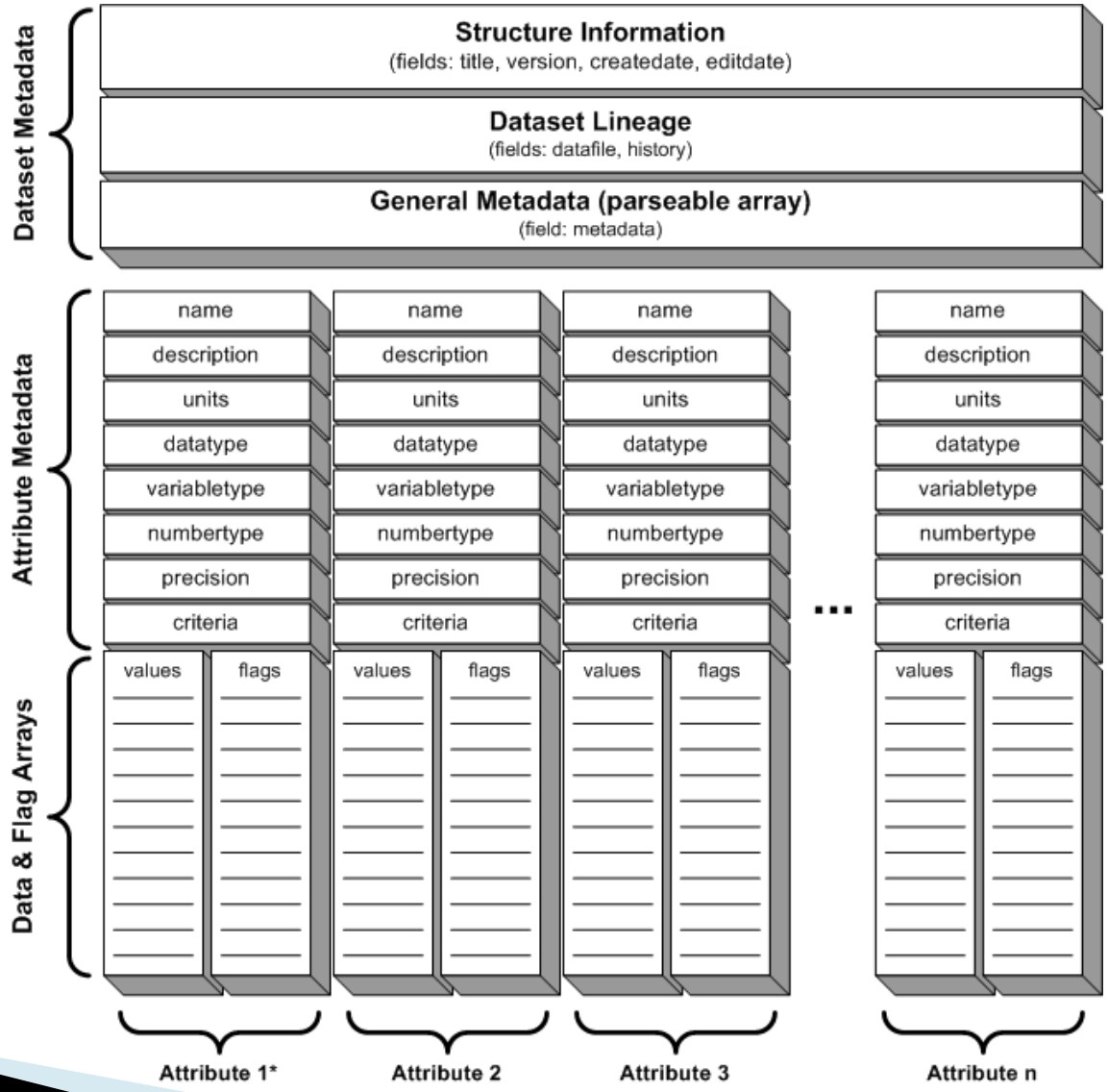
- Benefits
  - Ubiquitous in engineering and many science branches
  - Site licensed by many institutions
  - Rapid development, pre-built functionality, Java integration
  - Code, GUIs and data formats multi-platform (Windows, Unix, Mac)
  - Stable: good support and backward compatibility (~30 year history)
  - Scalable (netbook to HPC) - good performance with huge data sets
  - Broad I/O support (serial ports to web services)

- Costs
  - Commercial ("licensed source") - limits flexibility, costs $-$$$
  - Some programming required for maximum use

Toolbox Development

- Designed generalized tabular data model in MATLAB
  - 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
- Requirements based on 1995 ESA FLED report
  - Gross, Katherine L. and Catherine E. Pake. 1995. Final report of the Ecological Society of America Committee on the Future of Long-term Ecological Data (FLED). Volume I: Text of the Report. The Ecological Society of America, Washington, D.C.

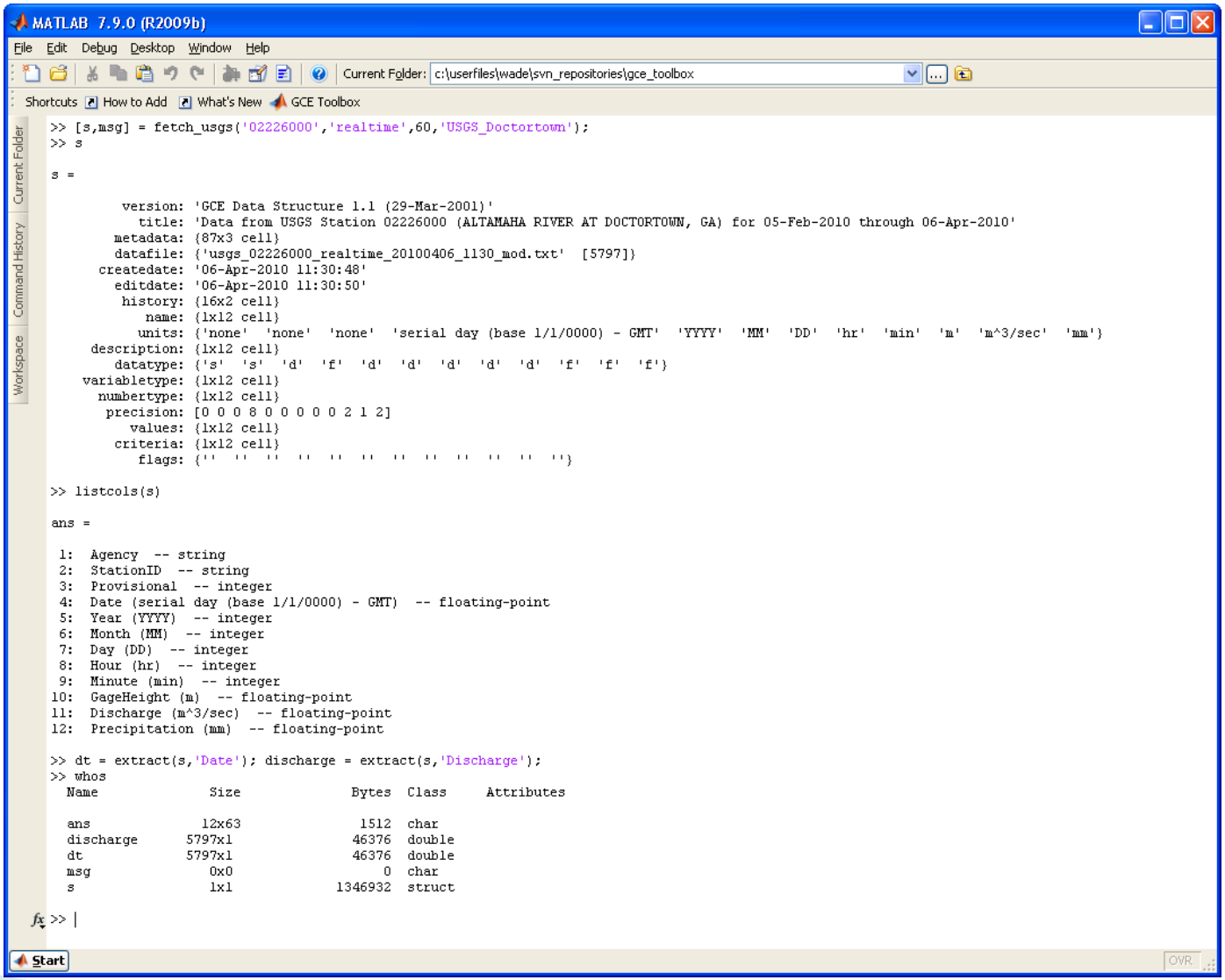
Data Model (GCE Data Structure)



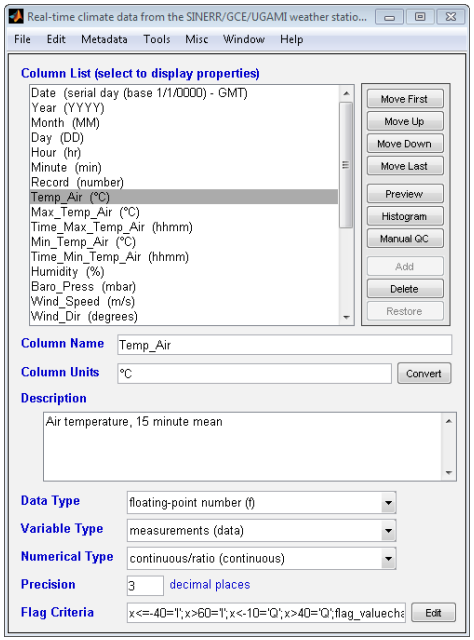
Toolbox Development

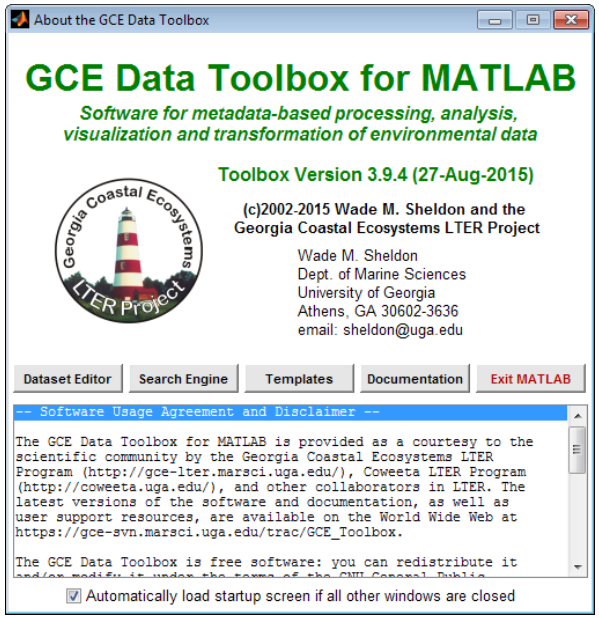
- Developed MATLAB software library ("toolbox")
  - Utility functions to abstract low-level operations (API)
    - Create structure, add/delete columns, copy/insert/delete rows
    - Extract, sort, query, update data, update flags
  - Analytical functions for high-level operations
    - Statistics, visualizations, geographic & date/time transformations
    - Unit inter-conversions, aggregation/re-sampling, joining data sets
  - GUI interface functions to simplify using the toolbox
  - All functions use metadata, data introspection to auto-parameterize and automate operations (semantic processing)
- Indexing and search support (and GUI search engine)
- Data harvest management tools

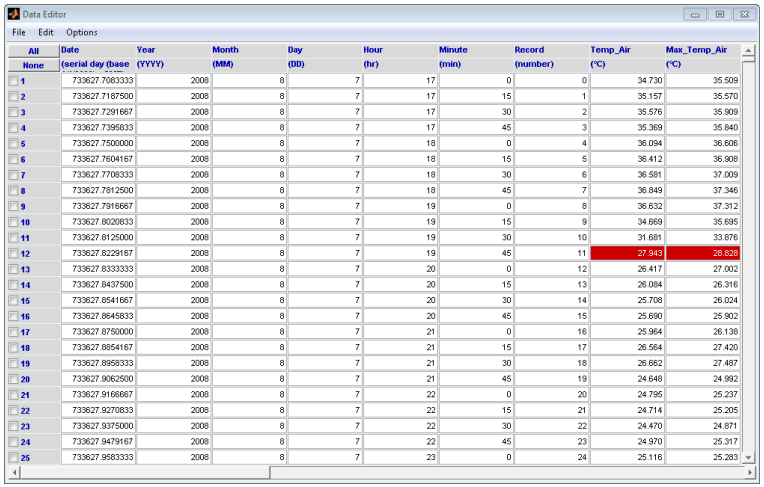
Command Line Interface (API)

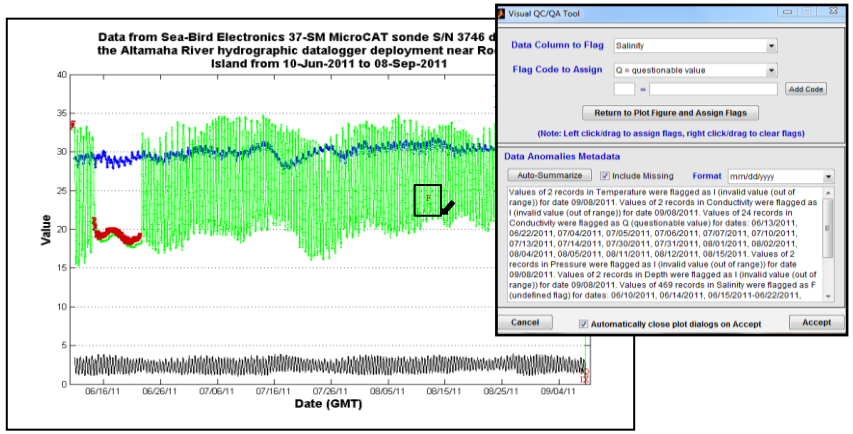


GUI Application Interface

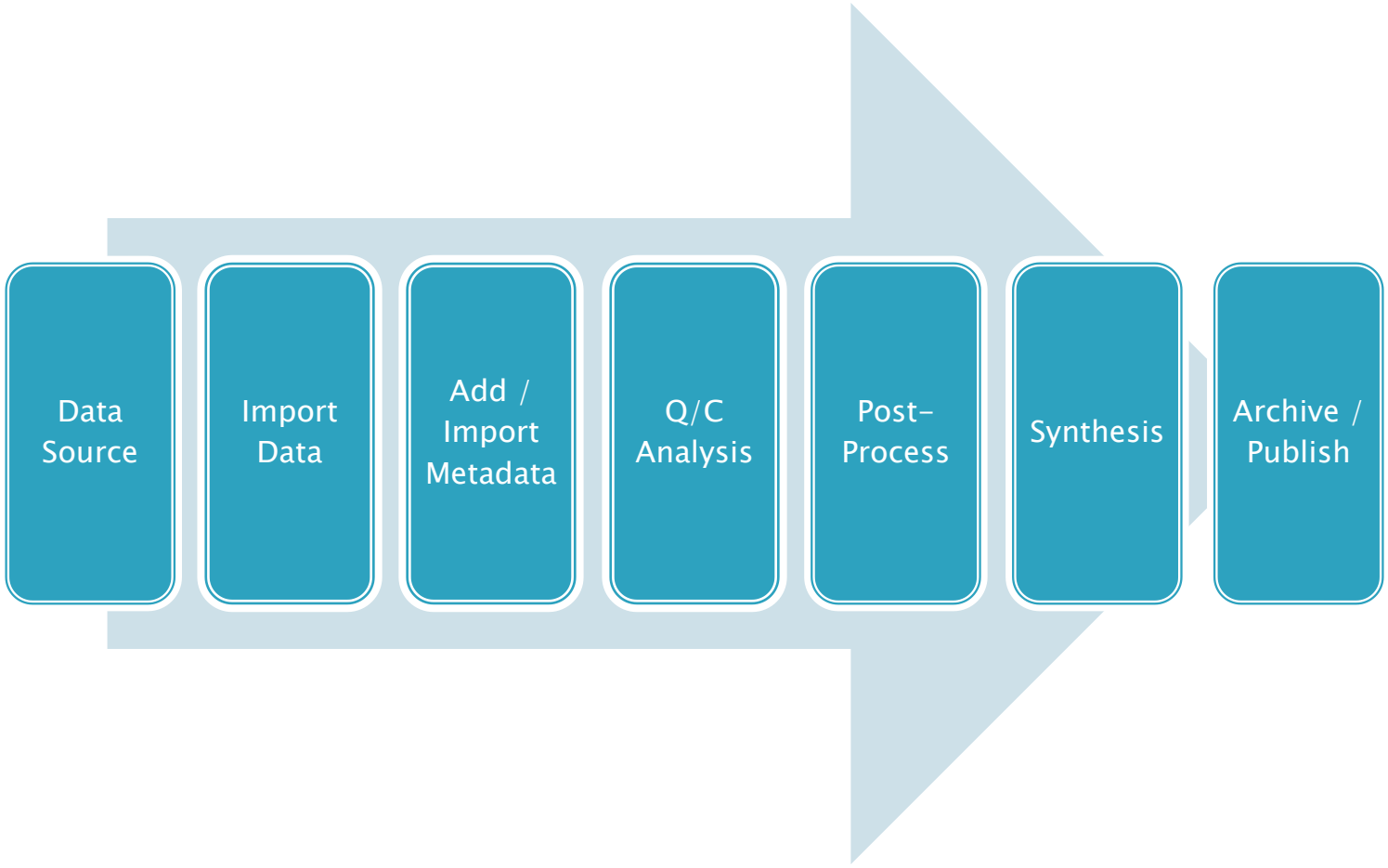








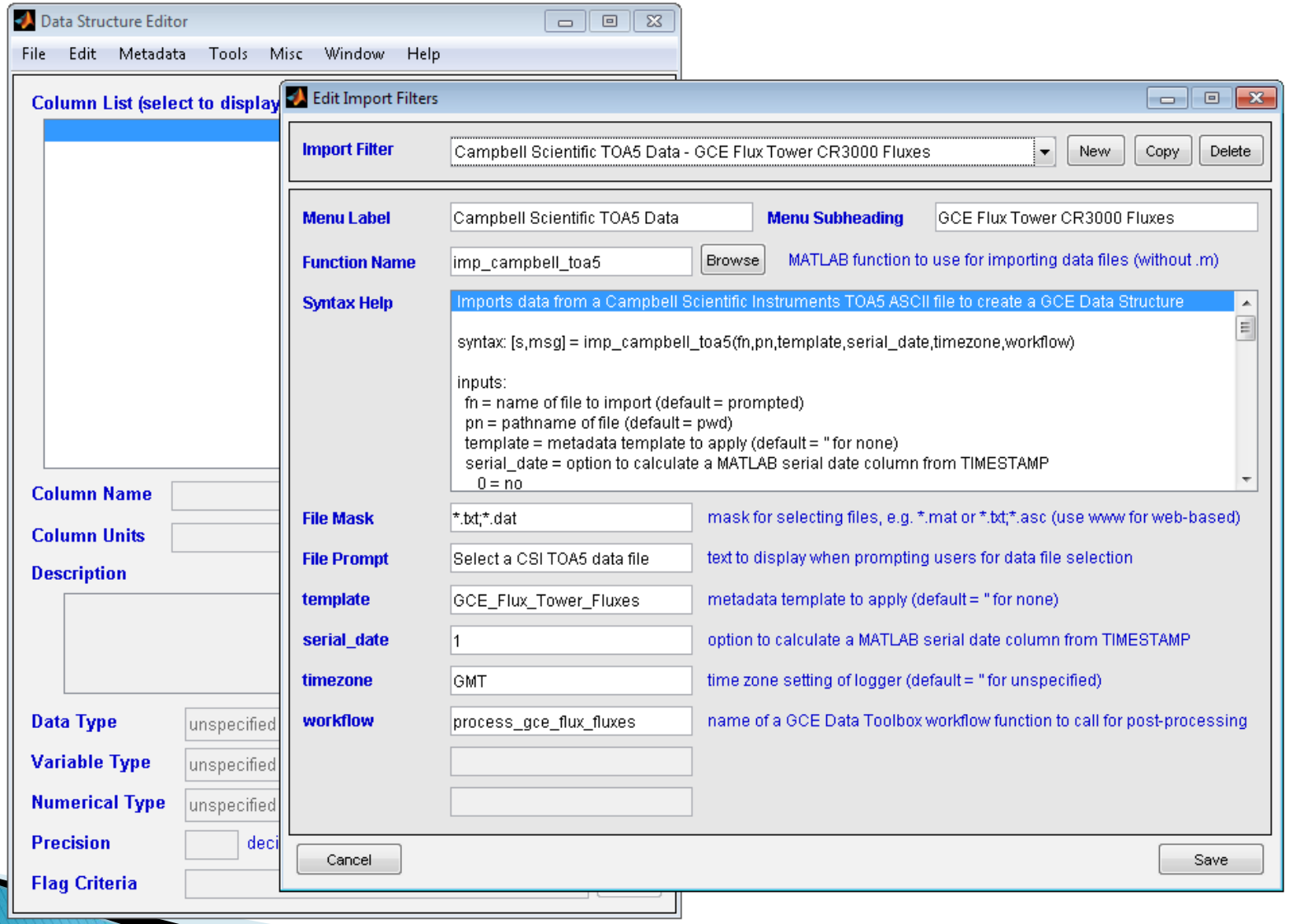
Data Management Cycle



Importing Data

- Generic parsers
  - Delimited text (CSV, space, tab)
  - MATLAB variables (arrays, matrices, structs)
- Specialized parsers
  - Vendor-specific logger files
    - Campbell Scientific Instruments (tables, arrays)
    - Sea-Bird CTD, sondes
    - Others (Hobo, Schlumberger, OSIL, ...)
  - Network data sources
    - SQL database queries (JDBC)
    - Federal databases (USGS NWIS, NOAA NCDC/GHCN, NOAA HADS)
    - LTER ClimDB/HydroDB
    - EML repositories (LTER NIS/PASTA, KNB)
    - Data Turbine servers
- Custom parsers can be added

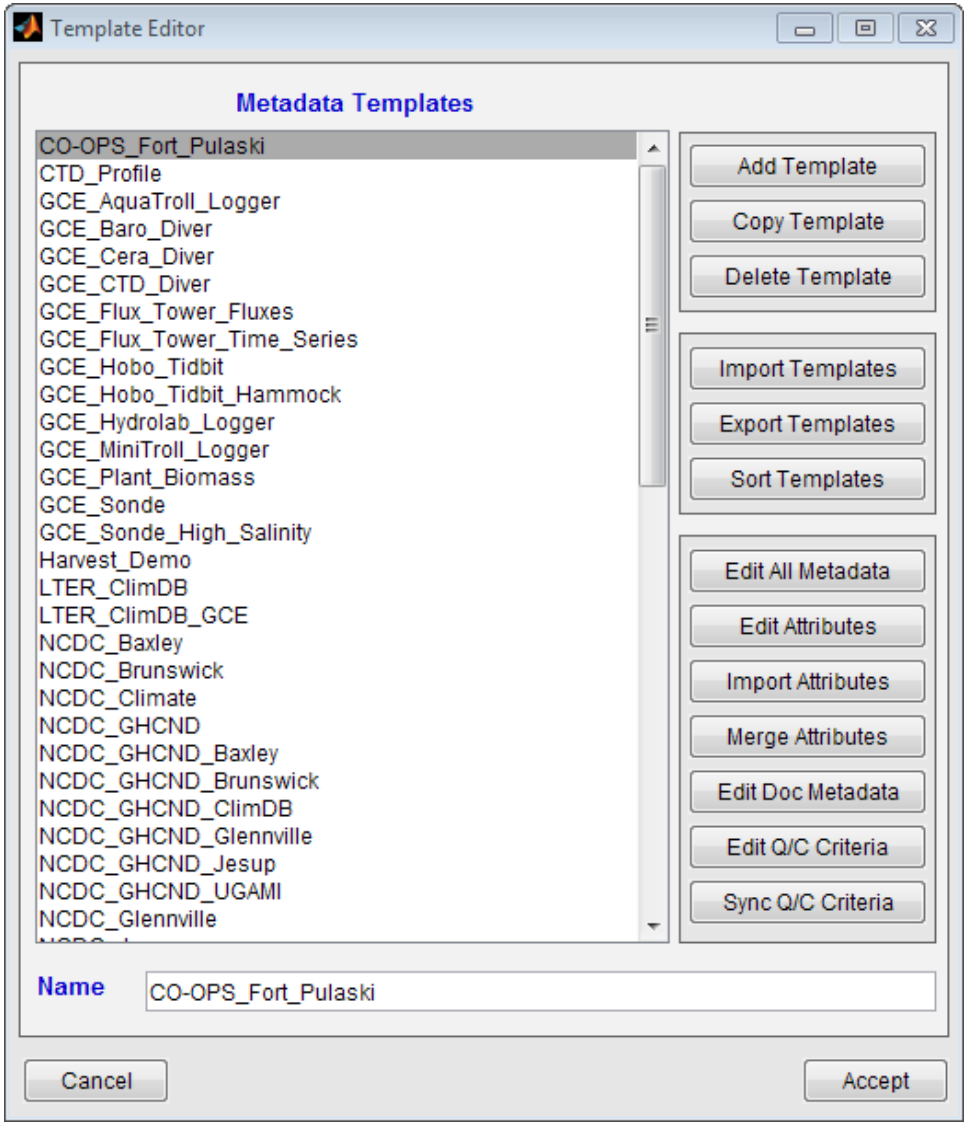
Importing Data



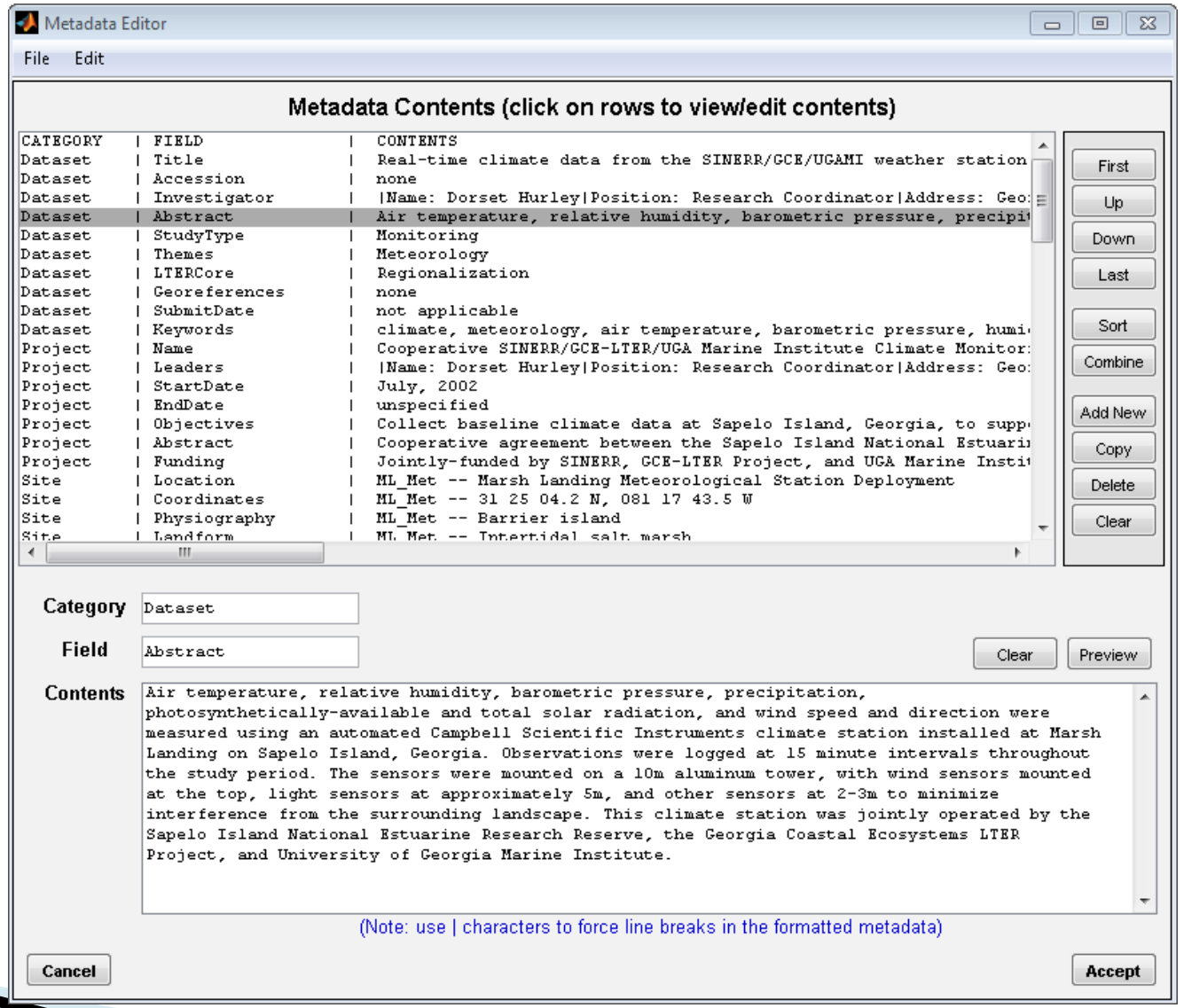
Adding/Importing Metadata

- Metadata entry is time-consuming (and tedious)!
  - Metadata capture, re-use strongly emphasized
  - Metadata auto-generated whenever possible
- Many pathways for building Metadata
  - Metadata can be imported along with data
    - Logger file headers (Campbell, Sea-bird)
    - Station, parameter information from USGS, NOAA
    - Tokenized headers from Data Submission Template
  - Metadata imported from other GCE toolbox data structures
  - Metadata imported from data repository (EML/XML)
  - Metadata added from user-defined “Templates”
    - Column (attribute) metadata matched to “variables”
    - Boilerplate documentation
    - GUI tool for creating/managing templates

Adding/Importing Metadata



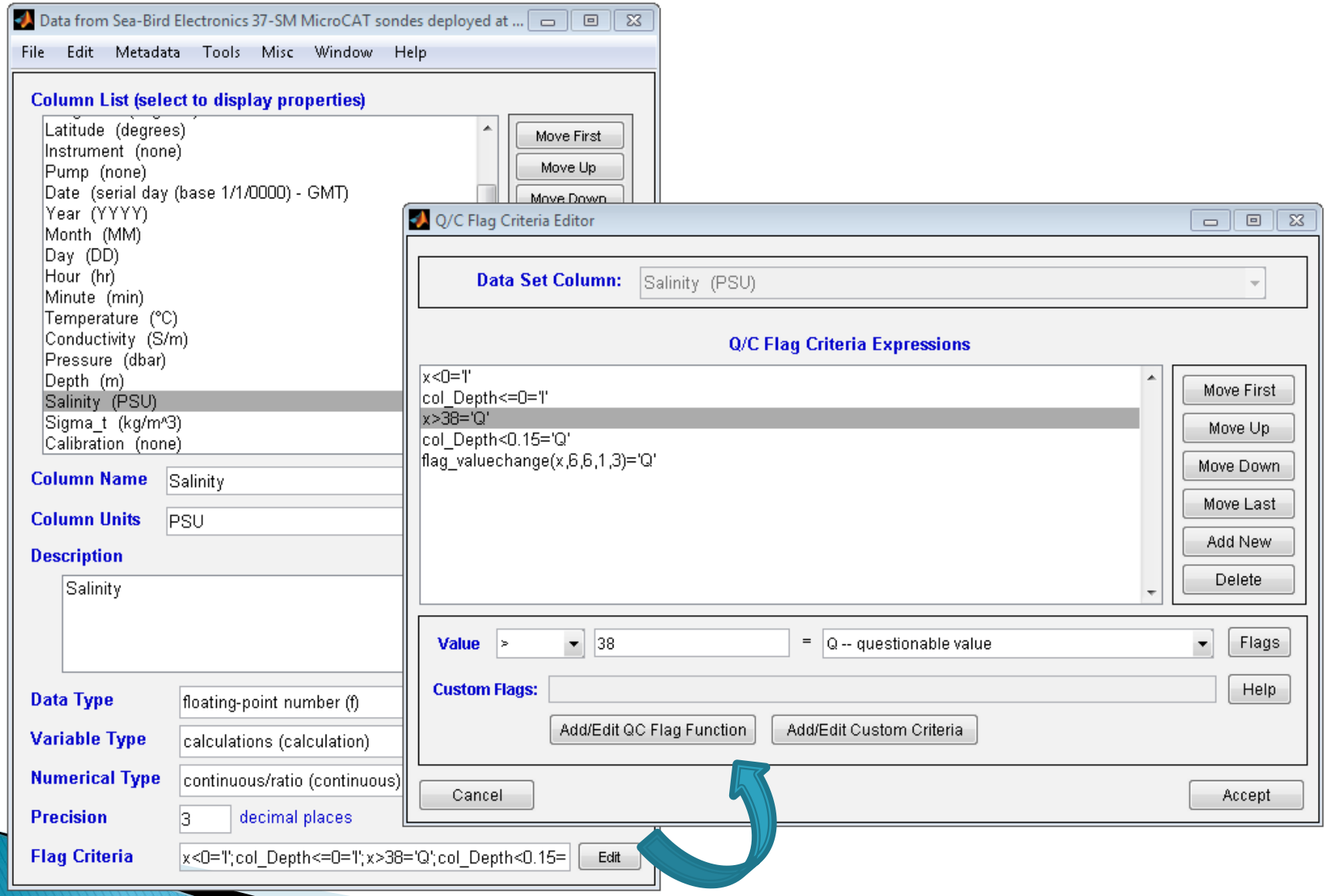
Adding/Importing Metadata



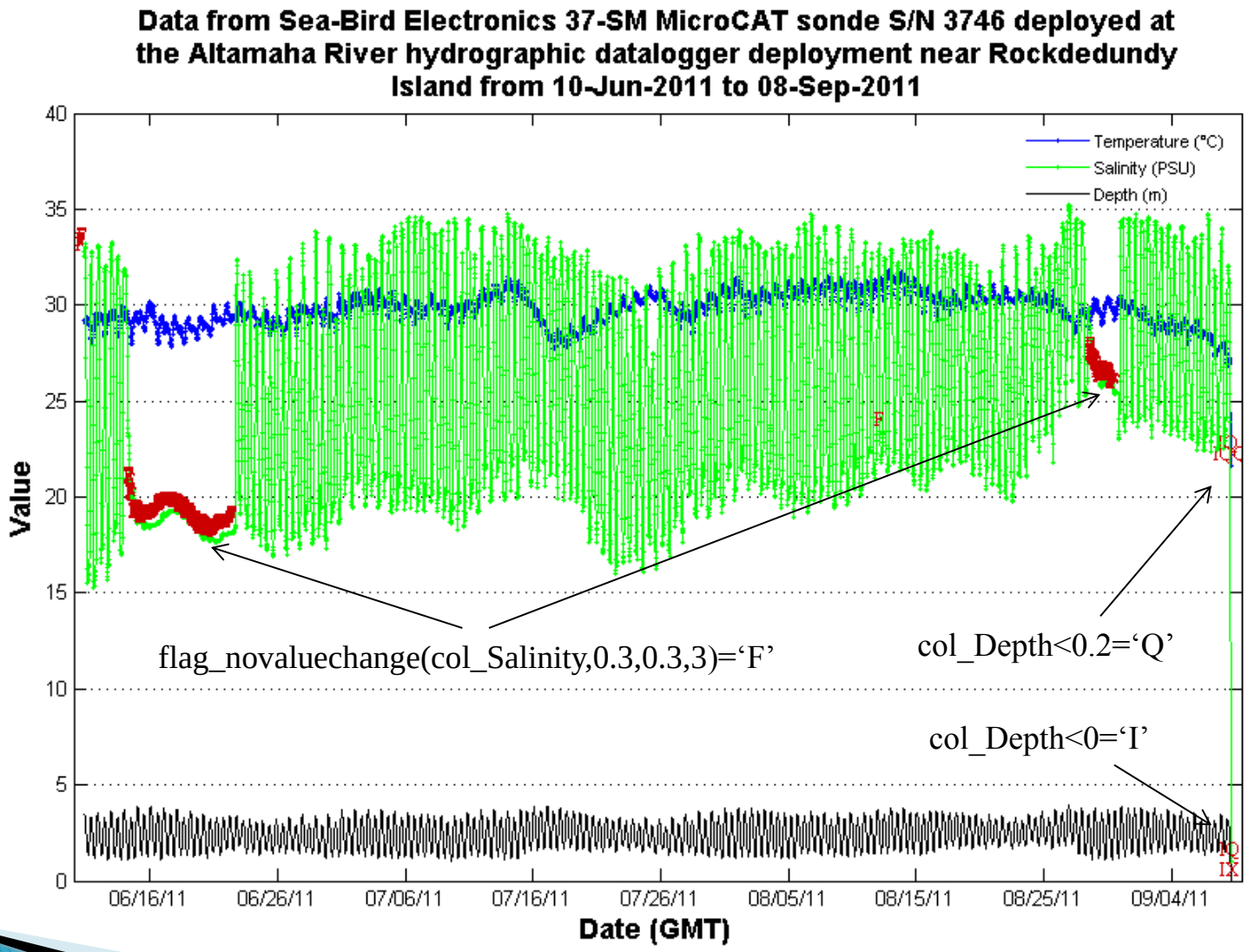
Q/C Analysis Framework

- Programmatic Q/C Analysis
  - "Rules" (i.e. criteria) 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
- Automatic Documentation of Q/C Steps
  - Q/C operations (including revisions) logged to processing lineage
  - Data anomalies reports can be auto-generated and annotated
- Data analysis, synthesis tools Q/C-aware
  - Qualified values can be filtered, summarized, visualized during analysis
  - Statistics about missing/qualified values tabulated, used to qualify derived data

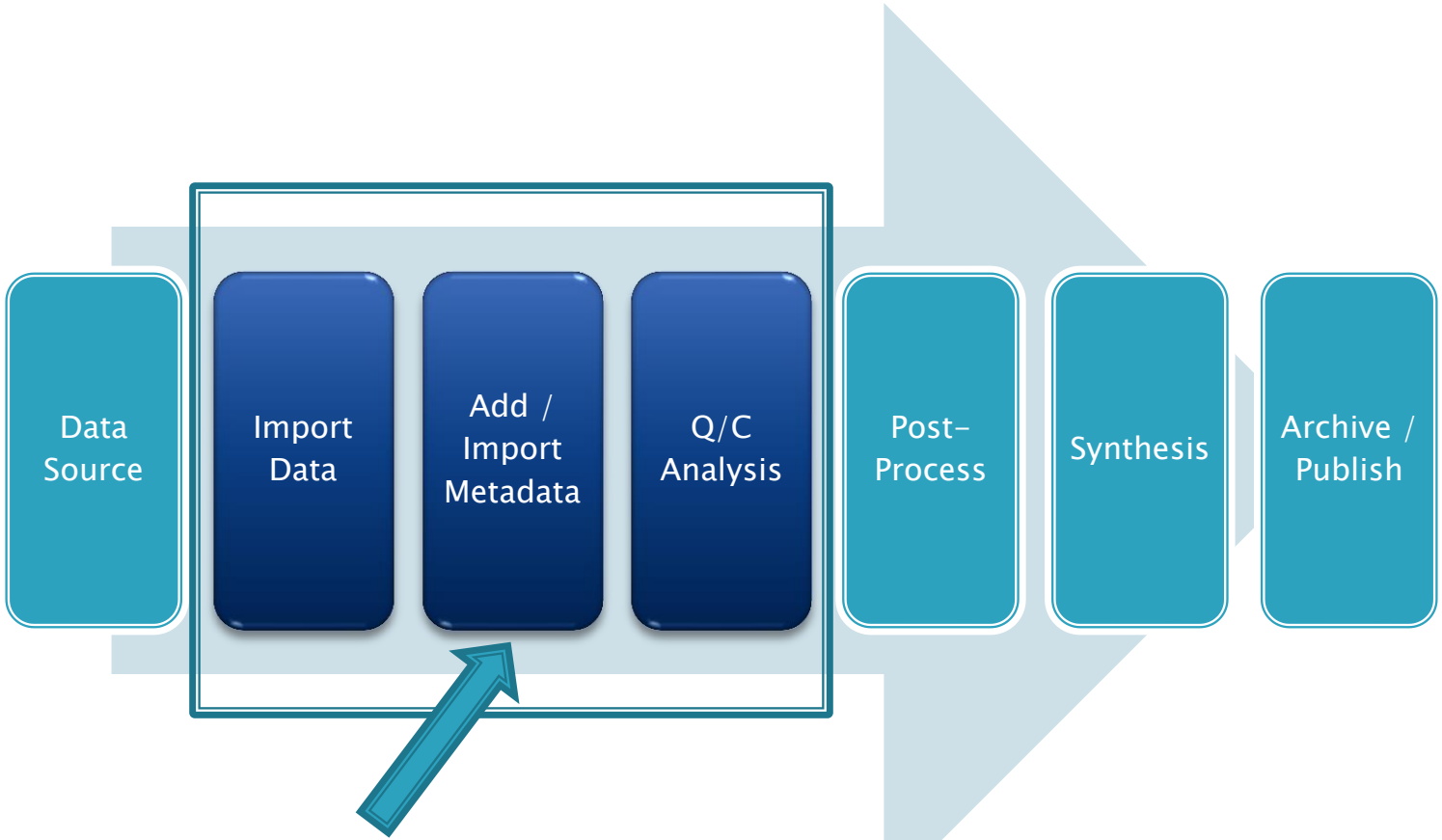
QA/QC Analysis Framework



Example – CTD mooring



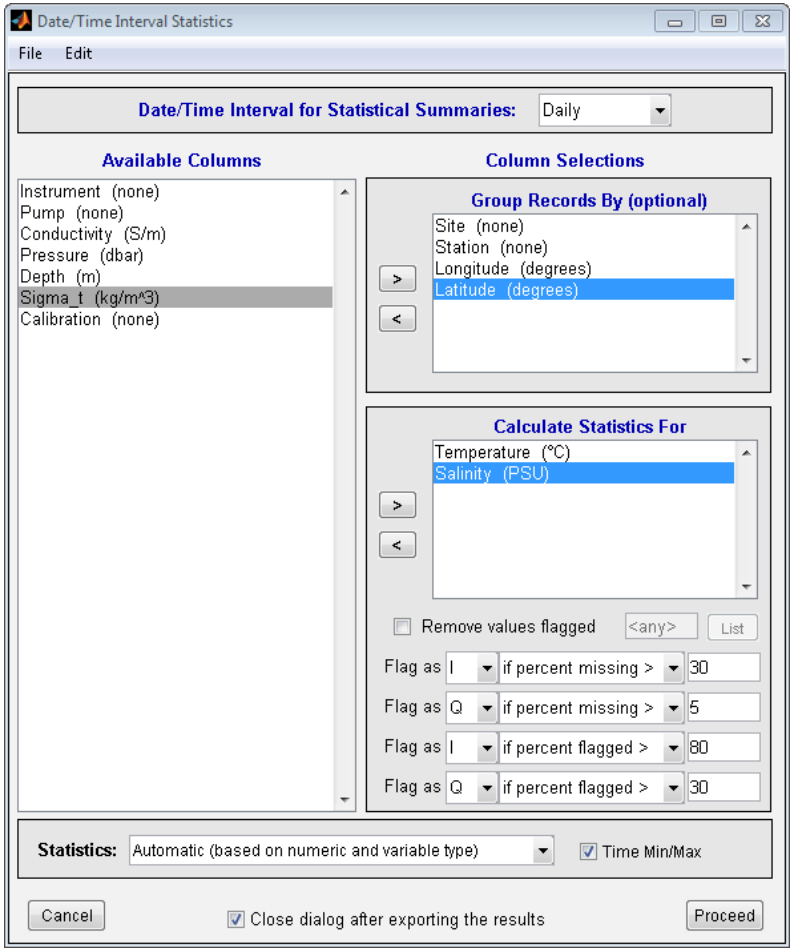
Data Management Cycle



Import Filter + Metadata Template
(One Step)

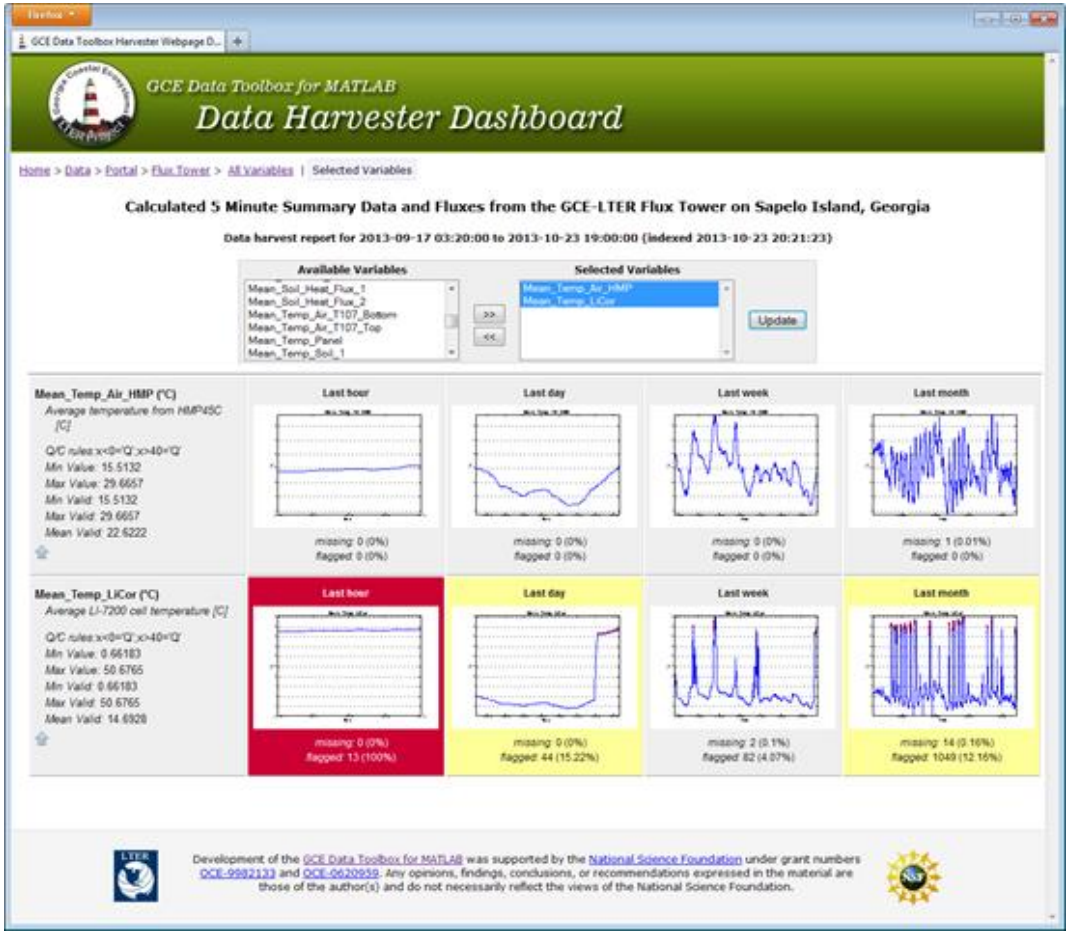
Post-Processing and Synthesis

- 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
- Multiple data sets can be combined by merging (union) and joining on key columns
- All derived data contain complete metadata describing the entire processing history
- Q/C rules can be generated for derived data columns automatically based on number or percent missing/flagged values



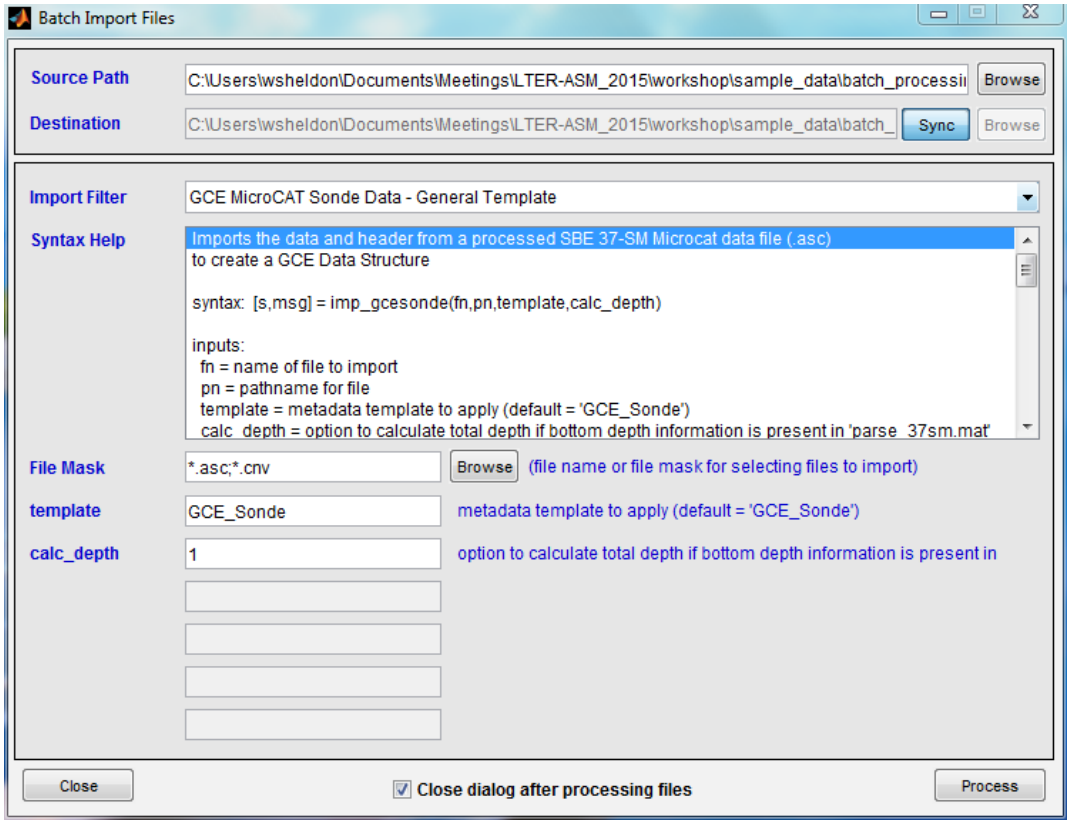
Archive/Publish Data

- Data and metadata can be exported in many formats
  - Text, MATLAB, XML/KML/HTML
  - RDBMS tables
- EML-described data packages can be generated/published in data archives
  - LTER NIS/PASTA
  - DataONE Member Nodes (e.g. KNB)
  - *DEIMS*
- Data can be refactored and published in CUAHSI ODM database, Hydroserver
- Data can be displayed on MATLAB-generated web pages and dashboards

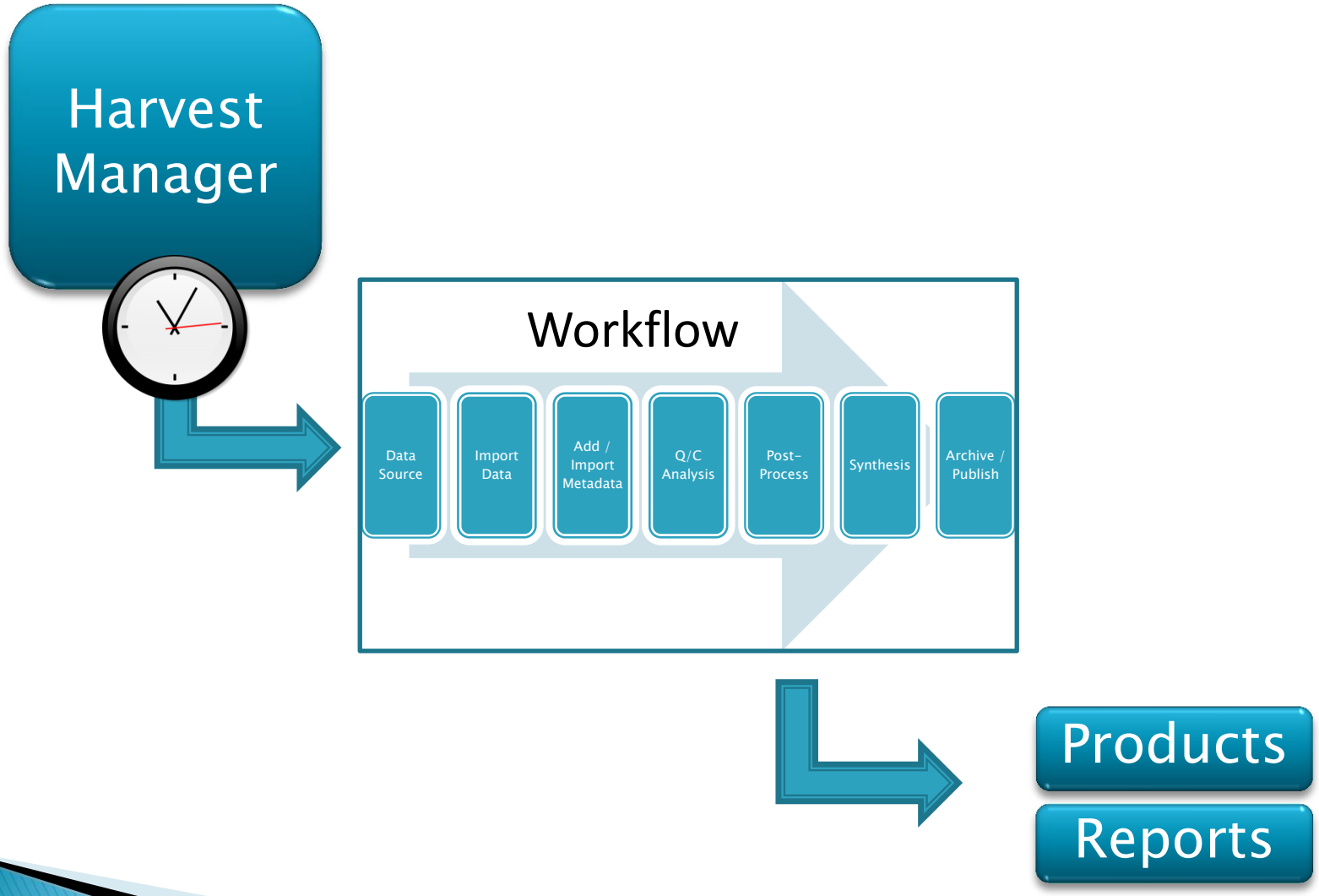


Automation – Batch Processing

- Entire directories of raw data files can be processed at once using an import filter and metadata template
- Multiple GCE Data Structure files in a directory can be batch-converted to TXT, CSV
- Multiple files can be merged via metadata-based union to create an integrated data set
- All directories containing GCE Data Structures can be indexed and searched, then merged, joined and exported



Automation – Data Harvesting



Key Concepts

- Operations performed in context of a "dataset"
  - Passing data columns to a tool transports metadata as well
  - Dataset metadata used to guide transformation, plotting, analysis
  - Metadata used to auto-parameterize functions
- Workflow steps generate new, complete datasets
  - Each step along a workflow results in a complete data set with metadata
  - Intermediate datasets can be saved or overwritten in workflows
- Processing history ("lineage") tracked
  - Each tool logs operations by date/time
  - Data revisions, deletions, flagging captured at user-specified detail
  - Lineage reported in metadata
- Metadata are "live", and updated automatically
  - Data column metadata and data revisions
  - Calculations and unit conversions
  - Code definitions
  - Metadata merged when multiple data sets are joined

Implementation Scenarios

- End-to-End Processing (logger-to-scientist)
  - Acquire raw data from logger, file system, network
  - Assign metadata from template or forms to validate and flag data
  - Review data and fine-tune flag assignments
  - Generate distribution files & plots, archive/publish data
  - Scientists can use toolbox on desktop to analyze, integrate data
- Data Processing and Q/C
  - Acquire, validate and flag raw data (on demand or timed/triggered)
  - Upload processed data files (e.g. csv) or values & flags to RDBMS (e.g. DEIMS, ODM Database)
- Workflow Step
  - Call toolbox from other software as part of workflow (tool-chaining)
  - Use toolbox as middleware between other systems (e.g. Data Turbine & ODM, Kepler, Taverna)

Resources

- MATLAB
  - Website: http://www.mathworks.com/products/matlab/
  - Version R13 (2002) or higher required (full or student version)
- Software Distribution
  - Website: https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox
    - Documentation, Tutorial, FAQ
    - Distribution Downloads (stable, beta)
    - Bug reporting (tickets)
  - SVN: https://gce-svn.marsci.uga.edu/svn/GCE_Toolbox/trunk
  - Code is open source (GPLv3) and cross-platform
- User Support
  - Peer-to-peer model
  - Sporadic training opportunities (LTER)
  - Email: gcetoolbox-l@listserv.uga.edu *(http://www.listserv.uga.edu/)*