GCE Data Toolbox for MATLAB

Wade Sheldon

Georgia Coastal Ecosystems LTER

University of Georgia

John Chamblee & Richard Cary

Coweeta LTER

University of Georgia

Background & Motivation

- Georgia Coastal Ecosystems LTER project started in Sept 2000
  - Large and varied data collection effort
  - NSF & LTER require data and metadata archiving, sharing
  - 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 constant network access
- Developed custom data management framework using 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

What is MATLAB?

- From Mathworks (www.mathworks.com):

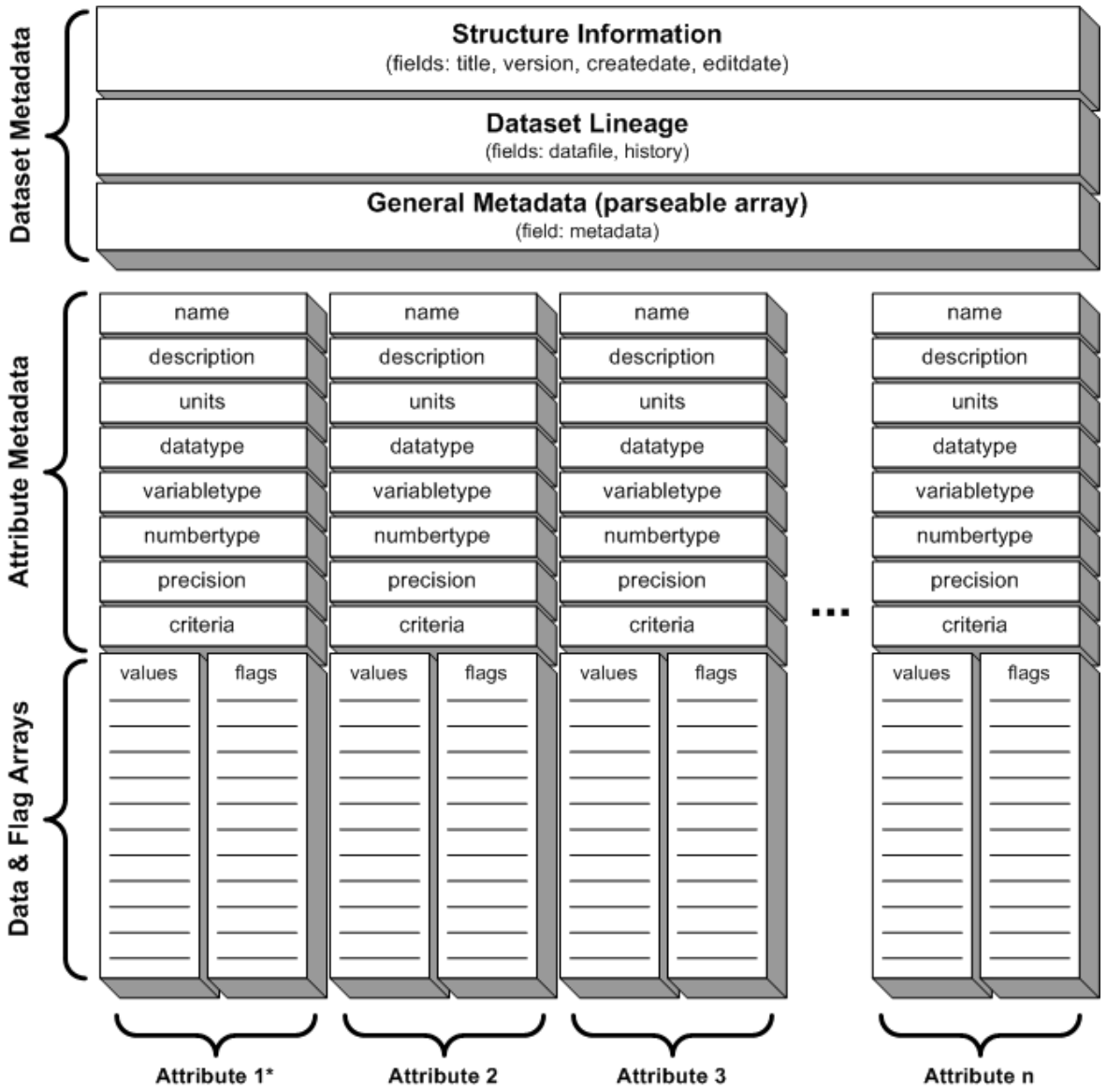
  "MATLAB is a programming environment for algorithm development, data analysis, visualization, and numerical computation."

- Benefits
  - Ubiquitous in engineering and many science branches
  - Rapid development with lots of pre-built functionality, Java integration
  - Code, GUIs and data formats multi-platform (Windows, *nix, Mac OS/x)
  - Stable: good support and backward compatibility (~30 year history)
  - Scalable (netbook to HPC cluster) - 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

- Used ESA's "FLED" report to guide approach
  - 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.
- Identified information storage requirements
  - Any number of numeric and text variables
  - Attribute metadata for each variable (name, units, description, 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
- Designed data model: "GCE Data Structure"
  - MATLAB "struct" array with named fields for each class of information
  - Detailed specifications for allowed content in each field
  - "Virtual table" design based on matched arrays for linking attribute metadata, data, flags
  - Same philosophy as relational database table plus additional descriptors

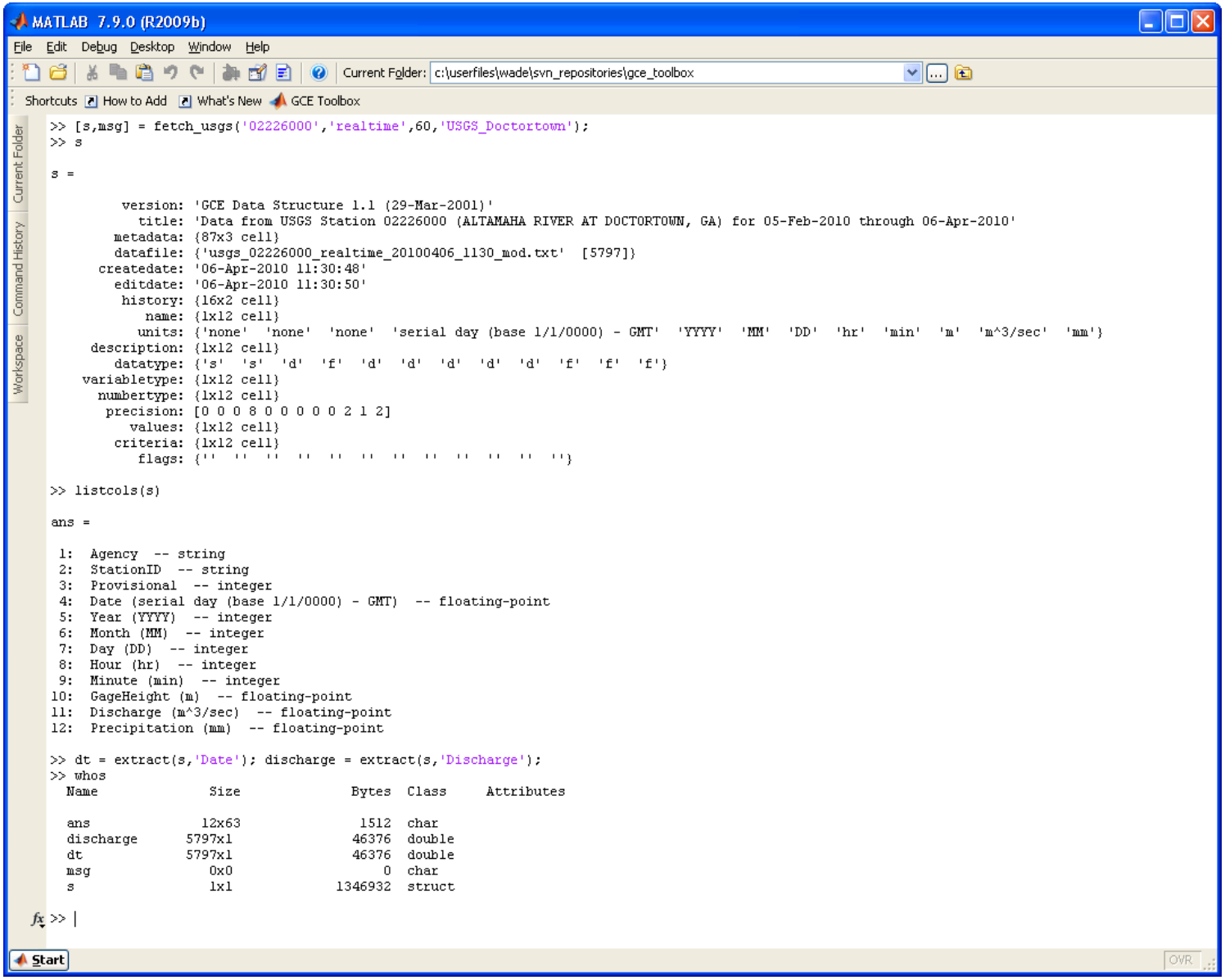
Data Model (GCE Data Structure)



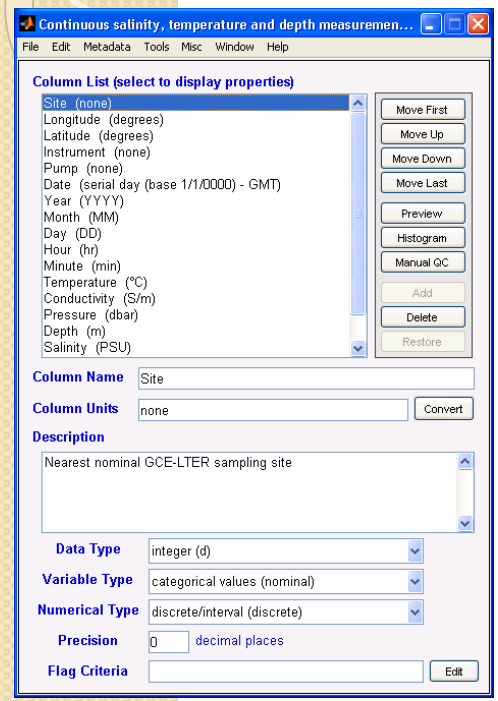
Toolbox Development

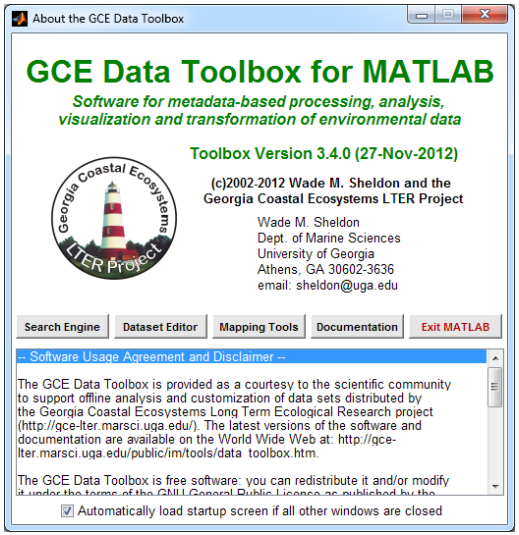
- Developed MATLAB software library to work with data structures
  - Utility functions to abstract low-level operations (API)
    - Create structure, add/delete columns, copy/insert/delete rows
    - Extract, sort, query, update data, update flags
  - Analytical functions for high-level operations
    - 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)
- Developed indexing and search support (and GUI search engine)
- Developing data harvest management tools

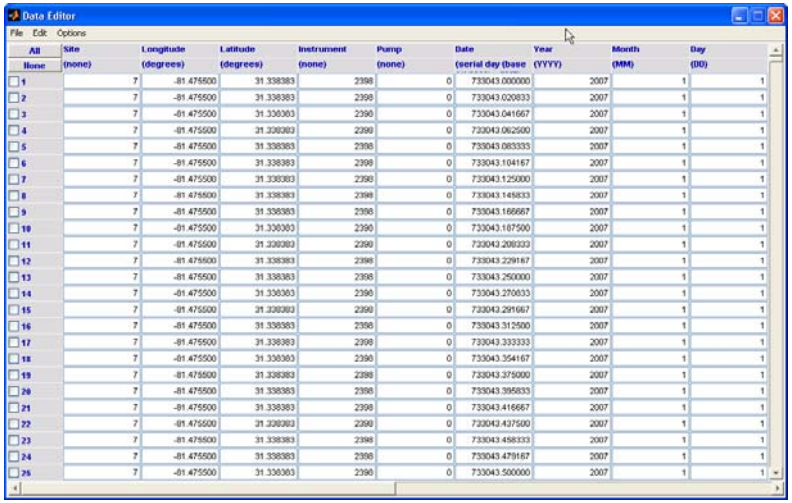
Command Line Interface (API)

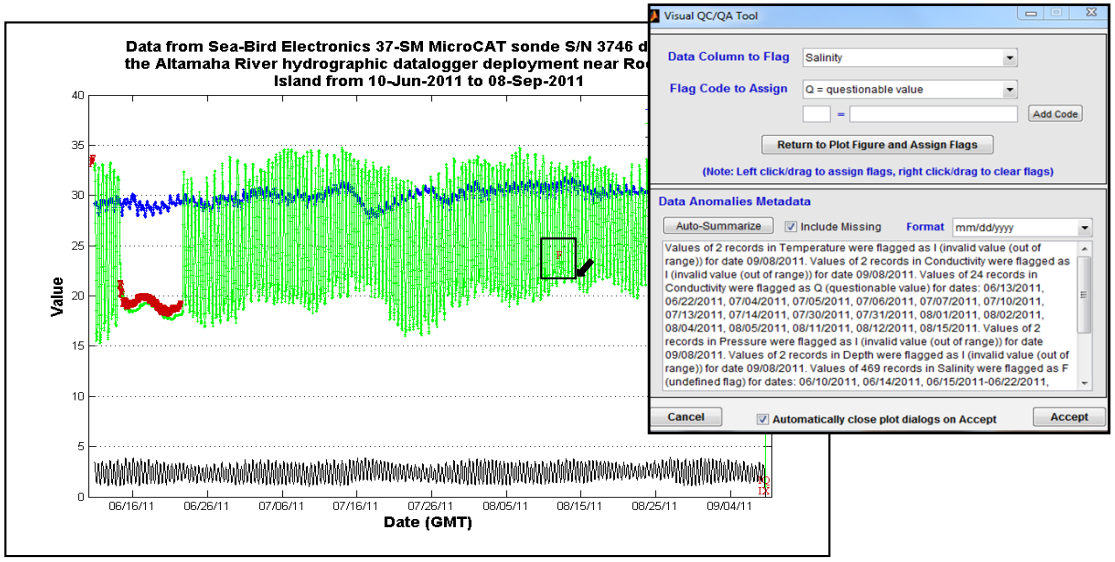


GUI Interface (Application)

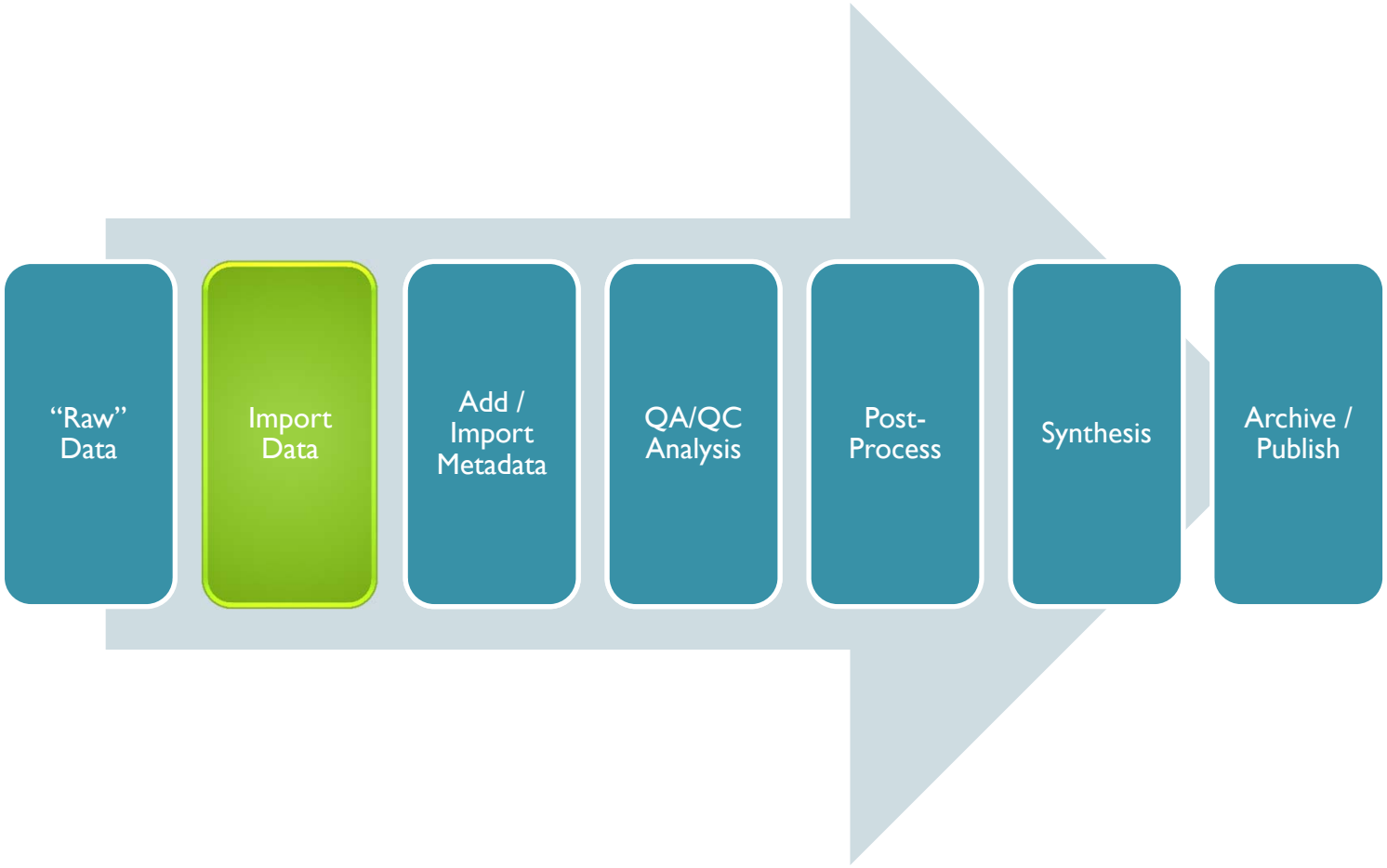








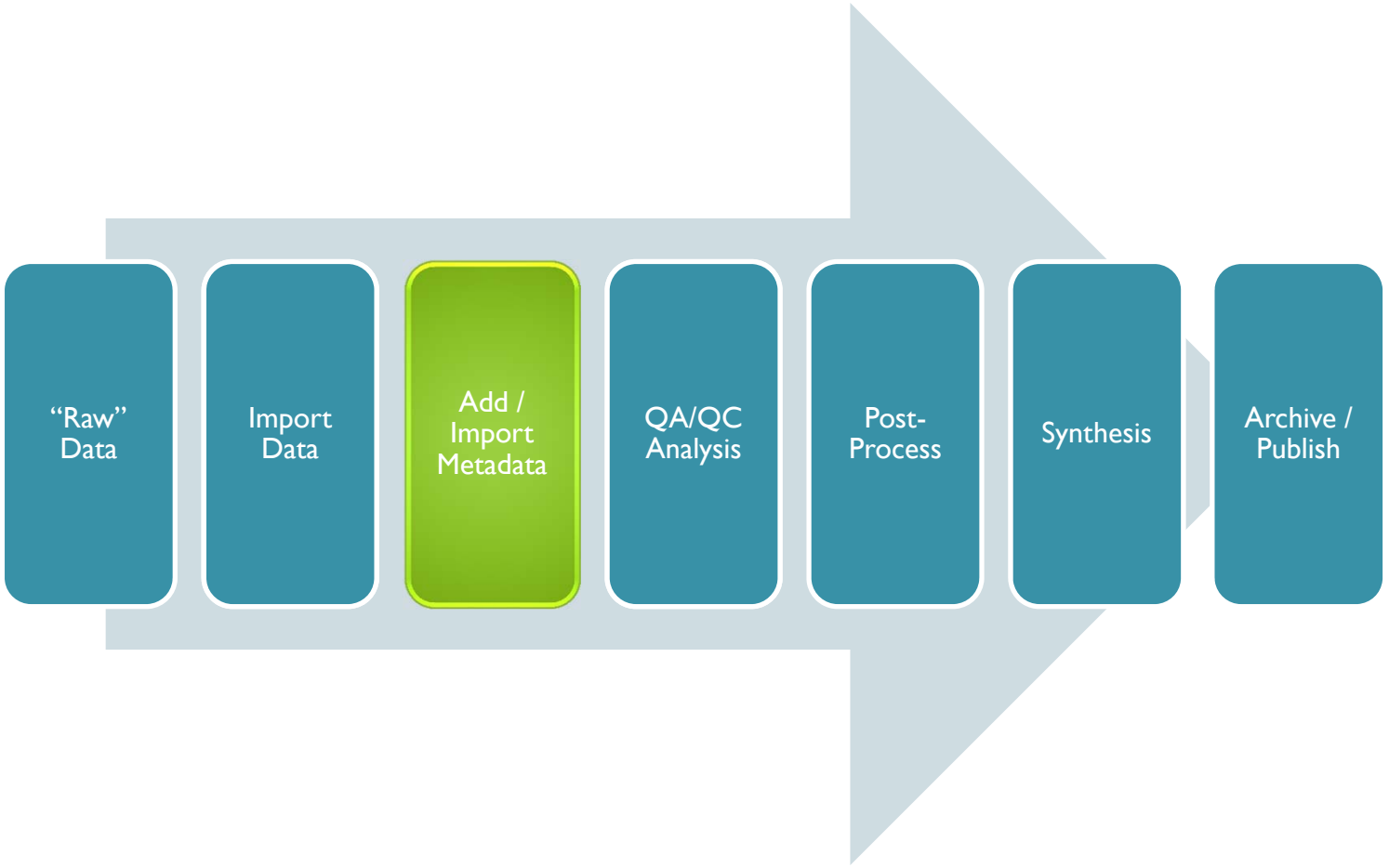
Data Management Cycle



Importing Data

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

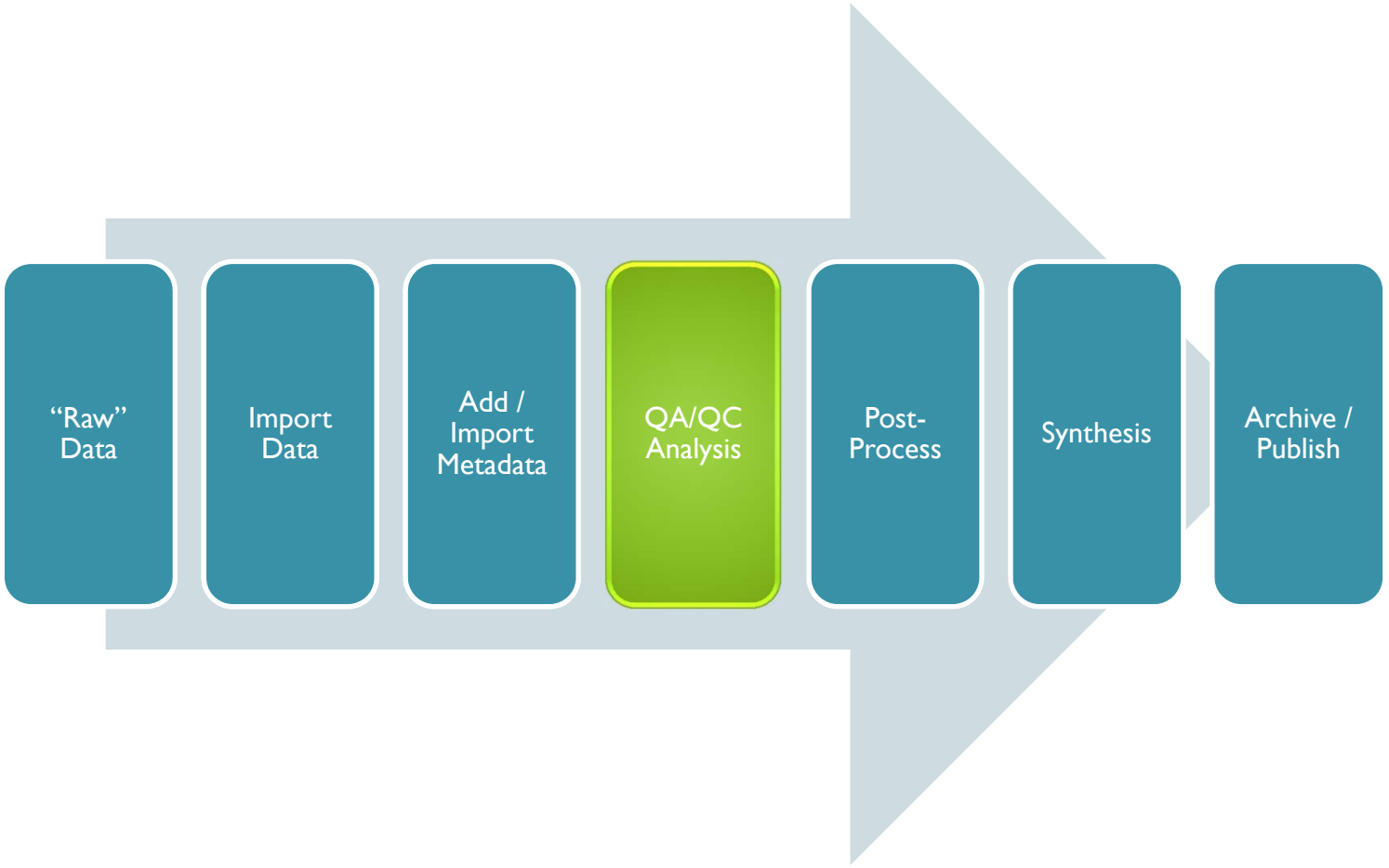
Data Management Cycle



Adding/Importing Metadata

- Metadata imported along with data
  - Logger file headers (Campbell, Sea-bird)
  - Tokenized headers from Data Submission Template
- Metadata imported from other data structures
- Metadata imported along with EML (XML)
- Metadata added from "Templates"
  - Column (attribute) metadata matched to "variables"
  - Boilerplate documentation
  - GUI tool for creating/managing templates
- Metadata capture, re-use emphasized

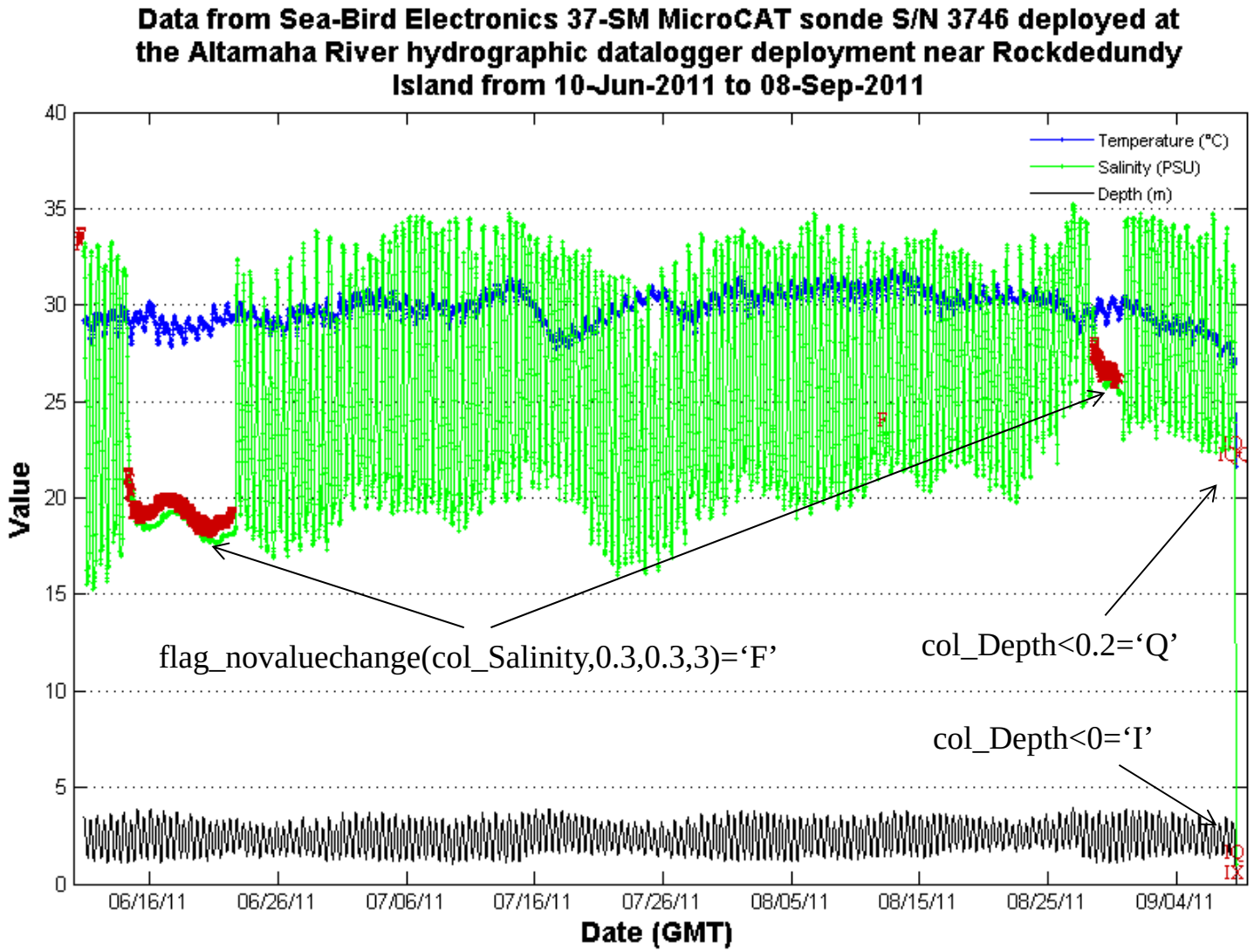
Data Management Cycle



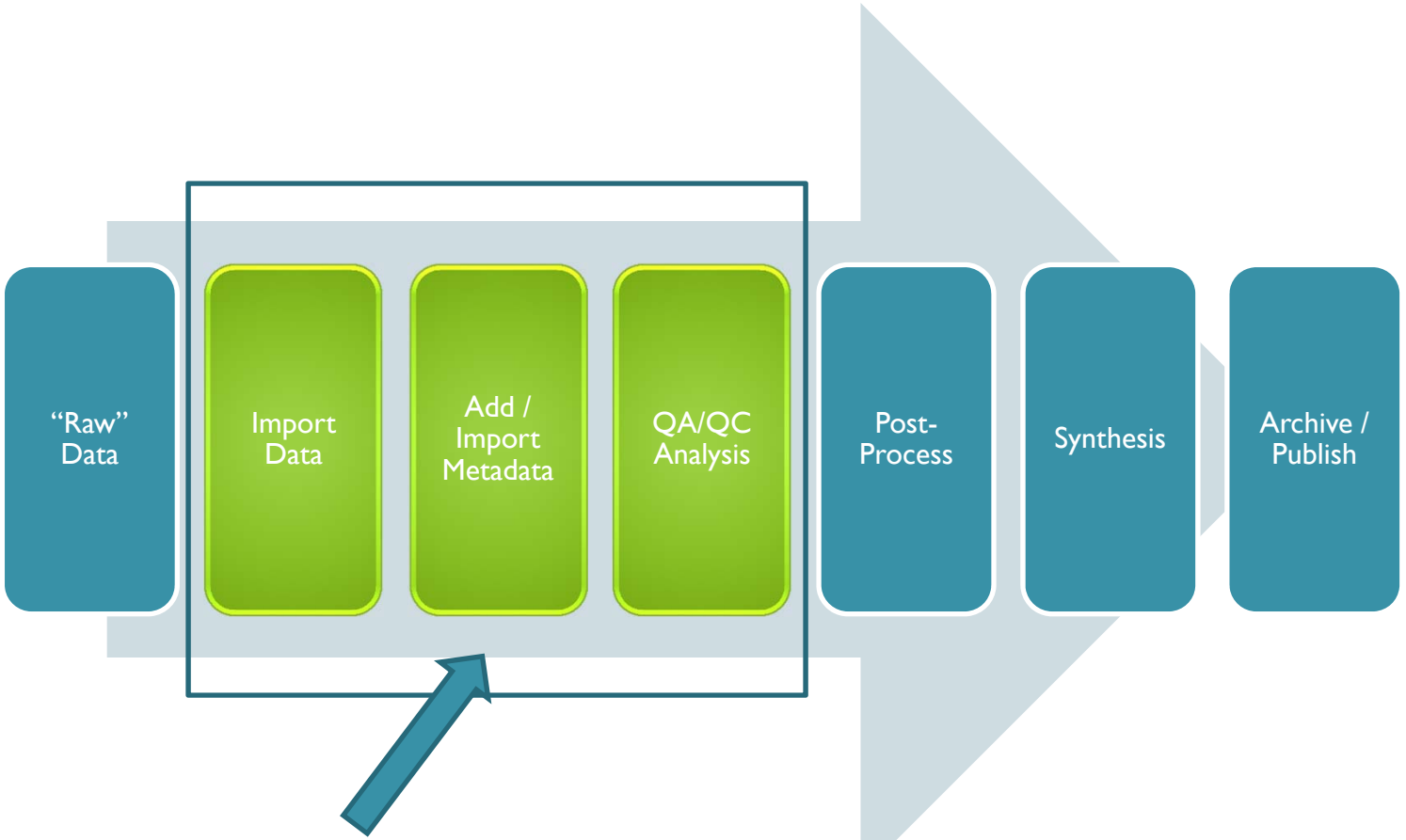
QA/QC Analysis Framework

- Programmatic Q/C Analysis
  - "Rules" that 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 can be 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
  - All Q/C operations (including revisions) logged to processing lineage
  - Data anomalies reports can be auto-generated and annotated to capture rationale
- 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

Example – CTD mooring

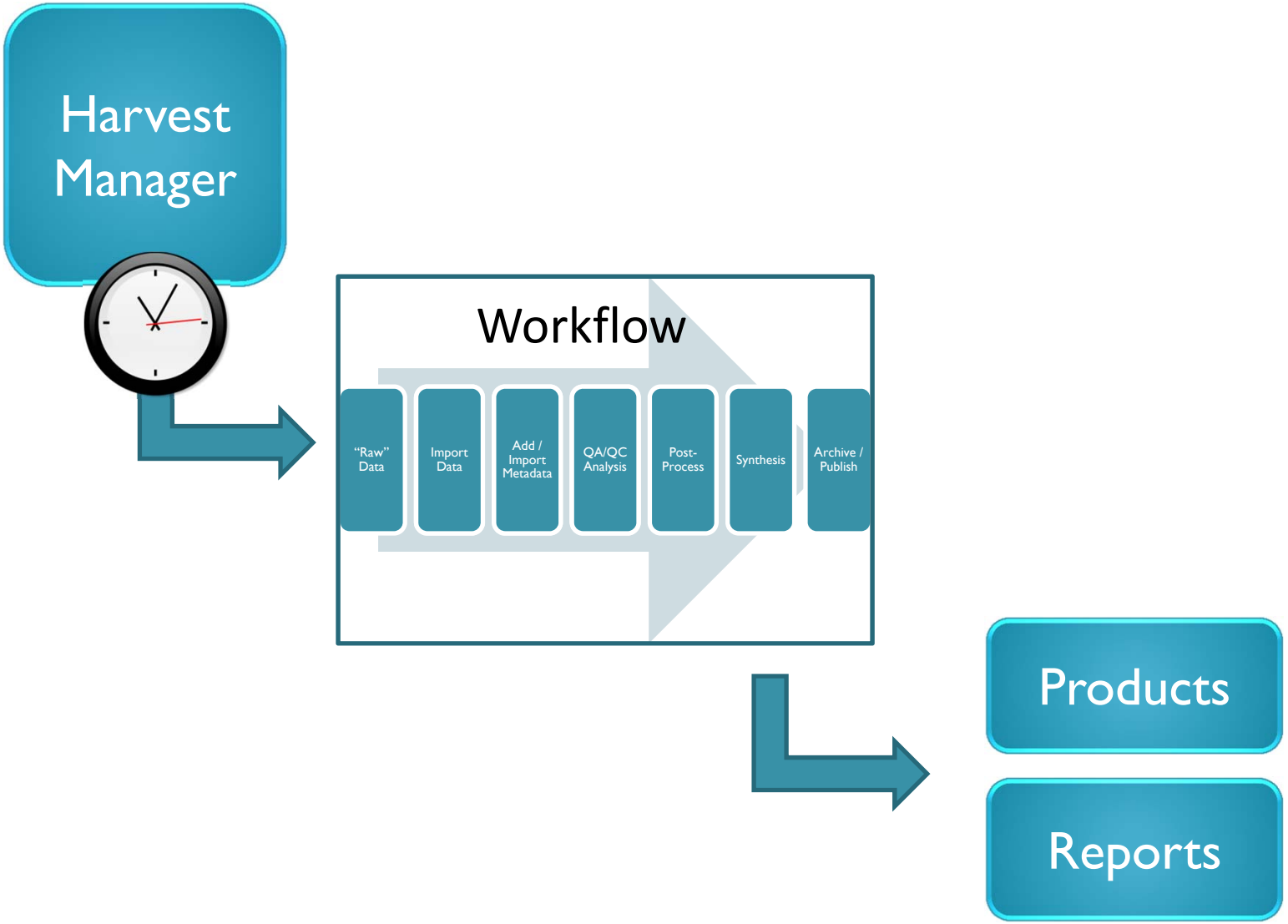


Data Management Cycle



Import Filter + Metadata Template
(One Step)

Data Harvesting Workflows



Key Concepts

- Every operation is 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 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 automatically
  - Each tool logs operations by date/time
  - Data revisions, deletions, flagging captured at user-specified detail
  - Lineage reported in metadata
- Metadata is "live", and updated automatically
  - Data column metadata and data revisions
  - Calculations and unit conversions
  - Code definitions

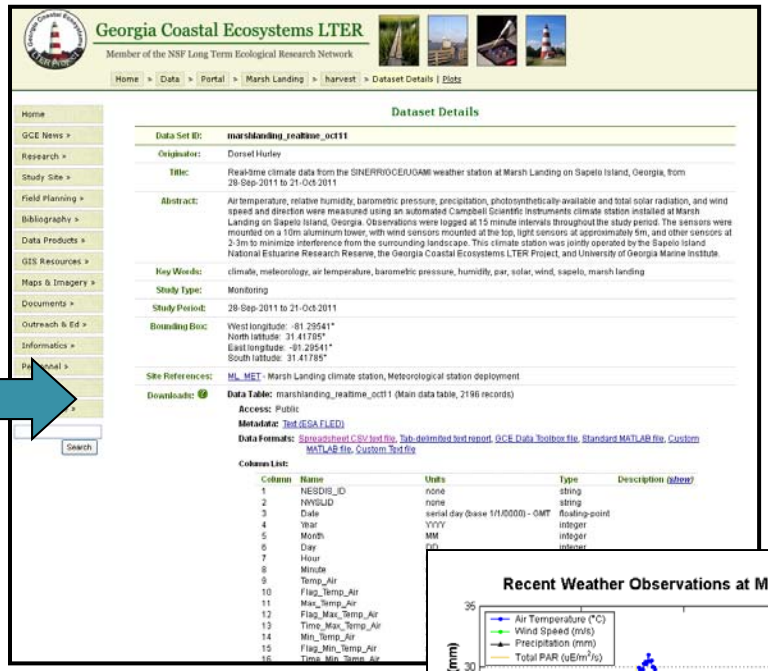
Implementation Scenarios

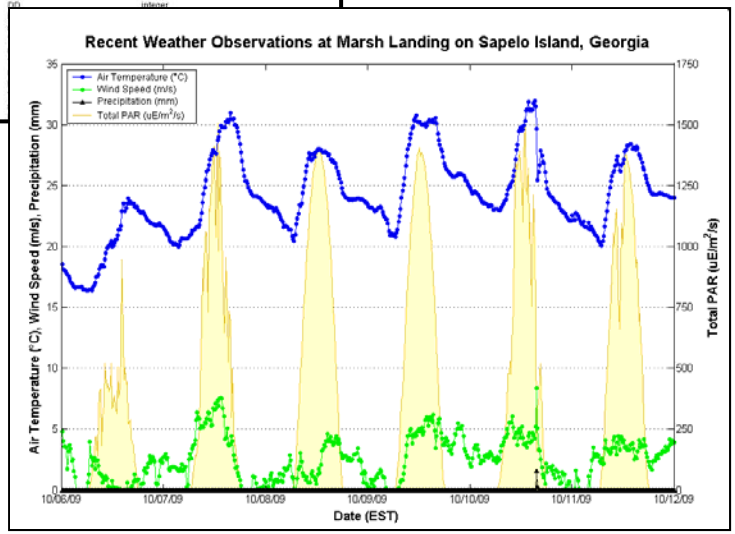
- End-to-End Processing (logger-to-scientist)
  - Acquire raw data from logger, file system, network
  - Assign metadata from template or using forms to validate and flag data
  - Review data and fine-tune flag assignments
  - Generate distribution files & plots, archive data, index for searching
  - Scientists can use toolbox on their desktop to analyze, integrate data
- Data Pre-processing
  - 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. ODM)
- 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)

Suitability for Real-Time Sensor Data

- Good Scalability
  - Data volumes only limited by computer memory (tested >1 GB data sets)
  - Multiple instances can be run on high-end, 64bit, clustered workstations
  - Good flag evaluation performance in use, testing with diverse rule sets
- Good scope for automation
  - Command-line API for unattended batch processing via workflow scripts
  - Timed and triggered workflow implementations easy to deploy
- Support for multiple I/O formats, transport protocols
  - Formats: ASCII, MATLAB, SQL, specialized (CSI, SBE, NWIS RDB, HADS, …)
  - Transport: local file system, UNC paths, HTTP, FTP, SOAP
- Already used at 6 LTER sites (5 sites evaluating)
- Running USGS data harvester for HydroDB since 2003

Real-Time GCE Data Harvesting (>10 years)





Ongoing Development

- Metadata model enhancements
  - Add "schema" support for better EML alignment
  - Add EML unit management
  - Add EML export to support LTER NIS (PASTA)
- Improve harvest management
  - Add GUI tools for configuring harvest info, plot info
  - Add GUI for managing timers
  - Add data harvesting "dashboard" for monitoring activity
- Improve documentation and training materials
- Time frame: July 2013

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
    - 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/)*
  - Tutorials, video training modules in development

Interactive Training

- Introduction to the Dataset Editor application
- Importing and exploring data
  - Generic ASCII
  - Native logger files (Campbell, SeaBird)
  - Data services (Data Turbine, USGS NWIS)
- Metadata management
  - Defining attribute metadata
  - Documentation metadata, templates
- QA/QC framework
  - Defining flags and "rules"
  - Visual QA/QC
  - Managing flagged data
- Creating and exporting products
  - Batch processing raw data
  - Integrating data (join, merge)
  - Summarizing data
  - Exporting data sets and plots