



GCE Data Toolbox for MATLAB

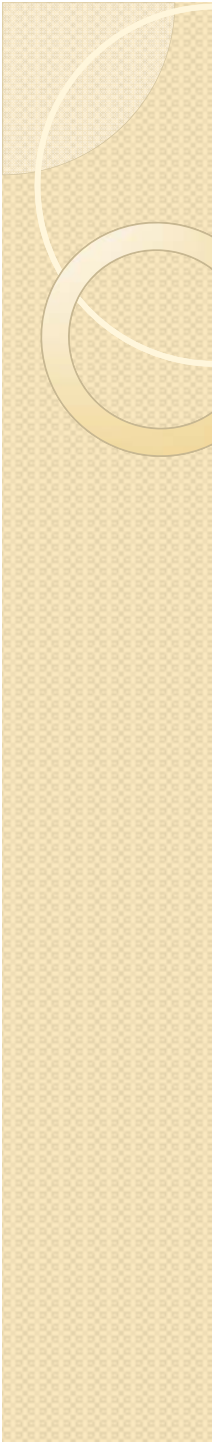
Wade Sheldon

Georgia Coastal Ecosystems LTER
University of Georgia

John Chamblee & Richard Cary

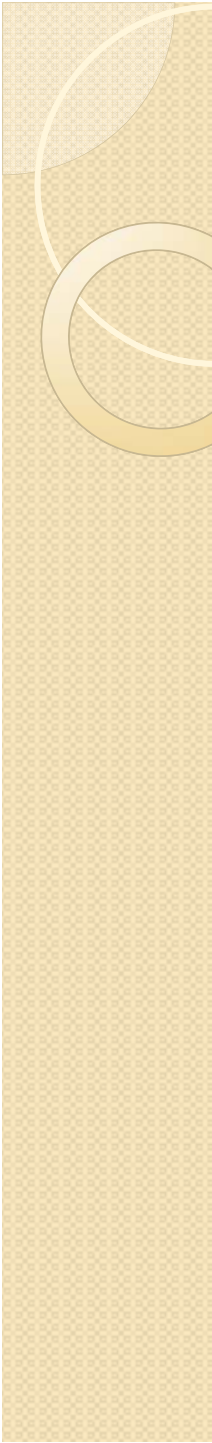
Coweeta LTER
University of Georgia





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):**

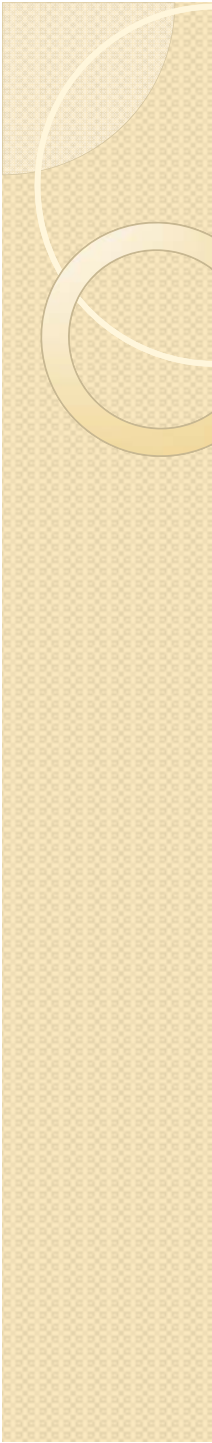
"MATLAB is a programming environment for algorithm development, data analysis, visualization, and numerical computation."

- **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**

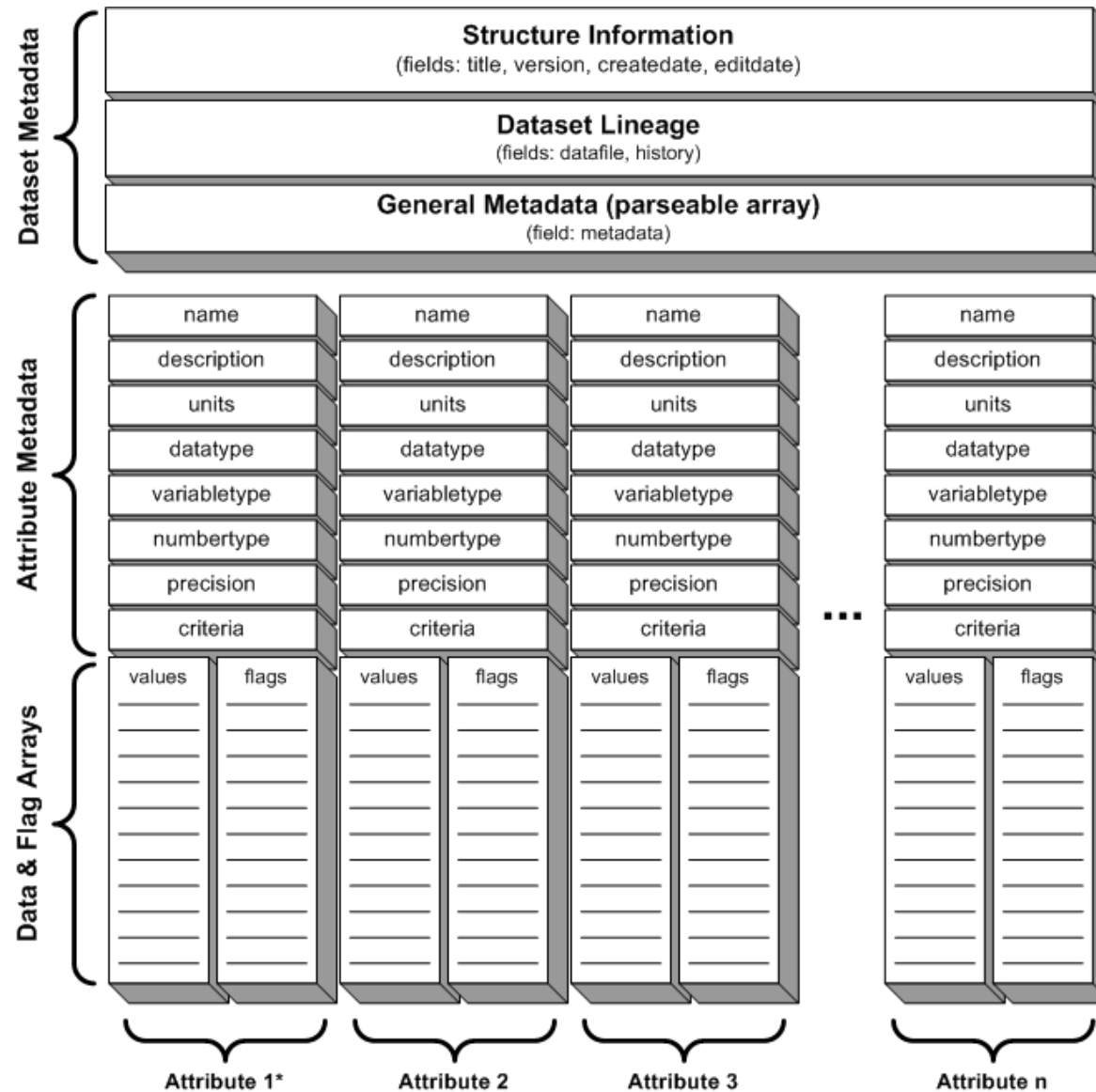
- Commercial ("licensed source") – limits flexibility, costs \$-\$\$\$
- Some programming required for maximum use

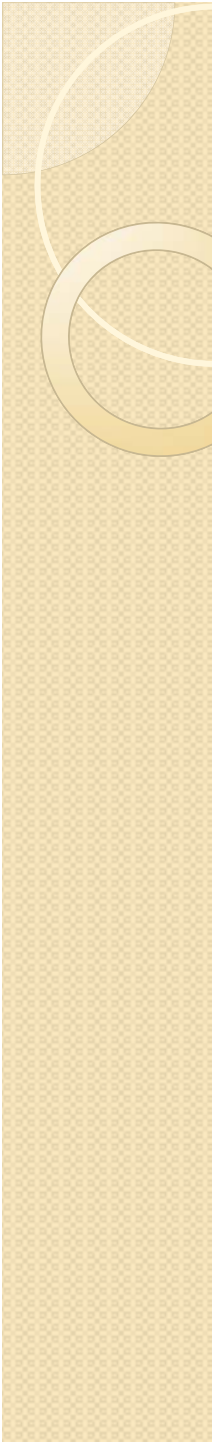


Toolbox Development

- 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

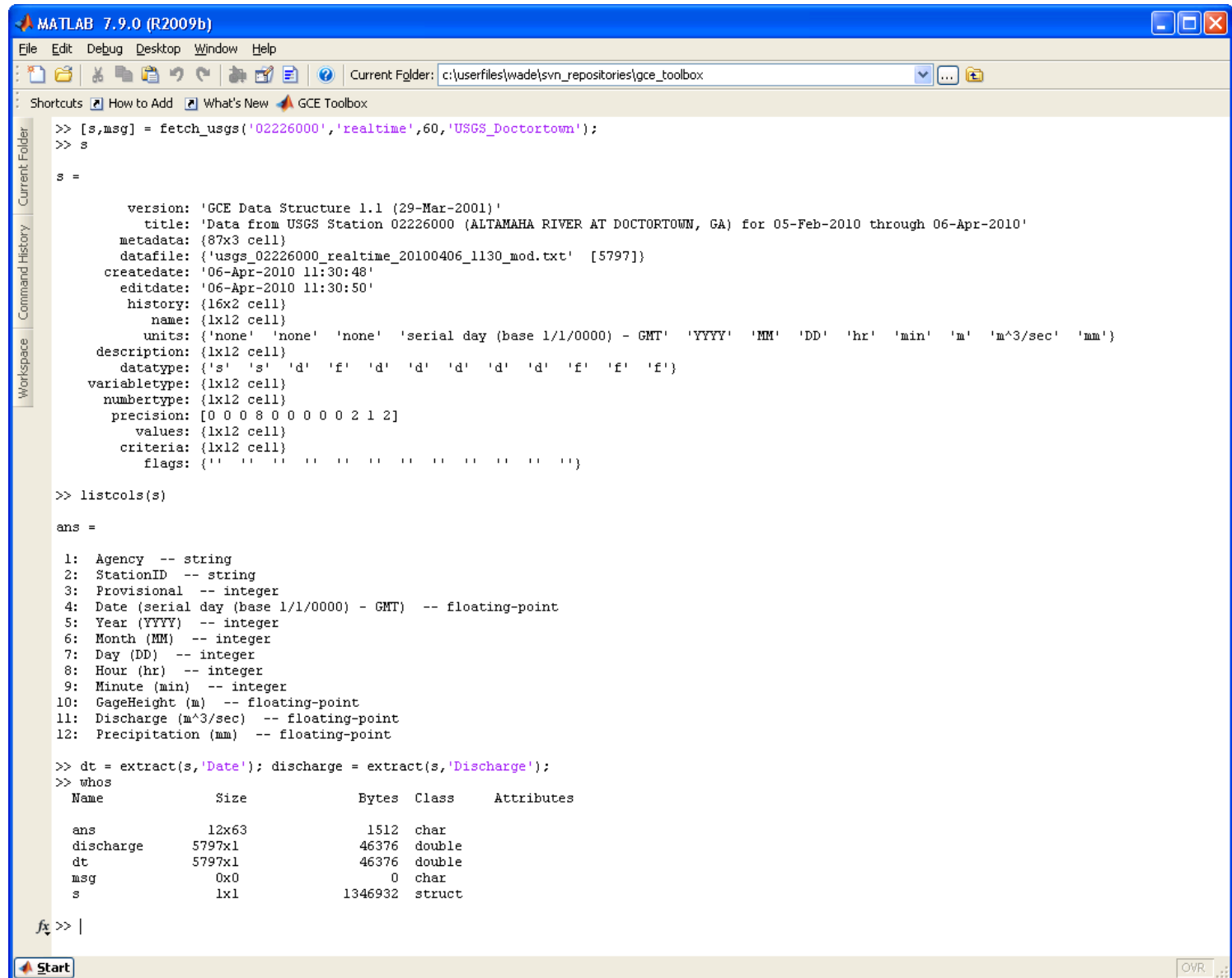
Data Model (GCE Data Structure)





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



GUI Interface (Application)

Continuous salinity, temperature and depth measuremen...

File Edit Metadata Tools Misc Window Help

Column List (select to display properties)

- Site (none)
- Longitude (degrees)
- Latitude (degrees)
- Instrument (none)
- Pump (none)
- Date (serial day (base 1/1.0000) - GMT)
- Year (YYYY)
- Month (MM)
- Day (DD)
- Hour (hr)
- Minute (min)
- Temperature (°C)
- Conductivity (S/m)
- Pressure (dbar)
- Depth (m)
- Salinity (PSU)

Move First Move Up Move Down Move Last Preview Histogram Manual QC Add Delete Restore

Column Name Site

Column Units none **Convert**

Description

Nearest nominal GCE-LTER sampling site

Data Type integer (d)

Variable Type categorical values (nominal)

Numerical Type discrete/interval (discrete)

Precision 0 decimal places

Flag Criteria **Edit**

About the GCE Data Toolbox

GCE Data Toolbox for MATLAB

Software for metadata-based processing, analysis, visualization and transformation of environmental data

Toolbox Version 3.4.0 (27-Nov-2012)

(c)2002-2012 Wade M. Sheldon and the Georgia Coastal Ecosystems LTER Project

Wade M. Sheldon
Dept. of Marine Sciences
University of Georgia
Athens, GA 30602-3636
email: sheldon@uga.edu

Search Engine Dataset Editor Mapping Tools Documentation Exit MATLAB

Software Usage Agreement and Disclaimer --

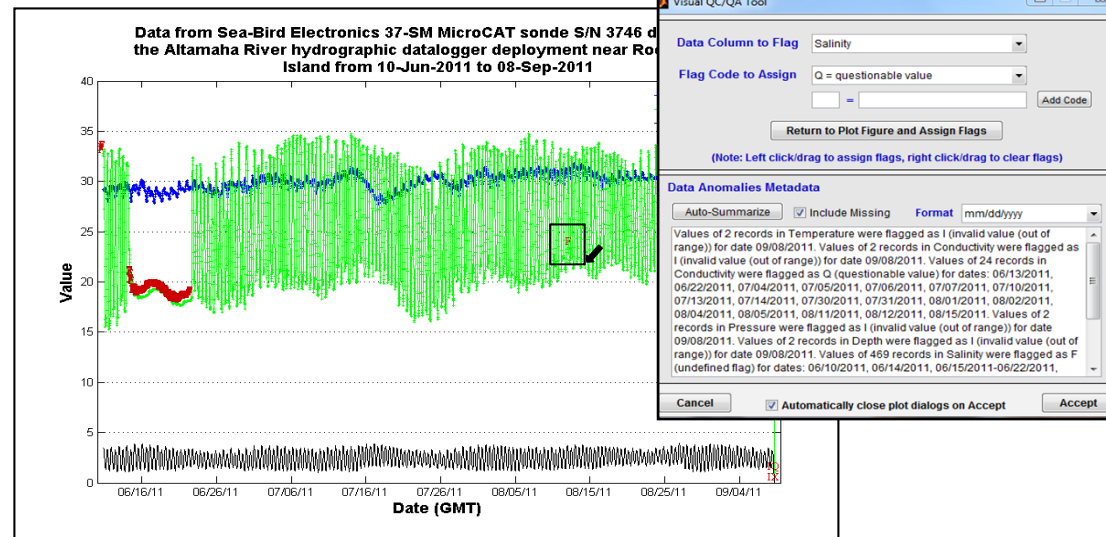
The GCE Data Toolbox is provided as a courtesy to the scientific community to support offline analysis and customization of data sets distributed by the Georgia Coastal Ecosystems Long Term Ecological Research project (<http://gce-lter.marsci.uga.edu/>). The latest versions of the software and documentation are available on the World Wide Web at: http://gce-lter.marsci.uga.edu/public/m/tools/data_toolbox.htm.

The GCE Data Toolbox is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.

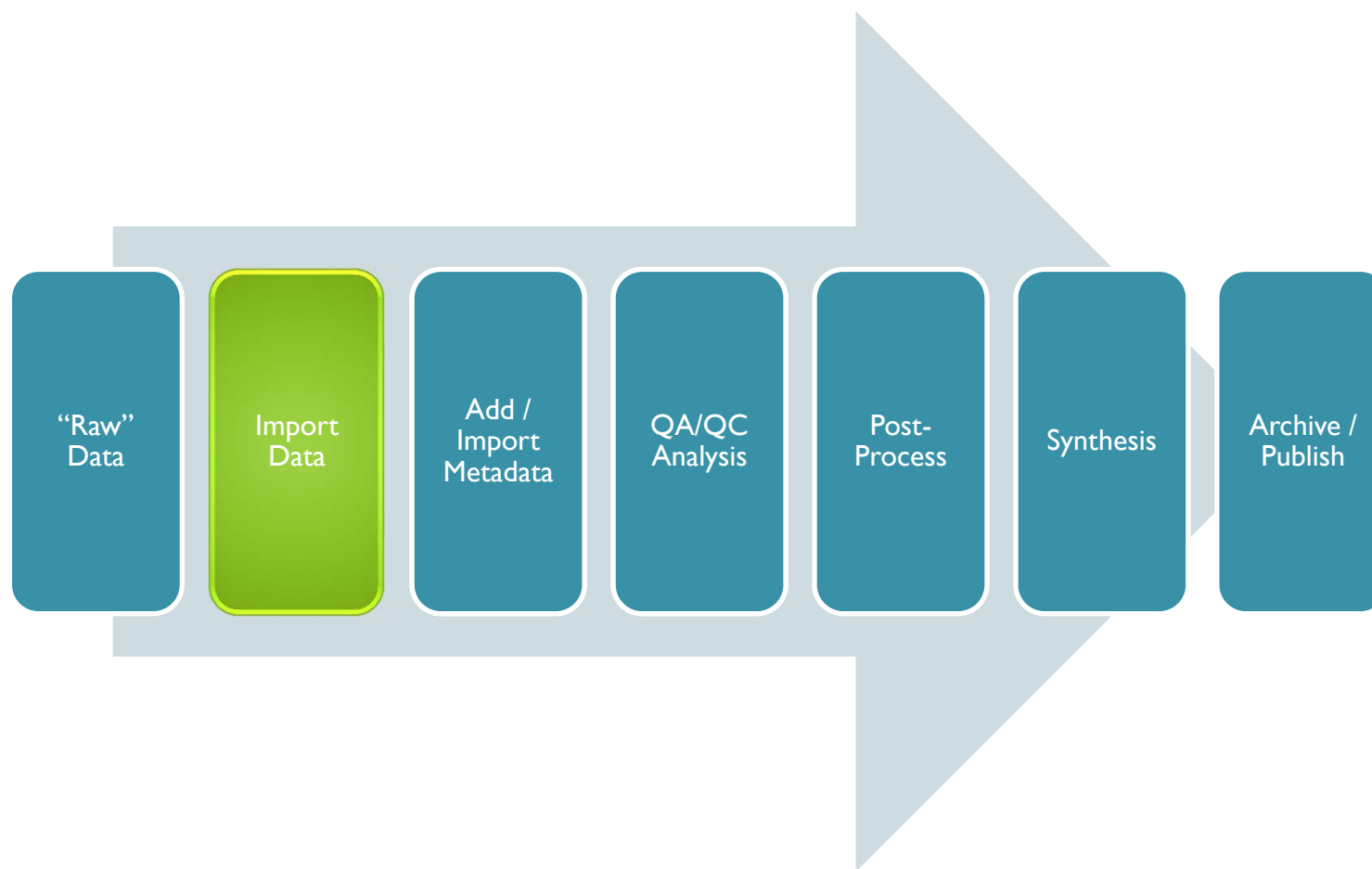
☒ Automatically load startup screen if all other windows are closed

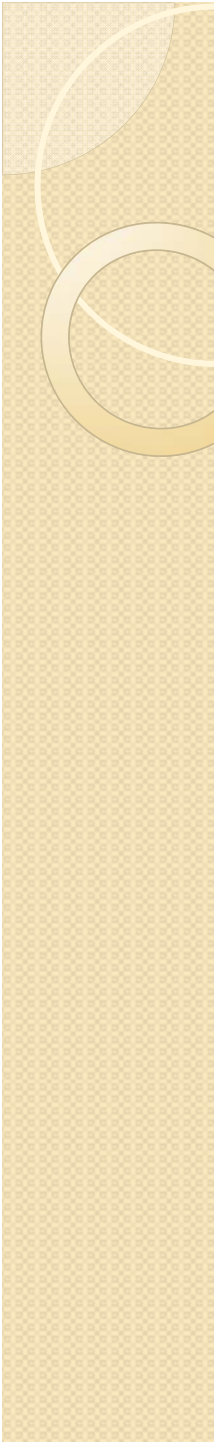
Data Editor

All	Site	Longitude	Latitude	Instrument	Pump	Date	Year	Month	Day
Base	(none)	(degrees)	(degrees)	(none)	(none)	(serial day (base 1/1.0000) - GMT)	(YYYY)	(MM)	(DD)
1	7	-81.475500	31.330383	2398	0	733043.000000	2007	1	1
2	7	-81.475500	31.330383	2398	0	733043.020033	2007	1	1
3	7	-81.475500	31.330383	2398	0	733043.041667	2007	1	1
4	7	-81.475500	31.330383	2398	0	733043.062500	2007	1	1
5	7	-81.475500	31.330383	2398	0	733043.083333	2007	1	1
6	7	-81.475500	31.330383	2398	0	733043.104167	2007	1	1
7	7	-81.475500	31.330383	2398	0	733043.125000	2007	1	1
8	7	-81.475500	31.330383	2398	0	733043.145833	2007	1	1
9	7	-81.475500	31.330383	2398	0	733043.166667	2007	1	1
10	7	-81.475500	31.330383	2398	0	733043.187500	2007	1	1
11	7	-81.475500	31.330383	2398	0	733043.208333	2007	1	1
12	7	-81.475500	31.330383	2398	0	733043.229167	2007	1	1
13	7	-81.475500	31.330383	2398	0	733043.250000	2007	1	1
14	7	-81.475500	31.330383	2398	0	733043.270833	2007	1	1
15	7	-81.475500	31.330383	2398	0	733043.291667	2007	1	1
16	7	-81.475500	31.330383	2398	0	733043.312500	2007	1	1
17	7	-81.475500	31.330383	2398	0	733043.333333	2007	1	1
18	7	-81.475500	31.330383	2398	0	733043.354167	2007	1	1
19	7	-81.475500	31.330383	2398	0	733043.375000	2007	1	1
20	7	-81.475500	31.330383	2398	0	733043.395833	2007	1	1
21	7	-81.475500	31.330383	2398	0	733043.416667	2007	1	1
22	7	-81.475500	31.330383	2398	0	733043.437500	2007	1	1
23	7	-81.475500	31.330383	2398	0	733043.458333	2007	1	1
24	7	-81.475500	31.330383	2398	0	733043.479167	2007	1	1
25	7	-81.475500	31.330383	2398	0	733043.500000	2007	1	1



Data Management Cycle

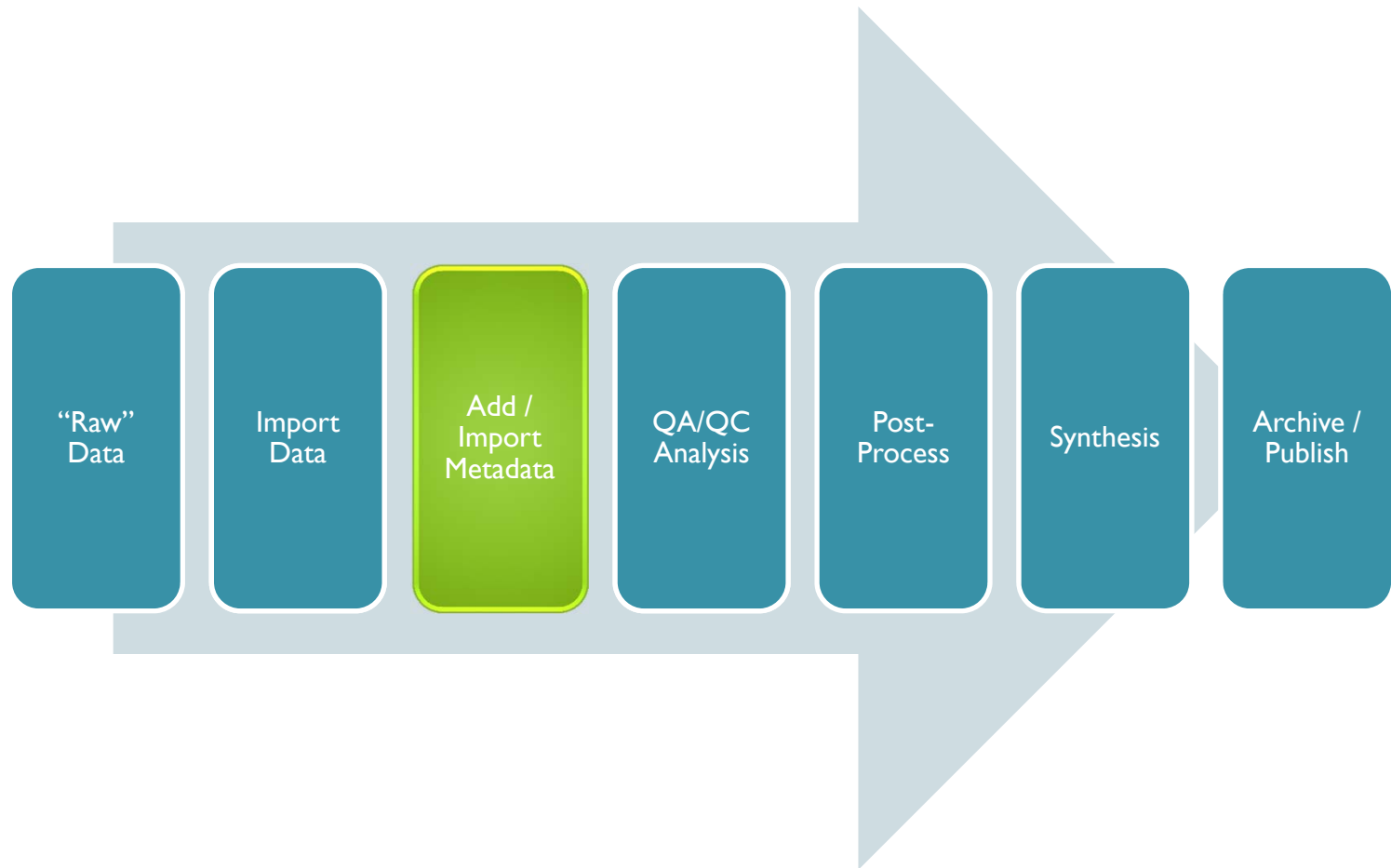


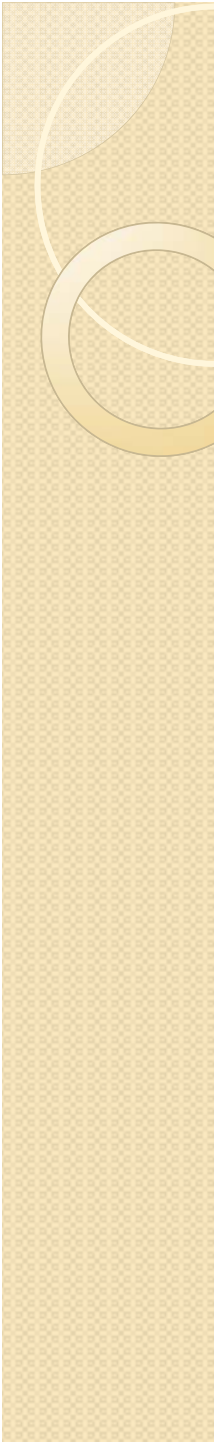


Importing Data

- 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

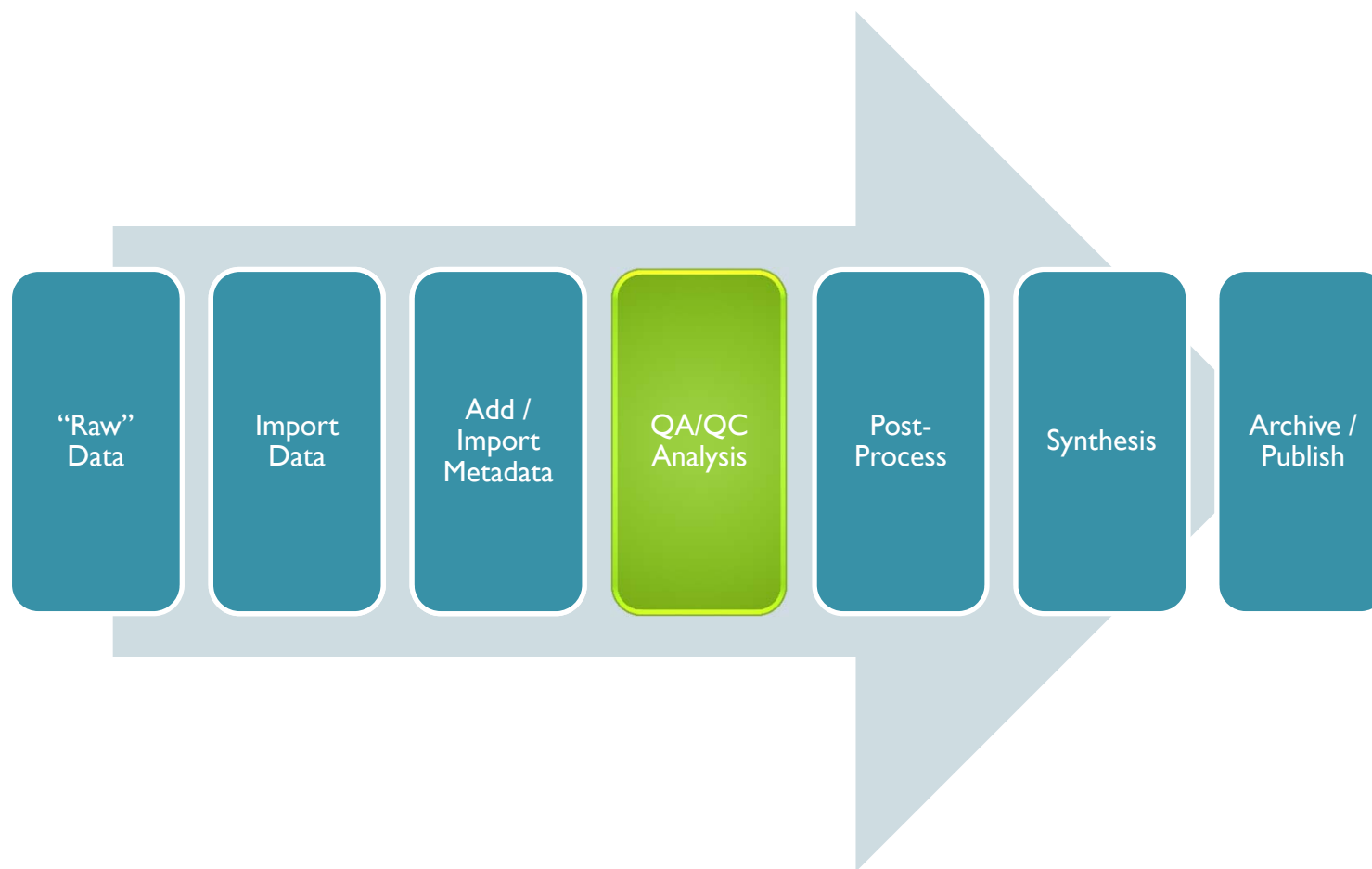


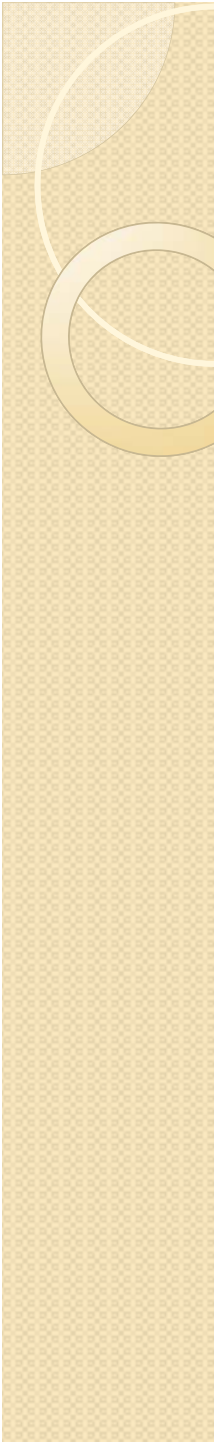


Adding/Importing Metadata

- 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



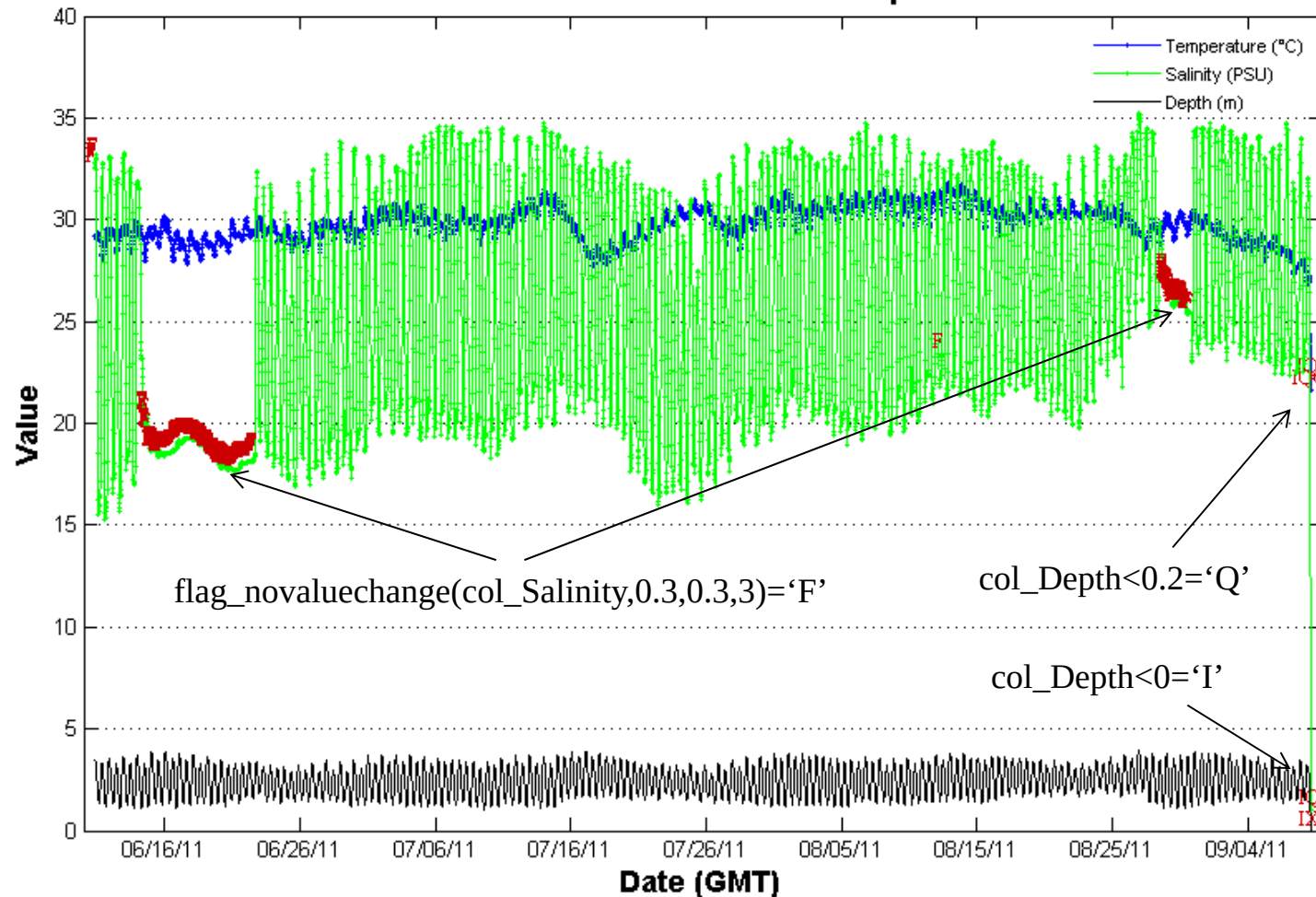


QA/QC Analysis Framework

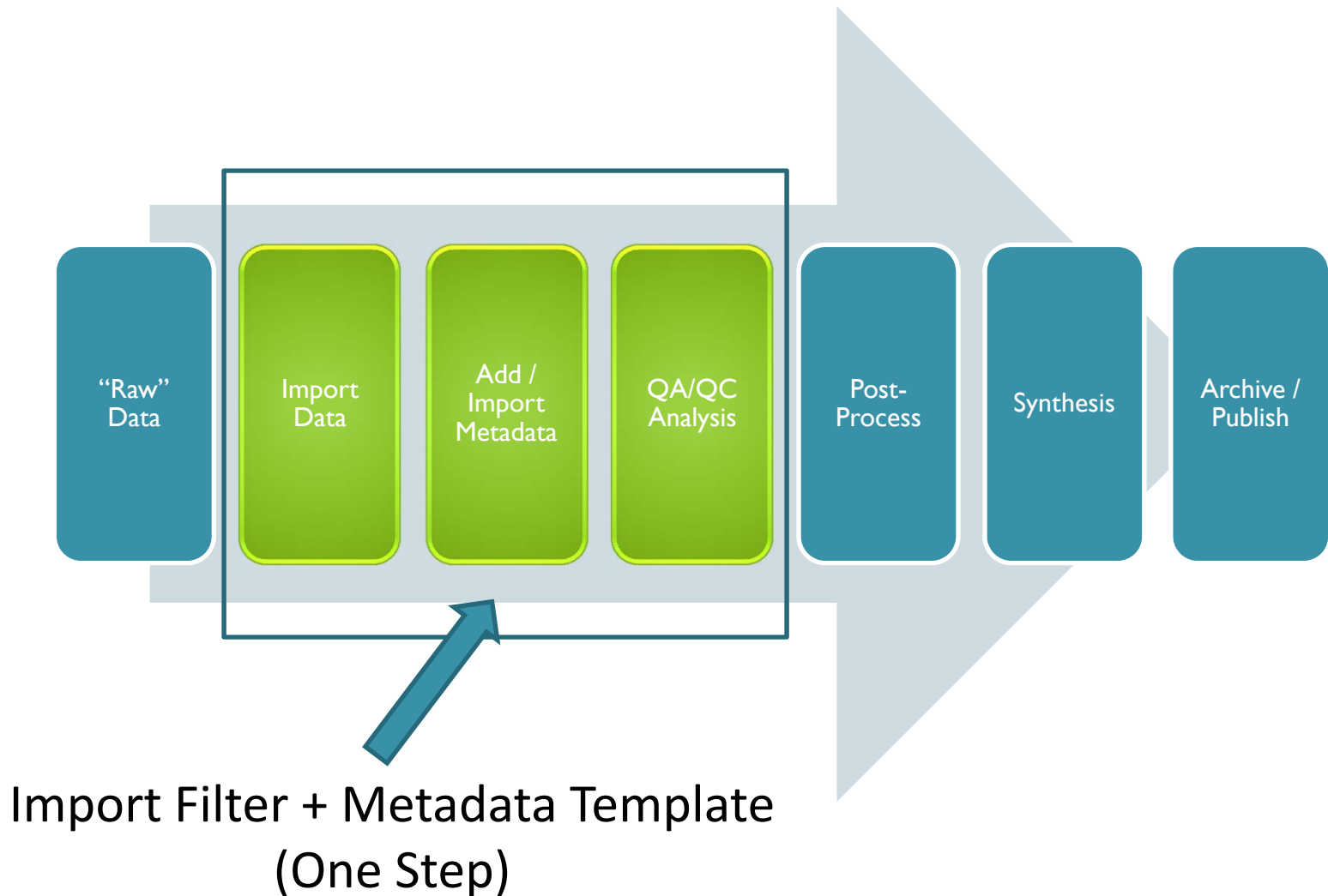
- Programmatic Q/C Analysis
 - “Rules” that define conditions in which values should be flagged
 - Unlimited Q/C rules for each variable
 - Rules evaluated when data loaded and when data or rules change
 - Rules can be predefined in metadata templates to automate Q/C on import
- Interactive Q/C Analysis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

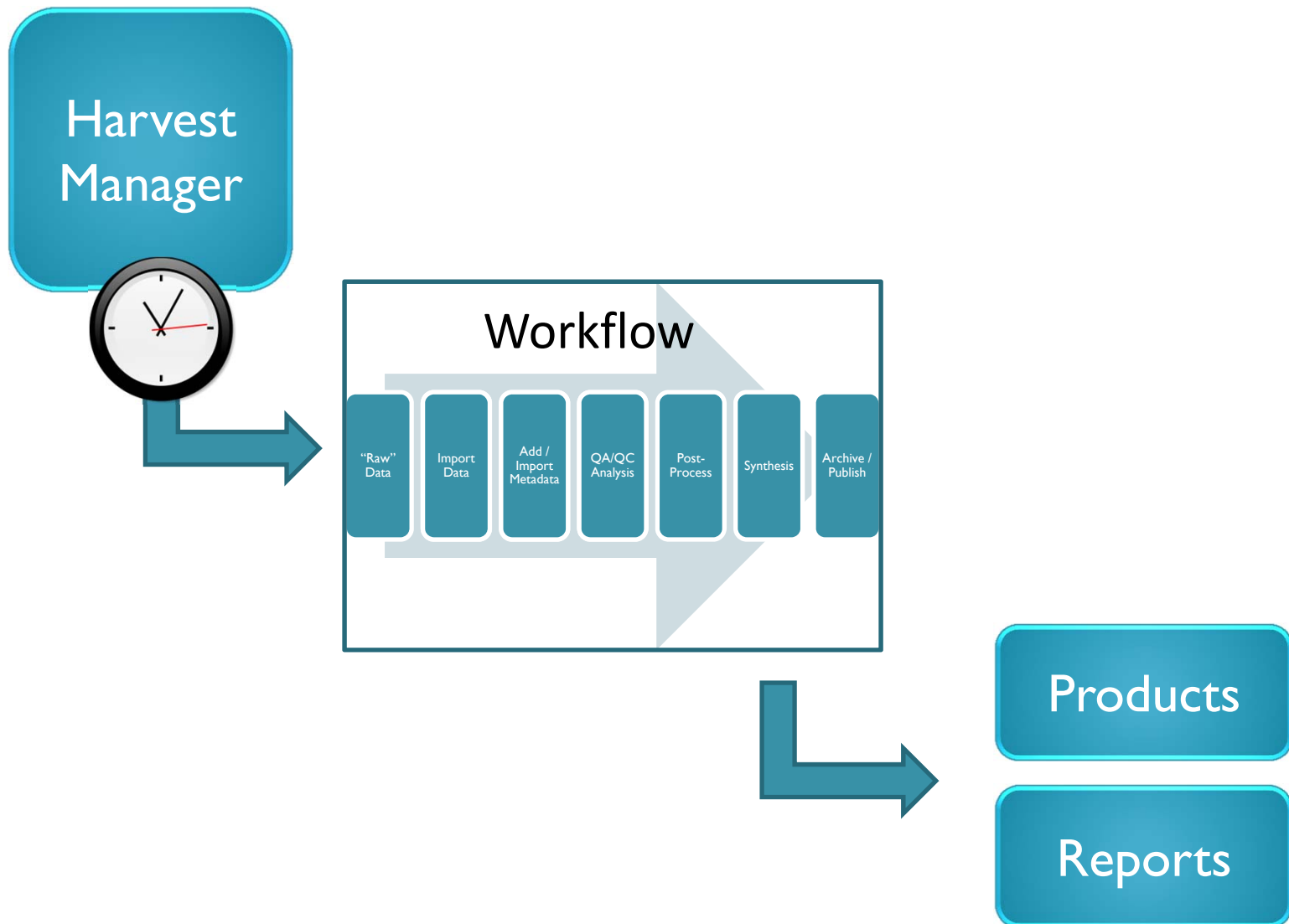
Data from Sea-Bird Electronics 37-SM MicroCAT sonde S/N 3746 deployed at the Altamaha River hydrographic datalogger deployment near Rockdedundy Island from 10-Jun-2011 to 08-Sep-2011

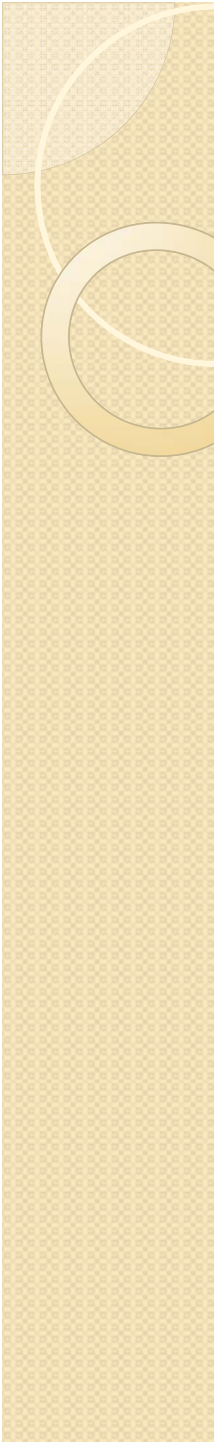


Data Management Cycle



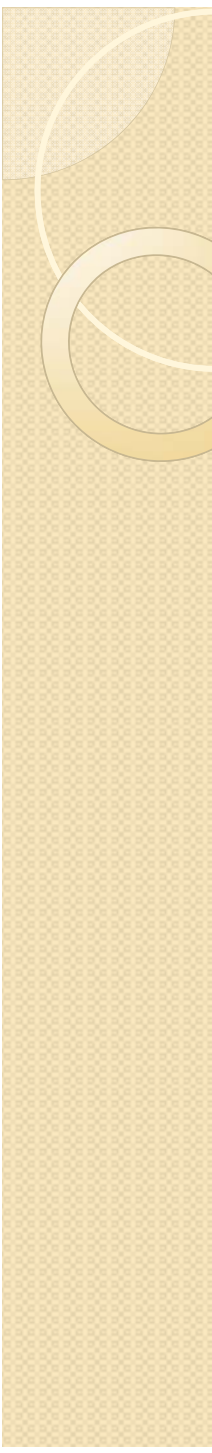
Data Harvesting Workflows





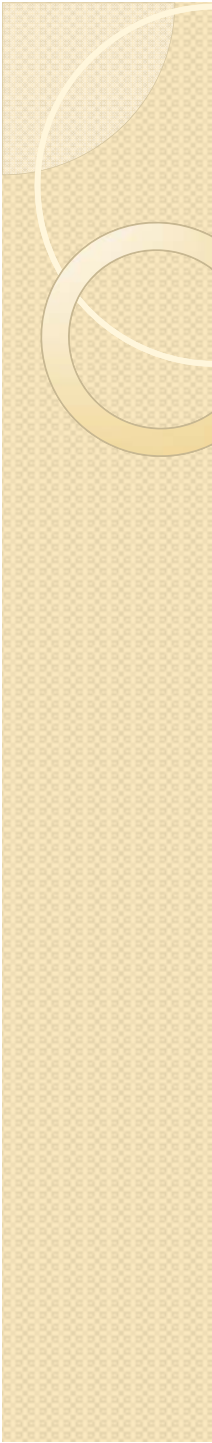
Key Concepts

- 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)



Georgia Coastal Ecosystems LTER
Member of the NSF Long Term Ecological Research Network

Home > Data > Portal > Marsh Landing > harvest > Dataset Details | Plots

Dataset Details

Data Set ID: marshlanding_realtime_oct11

Originator: Dorset Hurley

Title: Real-time climate data from the SNERVOCE/UOAM weather station at Marsh Landing on Sapelo Island, Georgia, from 28 Sep-2011 to 21-Oct-2011

Abstract: Air temperature, relative humidity, barometric pressure, precipitation, photosynthetically available and total solar radiation, and wind speed and direction were measured using an automated Campbell Scientific Instruments climate station installed at Marsh Landing on Sapelo Island, Georgia. Observations were logged at 15 minute intervals throughout the study period. The sensors were mounted on a 10m aluminum tower, with wind sensors mounted at the top, light sensors at approximately 5m, and other sensors at 2.3m to minimize interference from the surrounding landscape. This climate station was jointly operated by the Sapelo Island National Estuarine Research Reserve, the Georgia Coastal Ecosystems LTER Project, and University of Georgia Marine Institute.

Key Words: climate, meteorology, air temperature, barometric pressure, humidity, par, solar, wind, sapelo, marsh landing

Study Type: Monitoring

Study Period: 28-Sep-2011 to 21-Oct-2011

Bounding Box: West longitude: -81.29541°
North latitude: 31.41795°
East longitude: -81.29541°
South latitude: 31.41795°

Site References: [ML_MET](#) - Marsh Landing climate station, Meteorological station deployment

Downloads: [marshlanding_realtime_oct11](#) (Main data table, 2196 records)

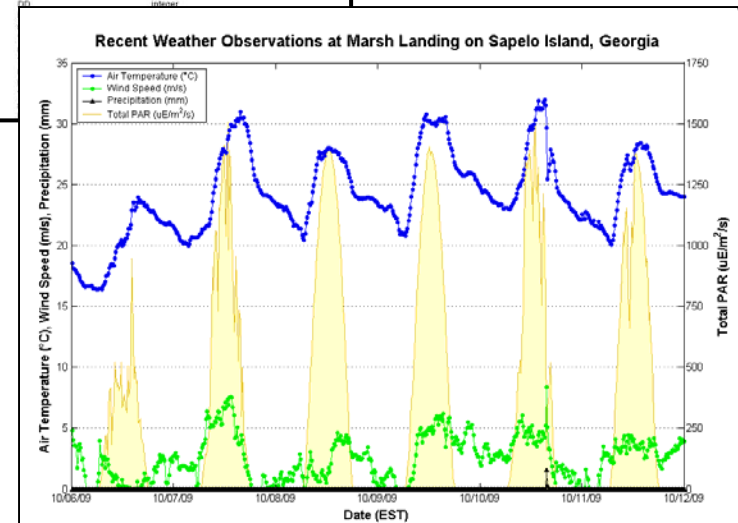
Access: Public

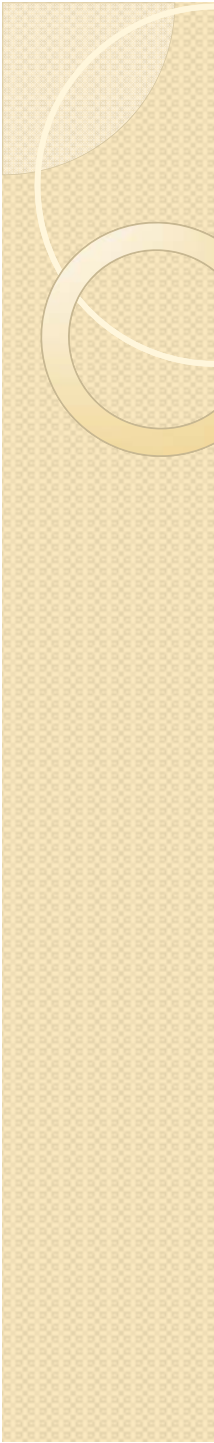
Metadata: [Text \(CSV FILED\)](#)

Data Formats: [Spreadsheet \(CSV text file\)](#), [Tab-delimited text report](#), [GCE Data Toolbox file](#), [Standard MATLAB file](#), [Custom MATLAB file](#), [Custom Text file](#)

Column List:

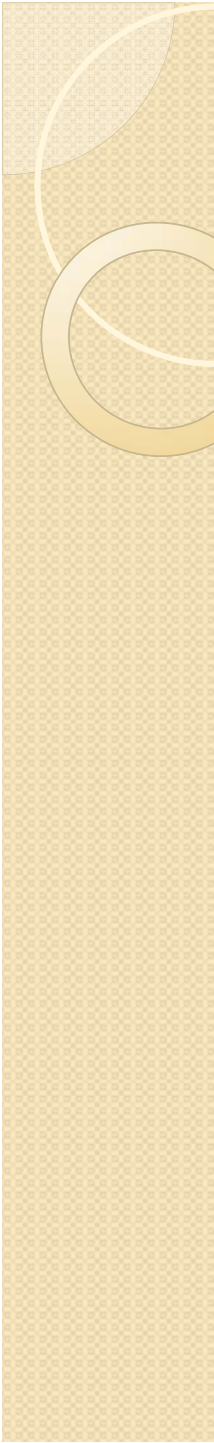
Column	Name	Units	Type	Description (abbrev)
1	NESDRID_ID	none	string	
2	NWGLID	none	string	
3	Date	serial day (base 1/1/2000) - GMT	floating-point	
4	Year	YYYY	integer	
5	Month	MM	integer	
6	Day	DD	integer	
7	Hour			
8	Minute			
9	Temp_Air			
10	Flag_Temp_Air			
11	Max_Temp_Air			
12	Flag_Max_Temp_Air			
13	Time_Max_Temp_Air			
14	Min_Temp_Air			
15	Flag_Min_Temp_Air			
16	Time_Min_Temp_Air			





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

