Overview of the GCE Data Toolbox for MATLAB

Wade Sheldon

Georgia Coastal Ecosystems LTER

University of Georgia

Background & Motivation

- Georgia Coastal Ecosystems LTER project started in Sept 2000
  - Large data collection effort (cruises, moorings, met stations, water quality, field surveys, ...)
  - NSF & LTER require data archiving and sharing
  - LTER requires detailed “metadata” for every data set
  - Needed to standardize data processing, quality control, documentation
- No ready-to-use software for LTER data management
  - Lab management software (LIMS) useless for field data, expensive
  - Most LTER sites were using “flat files” – limiting
  - A few sites using relational databases, client/server apps – proprietary, complex, unfamiliar, require constant network access
- Chose to develop custom data management software (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

Georgia Coastal Ecosystems LTER Project

What is MATLAB?

- From Mathworks: (http://www.mathworks.com/products/matlab/)

  "MATLAB is a programming environment for algorithm development, data analysis, visualization, and numerical computation. Using MATLAB, you can solve technical computing problems faster than with traditional programming languages, such as C, C++, and Fortran."

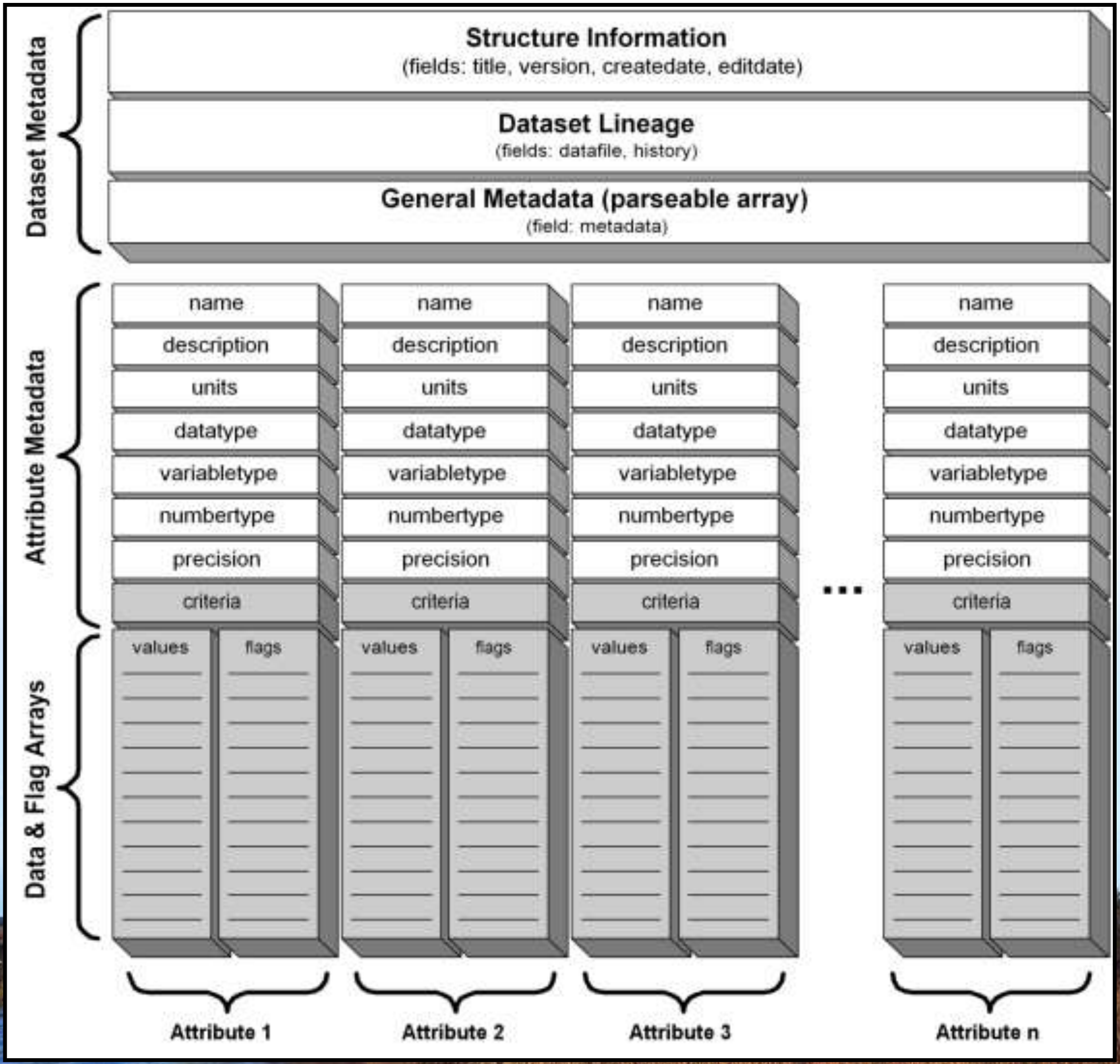
- Benefits:
  - ➢ Ubiquitous in engineering and many science branches
    - Geophysics, Oceanography, Chemistry, ecological modeling, sensor networking
  - ➢ Rapid development with lots of pre-built functionality, Java integration
  - ➢ Cross-platform code, GUIs and data formats (Windows, *nix, Mac OS/x)
  - ➢ Stable: good support and backward compatibility (28 year history)
  - ➢ Scalable (netbook to cluster) - great performance with huge data sets
  - ➢ Broad I/O support (serial ports to web services)

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

Toolbox Development

- Started by reviewing ESA's "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.
- Identified information storage requirements
  - Any number of numeric (integer, float, exponential) and text variables
  - Structured attribute metadata for each variable (name, units, desc., type, precision, ...)
  - Structured documentation (dataset 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

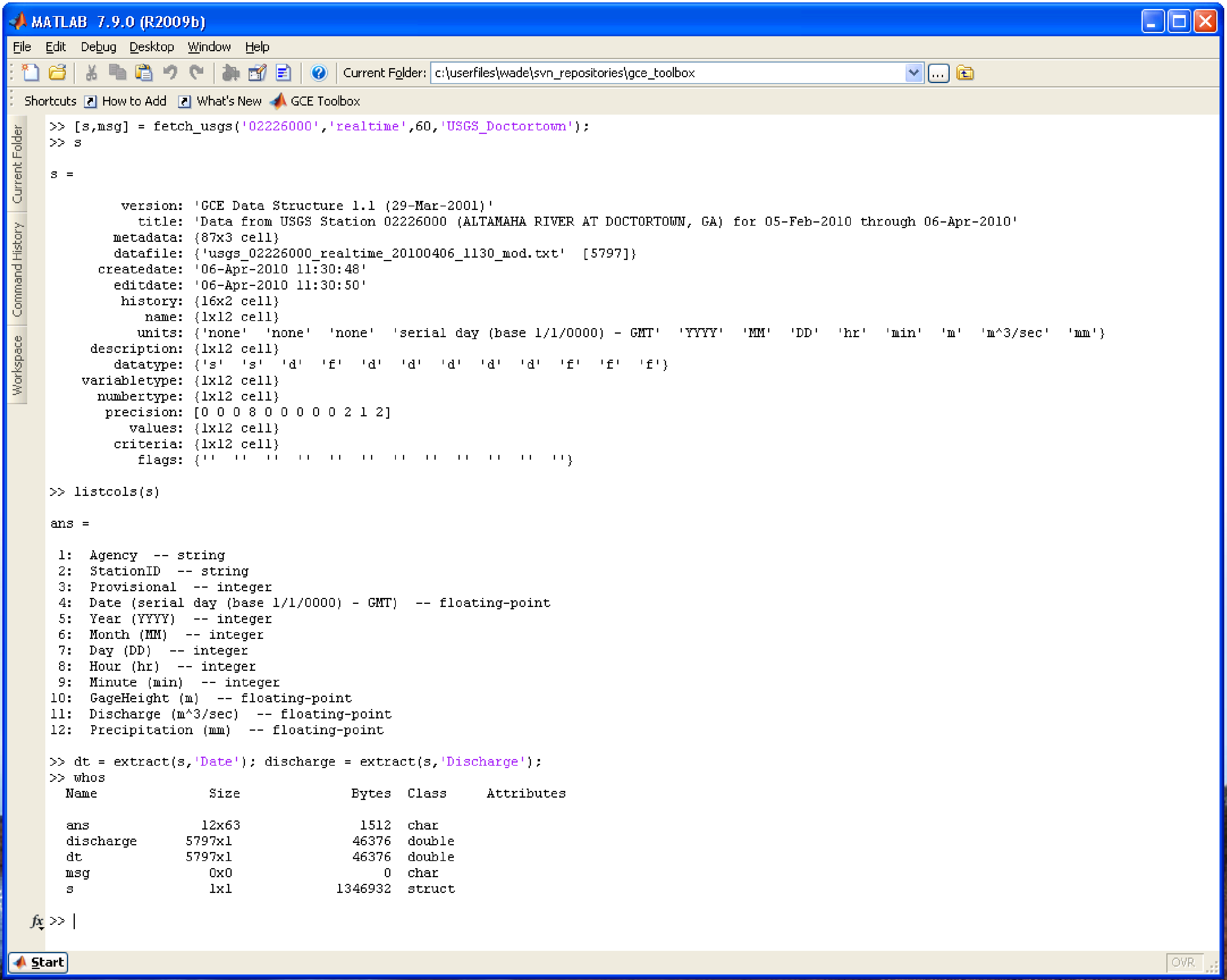
Data Model (GCE Data Structure)



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/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)

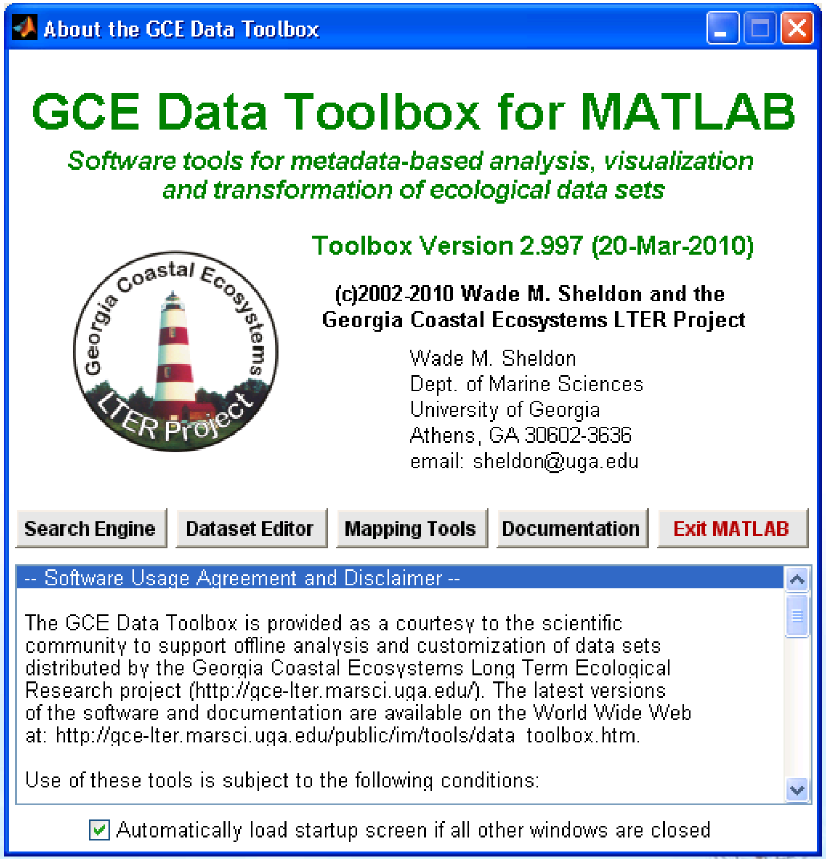
Command Line



Current Folder: c:\userfiles\wade\svn_repositories\gce_toolbox

```
>> [s,msg] = fetch_usgs('02226000','realtime',60,'USGS_Doctortown');
>> s

s =

          version: 'GCE Data Structure 1.1 (29-Mar-2001)'
            title: 'Data from USGS Station 02226000 (ALTAMAHA RIVER AT DOCTORTOWN, GA) for 05-Feb-2010 through 06-Apr-2010'
         metadata: {87x3 cell}
         datafile: {'usgs_02226000_realtime_20100406_1130_mod.txt'  [5797]}
       createdate: '06-Apr-2010 11:30:48'
         editdate: '06-Apr-2010 11:30:50'
          history: {16x2 cell}
             name: {1x12 cell}
            units: {'none'  'none'  'none'  'serial day (base 1/1/0000) - GMT'  'YYYY'  'MM'  'DD'  'hr'  'min'  'm'  'm^3/sec'  'mm'}
      description: {1x12 cell}
         datatype: {'s'  's'  'd'  'f'  'd'  'd'  'd'  'd'  'd'  'f'  'f'  'f'}
     variabletype: {1x12 cell}
       numbertype: {1x12 cell}
        precision: [0 0 0 8 0 0 0 0 0 2 1 2]
           values: {1x12 cell}
         criteria: {1x12 cell}
            flags: {''  ''  ''  ''  ''  ''  ''  ''  ''  ''  ''  ''}

>> listcols(s)

ans =

 1:  Agency  -- string
 2:  StationID  -- string
 3:  Provisional  -- integer
 4:  Date (serial day (base 1/1/0000) - GMT)  -- floating-point
 5:  Year (YYYY)  -- integer
 6:  Month (MM)  -- integer
 7:  Day (DD)  -- integer
 8:  Hour (hr)  -- integer
 9:  Minute (min)  -- integer
10:  GageHeight (m)  -- floating-point
11:  Discharge (m^3/sec)  -- floating-point
12:  Precipitation (mm)  -- floating-point

>> dt = extract(s,'Date'); discharge = extract(s,'Discharge');
>> whos
  Name              Size                 Bytes  Class     Attributes

  ans              12x63                  1512  char
  discharge      5797x1                  46376  double
  dt             5797x1                  46376  double
  msg               0x0                      0  char
  s                 1x1                1346932  struct

fx >>
```

Startup Dialog



GCE Data Toolbox for MATLAB

Software tools for metadata-based analysis, visualization and transformation of ecological data sets

Toolbox Version 2.997 (20-Mar-2010)

(c)2002-2010 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

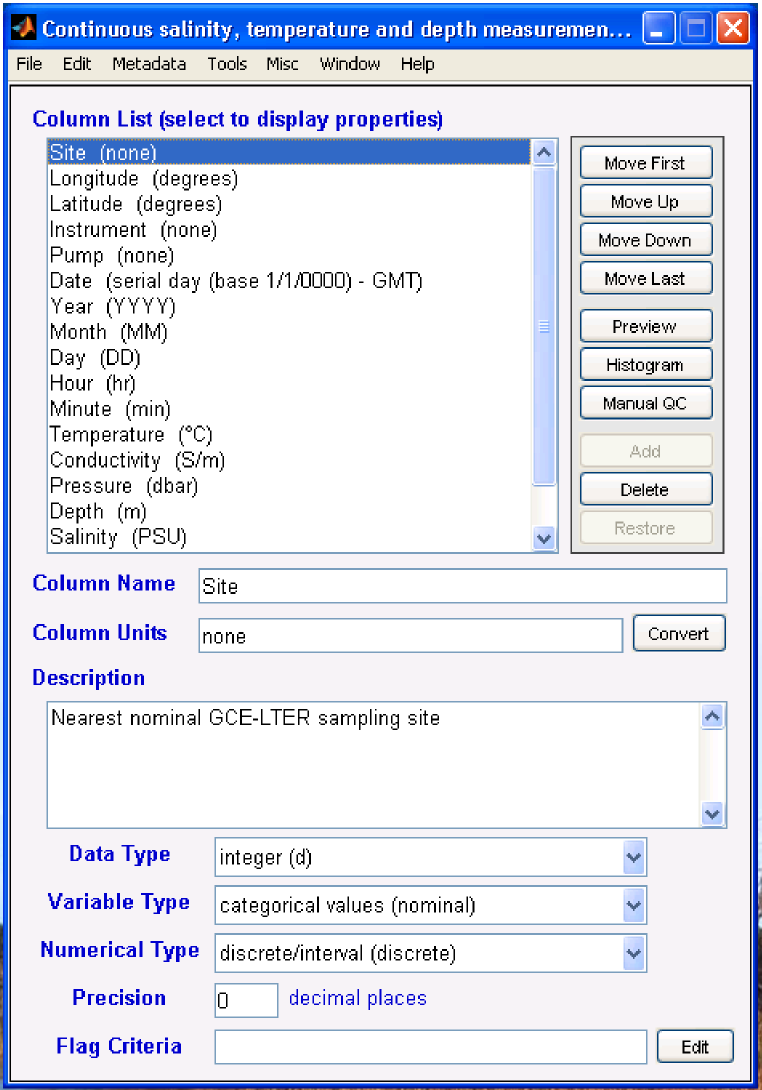
-- Software Usage Agreement and Disclaimer --

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/im/tools/data_toolbox.htm.

Use of these tools is subject to the following conditions:

☑ Automatically load startup screen if all other windows are closed

Dataset Editor



Continuous salinity, temperature and depth measuremen...

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)

Column Name Site

Column Units none

Description

Nearest nominal GCE-LTER sampling site

Data Type integer (d)

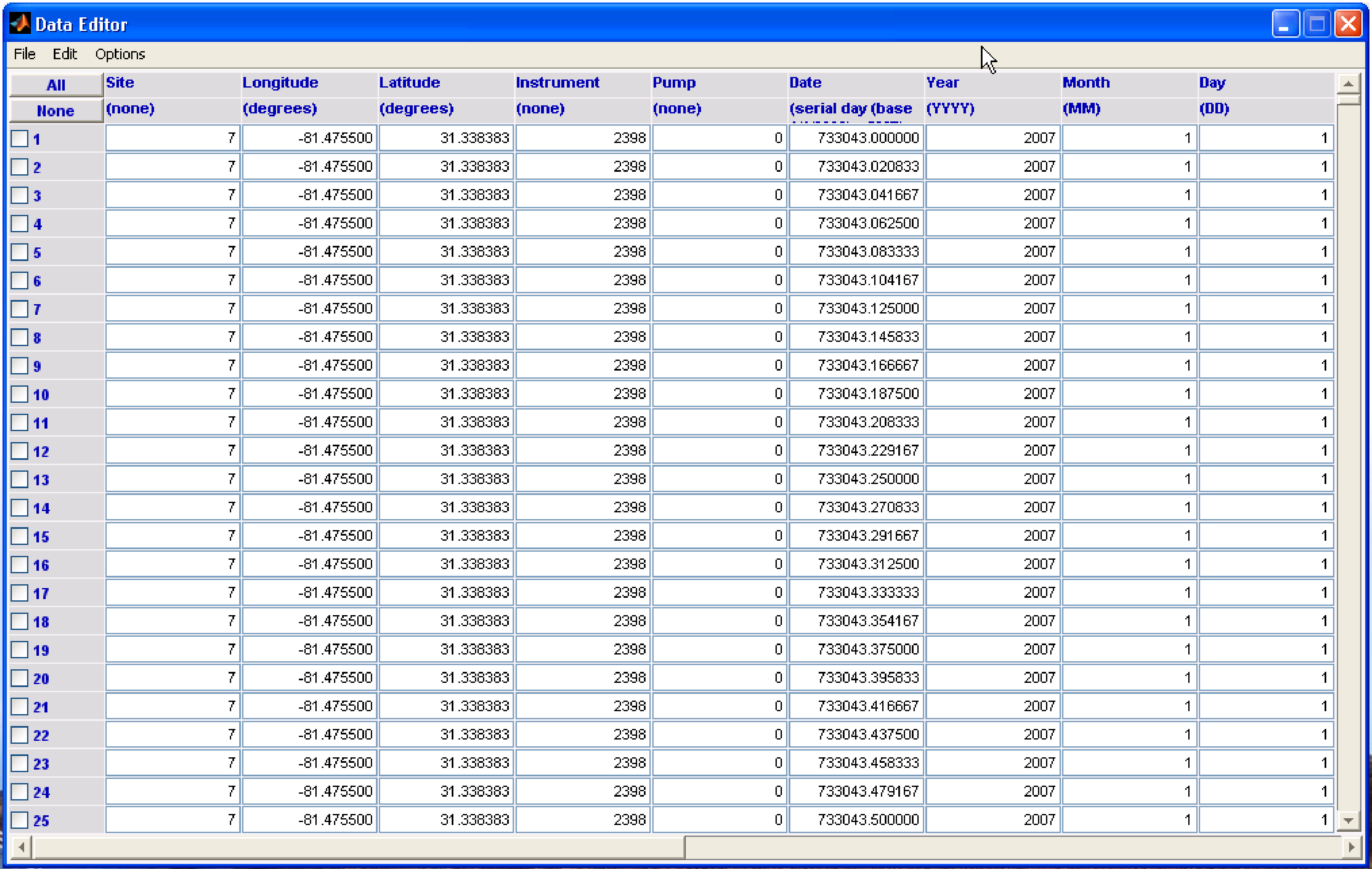
Variable Type categorical values (nominal)

Numerical Type discrete/interval (discrete)

Precision 0 decimal places

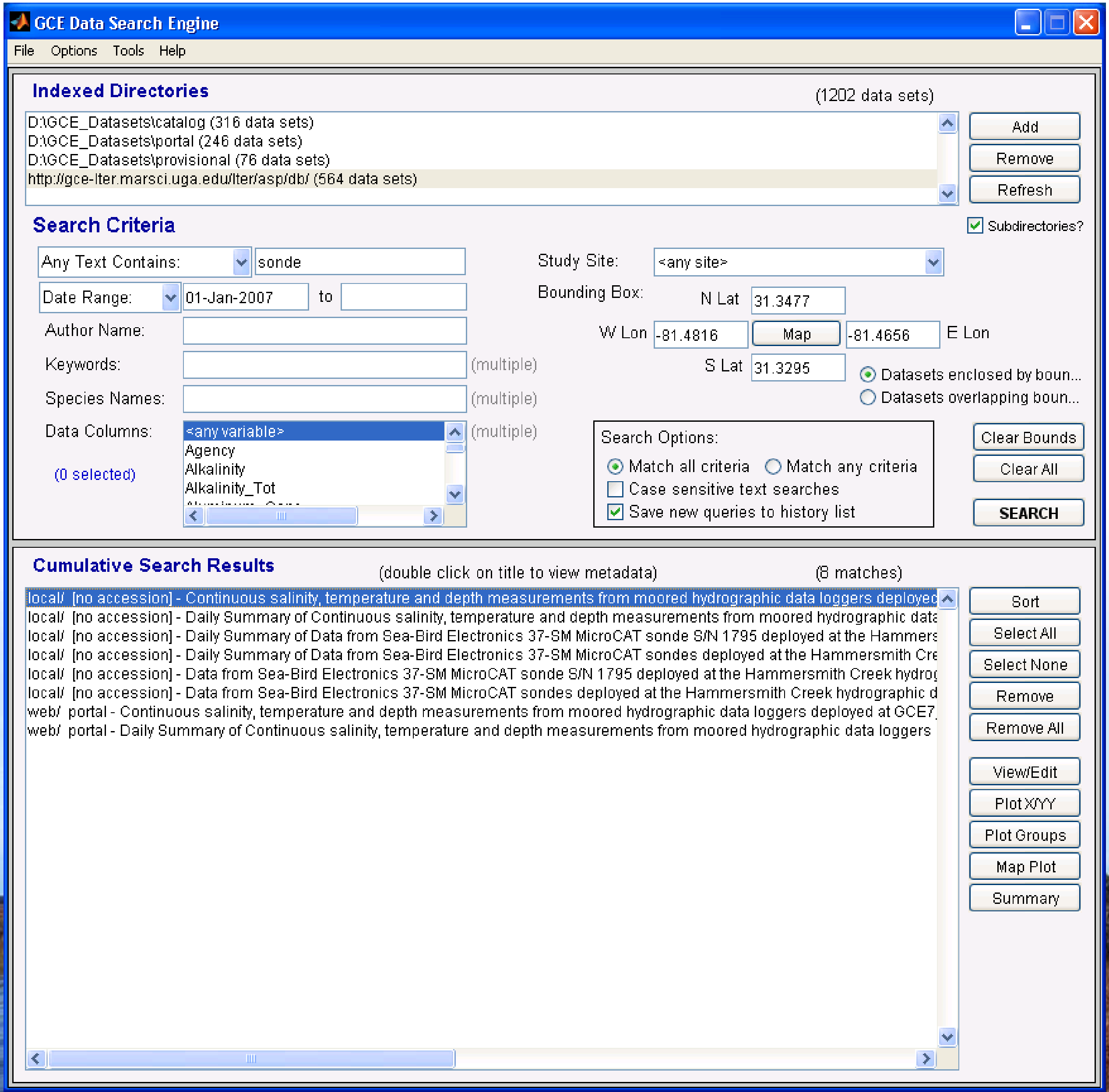
Flag Criteria

Data Viewer/Editor



All	Site	Longitude	Latitude	Instrument	Pump	Date	Year	Month	Day
None	(none)	(degrees)	(degrees)	(none)	(none)	(serial day (base	(YYYY)	(MM)	(DD)
□ 1	7	-81.475500	31.338383	2398	0	733043.000000	2007	1	1
□ 2	7	-81.475500	31.338383	2398	0	733043.020833	2007	1	1
□ 3	7	-81.475500	31.338383	2398	0	733043.041667	2007	1	1
□ 4	7	-81.475500	31.338383	2398	0	733043.062500	2007	1	1
□ 5	7	-81.475500	31.338383	2398	0	733043.083333	2007	1	1
□ 6	7	-81.475500	31.338383	2398	0	733043.104167	2007	1	1
□ 7	7	-81.475500	31.338383	2398	0	733043.125000	2007	1	1
□ 8	7	-81.475500	31.338383	2398	0	733043.145833	2007	1	1
□ 9	7	-81.475500	31.338383	2398	0	733043.166667	2007	1	1
□ 10	7	-81.475500	31.338383	2398	0	733043.187500	2007	1	1
□ 11	7	-81.475500	31.338383	2398	0	733043.208333	2007	1	1
□ 12	7	-81.475500	31.338383	2398	0	733043.229167	2007	1	1
□ 13	7	-81.475500	31.338383	2398	0	733043.250000	2007	1	1
□ 14	7	-81.475500	31.338383	2398	0	733043.270833	2007	1	1
□ 15	7	-81.475500	31.338383	2398	0	733043.291667	2007	1	1
□ 16	7	-81.475500	31.338383	2398	0	733043.312500	2007	1	1
□ 17	7	-81.475500	31.338383	2398	0	733043.333333	2007	1	1
□ 18	7	-81.475500	31.338383	2398	0	733043.354167	2007	1	1
□ 19	7	-81.475500	31.338383	2398	0	733043.375000	2007	1	1
□ 20	7	-81.475500	31.338383	2398	0	733043.395833	2007	1	1
□ 21	7	-81.475500	31.338383	2398	0	733043.416667	2007	1	1
□ 22	7	-81.475500	31.338383	2398	0	733043.437500	2007	1	1
□ 23	7	-81.475500	31.338383	2398	0	733043.458333	2007	1	1
□ 24	7	-81.475500	31.338383	2398	0	733043.479167	2007	1	1
□ 25	7	-81.475500	31.338383	2398	0	733043.500000	2007	1	1

Data Search Engine



Indexed Directories (1202 data sets)

D:\GCE_Datasets\catalog (316 data sets)
D:\GCE_Datasets\portal (246 data sets)
D:\GCE_Datasets\provisional (76 data sets)
http://gce-lter.marsci.uga.edu/lter/asp/db/ (564 data sets)

☑ Subdirectories?

Search Criteria

Any Text Contains: sonde

Date Range: 01-Jan-2007 to

Author Name:

Keywords: (multiple)

Species Names: (multiple)

Data Columns: (multiple)
<any variable>
Agency
Alkalinity
Alkalinity_Tot

(0 selected)

Study Site: <any site>

Bounding Box: N Lat 31.3477

W Lon -81.4816 E Lon -81.4656

S Lat 31.3295

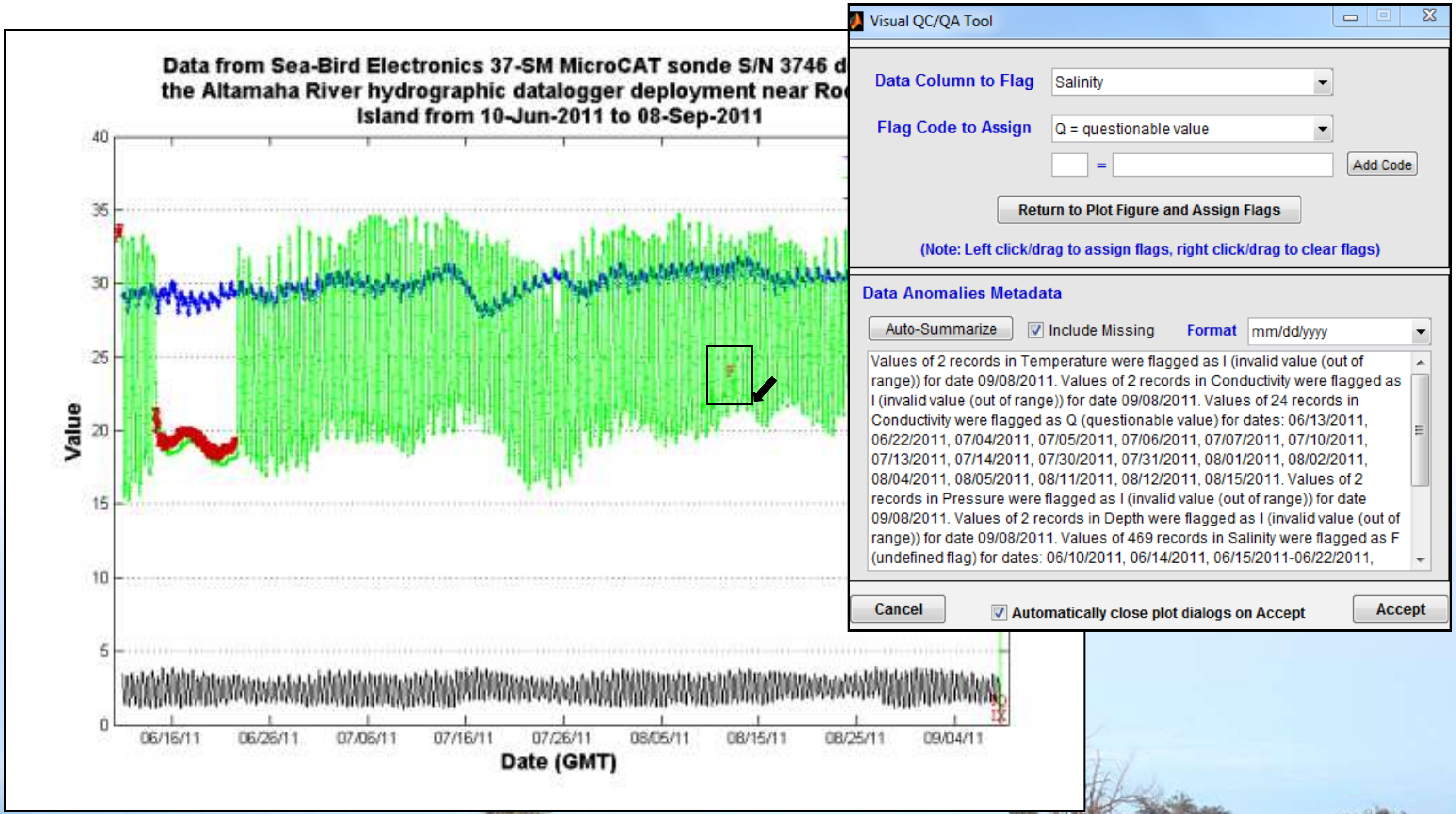
● Datasets enclosed by boun...
○ Datasets overlapping boun...

Search Options:
● Match all criteria ○ Match any criteria
□ Case sensitive text searches
☑ Save new queries to history list

Cumulative Search Results (double click on title to view metadata) (8 matches)

local/ [no accession] - Continuous salinity, temperature and depth measurements from moored hydrographic data loggers deployed
local/ [no accession] - Daily Summary of Continuous salinity, temperature and depth measurements from moored hydrographic data
local/ [no accession] - Daily Summary of Data from Sea-Bird Electronics 37-SM MicroCAT sonde S/N 1795 deployed at the Hammers
local/ [no accession] - Daily Summary of Data from Sea-Bird Electronics 37-SM MicroCAT sondes deployed at the Hammersmith Cre
local/ [no accession] - Data from Sea-Bird Electronics 37-SM MicroCAT sonde S/N 1795 deployed at the Hammersmith Creek hydrog
local/ [no accession] - Data from Sea-Bird Electronics 37-SM MicroCAT sondes deployed at the Hammersmith Creek hydrographic d
web/ portal - Continuous salinity, temperature and depth measurements from moored hydrographic data loggers deployed at GCE7
web/ portal - Daily Summary of Continuous salinity, temperature and depth measurements from moored hydrographic data loggers

Interactive Plotting & Q/C Tools


Data from Sea-Bird Electronics 37-SM MicroCAT sonde S/N 3746 d
the Altamaha River hydrographic datalogger deployment near Ro
Island from 10-Jun-2011 to 08-Sep-2011
Value
40
35
30
25
20
15
10
5
0
06/16/11
06/26/11
07/06/11
07/16/11
07/26/11
08/05/11
08/15/11
08/25/11
09/04/11
Date (GMT)

Visual QC/QA Tool

Data Column to Flag: Salinity

Flag Code to Assign: Q = questionable value

= Add Code

Return to Plot Figure and Assign Flags

(Note: Left click/drag to assign flags, right click/drag to clear flags)

Data Anomalies Metadata

Auto-Summarize ☑ Include Missing Format: mm/dd/yyyy

Values of 2 records in Temperature were flagged as I (invalid value (out of range)) for date 09/08/2011. Values of 2 records in Conductivity were flagged as I (invalid value (out of range)) for date 09/08/2011. Values of 24 records in Conductivity were flagged as Q (questionable value) for dates: 06/13/2011, 06/22/2011, 07/04/2011, 07/05/2011, 07/06/2011, 07/07/2011, 07/10/2011, 07/13/2011, 07/14/2011, 07/30/2011, 07/31/2011, 08/01/2011, 08/02/2011, 08/04/2011, 08/05/2011, 08/11/2011, 08/12/2011, 08/15/2011. Values of 2 records in Pressure were flagged as I (invalid value (out of range)) for date 09/08/2011. Values of 2 records in Depth were flagged as I (invalid value (out of range)) for date 09/08/2011. Values of 469 records in Salinity were flagged as F (undefined flag) for dates: 06/10/2011, 06/14/2011, 06/15/2011-06/22/2011,

Cancel ☑ Automatically close plot dialogs on Accept Accept

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
- Data structure instances are independent
  - 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") information captured for all steps
  - Each tool logs operations by date/time
  - Data revisions, deletions, flagging captured at user-specified detail
  - Lineage reported in metadata
- Dataset metadata is "live", and updated automatically
  - Attribute changes
  - Calculations, unit conversions
  - Code definitions

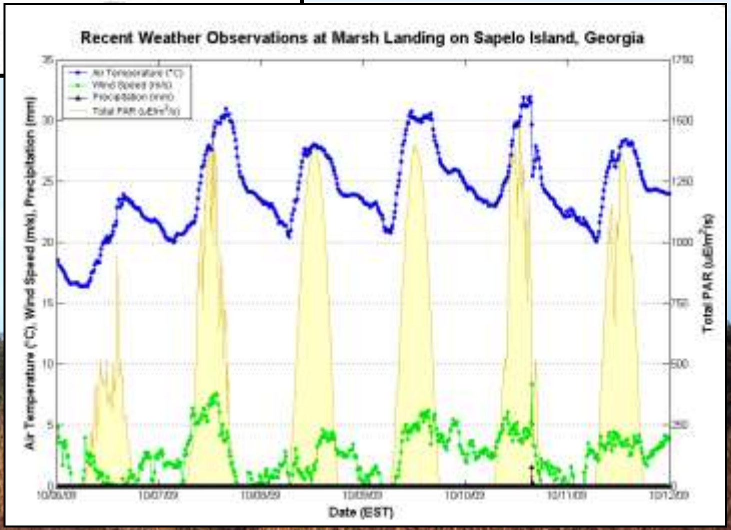
Suitability for Real-Time Sensor Data

- Good Scalability
  - Data volumes only limited by computer memory (tested >2 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 for real-time GCE data, USGS data harvesting service (LTER HydroDB, CWT)

Real-Time GCE Data Harvesting

Georgia Coastal Ecosystems LTER

Dataset Details



Recent Weather Observations at Marsh Landing on Sapelo Island, Georgia

Implementation Scenarios

- End-to-End Processing (logger-to-scientist)
  - Acquire raw data from logger, file system, network (CIFS,HTTP,FTP,SOAP)
  - 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
- Data Pre-processing
  - Acquire, validate and flag raw data (on demand or timed/triggered)
  - Upload processed data files (e.g. csv) or value & flag arrays to RDBMS (e.g HIS)
- Workflow Step
  - Call toolbox from other software as part of workflow (e.g. LoggerNet)
  - Kepler via MATLAB actor
  - DataTurbine via MATLAB off-ramp or Java API

Toolbox History

- 2001 – Initial toolbox development in Feb 2001 (revised in May to add dynamic QA/QC)
- 2002 – Added GUI, released code to GCE affiliates
- 2003 – Added dynamic data harvesting support (USGS, NOAA, CSI LoggerNet); automated USGS harvesting service for ClimDB/HydroDB
- 2004 – Added "search engine" tool for local search/integration of data
- 2005 – First public distribution of "compiled" code; source code on request to LTER sites
- 2006 – Added ClimDB data mining GUI
- 2007 – Added enhanced data sythesis, refactoring tools
- 2008 – Added GUI for managing QA/QC rules in metadata templates, additional flag tools
- 2009 – Refined XML schema for formatted metadata; code moved to SVN
- 2010 – Toolbox released as open source (GPLv3); Trac support site established
- 2011 – Expanded QA/QC tool options, refinements; focused on usability
- 2012 – Added prototype EML support, GUI for batch processing (import/export)

Concluding Remarks

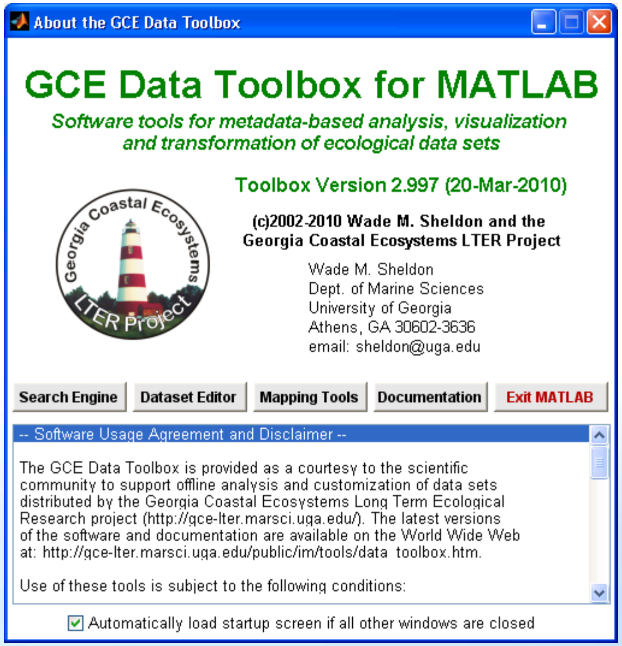
- "Fine Print"
  - Requires MATLAB ($ academic, $$$ government/industry)
  - Software documented, but more tutorial and training materials needed (planned)
  - Support is limited (unfunded outreach)
- Benefits
  - Fully cross-platform (Windows, MacOS, Linux, Solaris)
  - Mature – used 24/7 for over 11 years for LTER data management (>3000 dl's)
  - GCE Data Toolbox is free and open source (GPL) – can customize, redistribute
- More information and downloads at:

  https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox

(This work was supported by NSF grants OCE-9982133 and OCE-0620959)

Interactive Demo & Discussion

- Installing and launching the GCE Data Toolbox
- Loading data
  - Pre-defined import filter and metadata template
  - Generic text file and new metadata template
- Performing QA/QC
  - Defining and editing "rules"
  - Interactive QA/QC flag visualization, revision
  - Managing flagged values
- Post-processing data
  - Filtering, basic gap filling, drift correction
  - Statistical re-sampling, binning
  - Integrating multiple data sets
- Batch processing & export


About the GCE Data Toolbox
Georgia Coastal Ecosystems LTER Project
Search Engine
Dataset Editor
Mapping Tools
Documentation
Exit MATLAB

GCE Data Toolbox for MATLAB

Software tools for metadata-based analysis, visualization and transformation of ecological data sets

Toolbox Version 2.997 (20-Mar-2010)

(c)2002-2010 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

-- 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/im/tools/data toolbox.htm.

Use of these tools is subject to the following conditions:

☑ Automatically load startup screen if all other windows are closed