GCE Data Toolbox for MATLAB – An Introduction

Wade Sheldon
Georgia Coastal Ecosystems LTER
Background & Motivation

- Georgia Coastal Ecosystems LTER project started in May 2000
  - Major data collection effort
  - 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
GCE Data Model

- Started by reviewing ESA’s “FLED” report

- 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)
  - Added later: quality control rules for every variable, 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
Conceptual Model of the GCE Data Structure Specification
(version 1.1, 29-Mar-2001)

Note: GCE Data Structures are created and managed using the GCE Data Toolbox, a MATLAB software library for metadata-based analysis, visualization, transformation and management of ecological data sets (url: http://gce-lter.marisci.uga.edu/lter/research/tools/data_toolbox.htm)

Dataset Metadata
- Structure Information (fields: title, version, createdate, edittedate)
- Dataset Lineage (fields: datafile, history)
- General Metadata (parseable array) (field: metadata)

Attribute Metadata
- name
- description
- units
- datatype
- variablenotype
- numbertype
- precision
- criteria
- values
- flags

Data & Flag Arrays
- Attribute 1
- Attribute 2
- Attribute 3
- Attribute n
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)
Startup Window (ui_aboutgce.m)

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 30602-3535
e-mail: 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://gcce-lter.marshsci.uga.edu). The latest versions of the software and documentation are available on the World Wide Web at: http://gcce-lter.marshsci.uga.edu/public/itm/tools/data_toolbox.htm.

Use of these tools is subject to the following conditions:

☑ Automatically load startup screen if all other windows are closed
Data Editor (ui_editor.m)
## Data Viewer (ui_datagrid.m)

![Data Viewer Image]

### Data Table

<table>
<thead>
<tr>
<th>All</th>
<th>Site</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Instrument</th>
<th>Pump</th>
<th>Date</th>
<th>Year</th>
<th>Month</th>
<th>Day</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(degrees)</td>
<td>(degrees)</td>
<td>(none)</td>
<td>(none)</td>
<td>(serial day base)</td>
<td>(YYYY)</td>
<td>(MM)</td>
<td>(DD)</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.00030000</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.020333</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.041567</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.062500</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.083333</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.104167</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.125000</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.145333</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.165967</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.187500</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.209333</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.229167</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.250000</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.270333</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.291567</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.312500</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.333333</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.354167</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.375000</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.395833</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.416567</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.437500</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.458333</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.479167</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>-81.475500</td>
<td>31.338383</td>
<td>2308</td>
<td>0</td>
<td>733043.500000</td>
<td>2007</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Command Line

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

   version: 'GCE Data Structure 1.1 (29-Mar-2001)'
title: 'Data from USGS Station 02226000 (ALUMASHA RIVER AT DOCTOR TOWN, GA) for 03-Feb-2010 through 05-Apr-2010'
actdata: (0x93 cell)
datfile: ('usgs_02226000_realtime_20100406_1130_mod.txt' [5797])
createdate: '06-Apr-2010 11:30:49'
editedate: '06-Apr-2010 11:30:49'
history: (1x82 cell)
names: (1x12 cell)
units: {'none' 'none' 'none' 'serial day (base 1/1/2000) - GMT' 'YYYY' 'MM' 'DD' 'hr' 'min' 'm' 'a' '3/sec' 'mm')
description: (1x12 cell)
datatype: {'g' 'g' 'd' 'd' 'd' 'd' 'd' 'd' 'd' 'd' 'd' 'f'}
variabletype: (1x12 cell)
numberof: (1x12 cell)
precision: [0 0 0 0 0 0 0 0 2 1 2]
values: (1x12 cell)
criteria: (1x12 cell)
flags: {''''''''''''''''''''''''''}

>> listclos(s)

ans =

1: Agency -- string
2: StationID -- string
3: Provisional -- integer
4: Date serial day (base 1/1/2000) - GMT -- floating-point
5: Year YYYY -- integer
6: Month MM -- integer
7: Day DD -- integer
8: Hour hh -- integer
9: Minute (min) -- integer
10: GaugeHeight (m) -- floating-point
11: Discharge m^3/sec -- floating-point
12: Precipitation (mm) -- floating-point

>> dt = extract(s, 'Date'); discharge = extract(s, 'Discharge');
```

```
Current Toolbox Use

- Major use is GCE data processing, harvesting, distribution
- Also used for general data mining, instrument data acquisition
  - Internet Import: USGS NWIS, NOAA HADS, LTER ClimDB
  - File Import: NOAA NCDC, NERR CDMO, generic MATLAB, generic text
  - Instruments: Seabird ctd/sonde, Campbell loggers, Hydrolab, Aquatroll, Hobo Tidbit
- Being used by other LTER sites (CWT, SEV, PIE?, VCR?)
- ~3000 downloads by non-GCE web visitors to date
- Potentially very beneficial for GCE data users, but under-used
  - Customize data set layout, format for analyses
  - Handle QA/QC flagged and missing values
  - Re-sample data (date-time aggregation, grouping, binning)
  - Convert units, harmonize column names for comparison with other data
Getting the Toolbox

- Public “compiled” & GCE “source” versions available on GCE web site
- Source code provided to other LTER sites, collaborators on request
- Code copyrighted to control redistribution
  - Download/usage tracking for NSF
  - Prevent “forking” development, external dependencies
  - Preserve funding opportunities (GCE not funded for software dev)

- New: Software development site/SVN repository
  - [https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox](https://gce-svn.marsci.uga.edu/trac/GCE_Toolbox)
  - Requires login - email sheldon@uga.edu for access