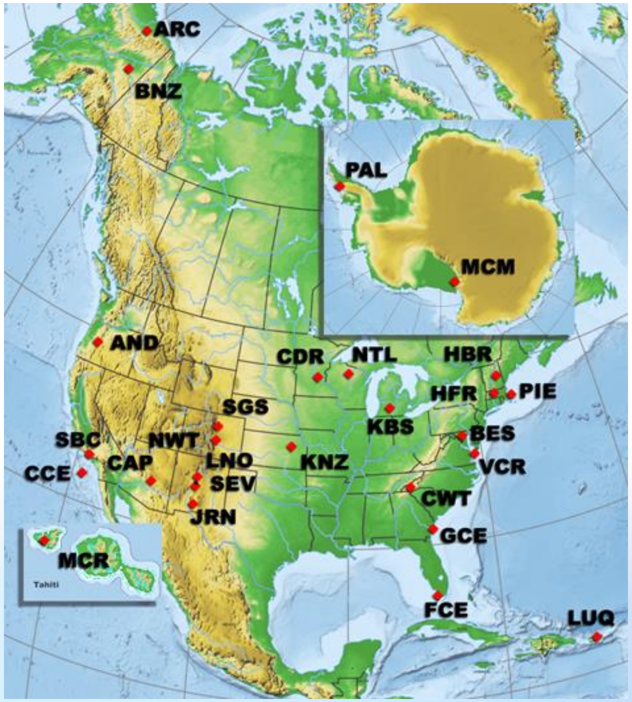
Managing Data, Provenance and Chaos through Standardization and Automation at the Georgia Coastal Ecosystems LTER Site

Wade Sheldon
Georgia Coastal Ecosystems LTER
University of Georgia

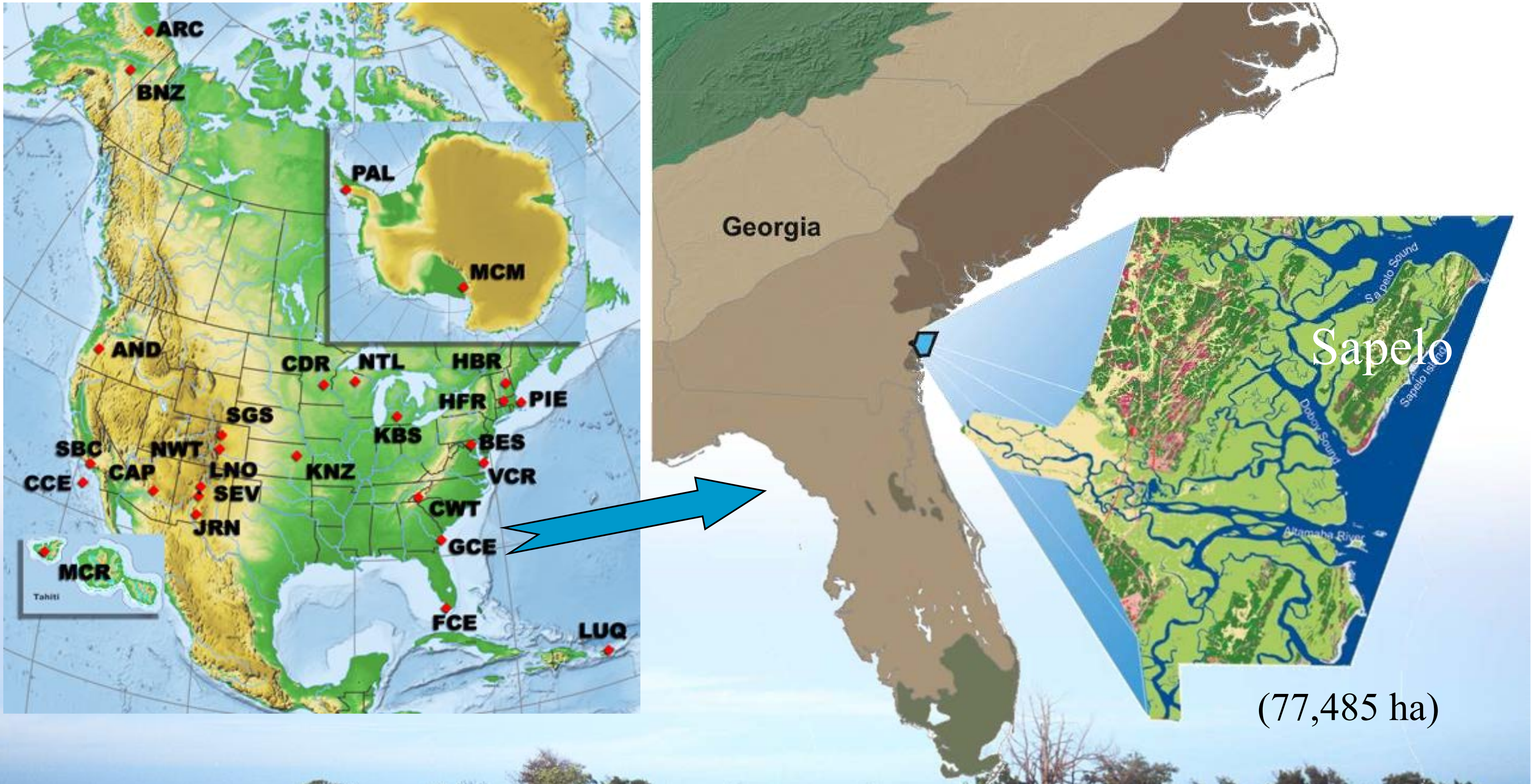
IN51D-05: Data Stewardship in Theory and in Practice
AGU Fall Meeting, 13-Dec-2013

Background

- Long Term Ecological Research Network (LTER) established by NSF in 1980
  - Research ecological issues that can last decades and span huge geographical areas
  - Site-based research in different biomes, unified by common themes (core areas)
  - 29 sites established over 33 years (25 active), plus Network Office
- Georgia Coastal Ecosystems LTER (GCE) funded in 2000
  - Originated from Georgia Rivers LMER (1994-1999): transport and transformation of organic and inorganic materials carried from the land into the sea
  - GCE-1 (2000-2006): patterns of variability in estuarine processes
  - GCE-2 (2006-2012): extent to which gradients in water inflow drive landscape patterns
  - GCE-3 (2012-2018): how variations in salinity and inundation, driven by climate change and anthropogenic factors, affect biotic and ecosystem responses at different spatial and temporal scales



Geographic Setting



(77,485 ha)

Data Stewardship Challenges

- Research is conducted within multiple, overlapping domains
  - Network of 25 LTER sites
  - Team of 21 investigators from 8 institutions
  - Field site operated by UGA, on state DNR-managed land within National Estuarine Research Reserve
  - Many related/leveraged projects
- Multidisciplinary research leads to highly diverse data
  - Analytical lab data
  - Ecological field/experiment data
  - Oceanographic cruise data
  - Sensor data (10 Hz – 1hr)
  - Remote sensing
  - Genomics analysis
  - Archeological data

Georgia Coastal Ecosystems LTER Project

Data Stewardship Challenges

- Change is the only constant
  - Changes in goals at the network, site level
  - Changes in expectations (NSF, LTER, scientific community, users)
  - Changes in standards, new standards
  - Changes in technology, security practices
- Information continually accrues
  - Long-term curation intrinsic to LTER mission
  - Need to add the new while keeping the old
- Resources never keep pace with needs
  - LTER sites flat-funded for 6+ year cycles
  - No additional resources to manage legacy data/information

Opportunities

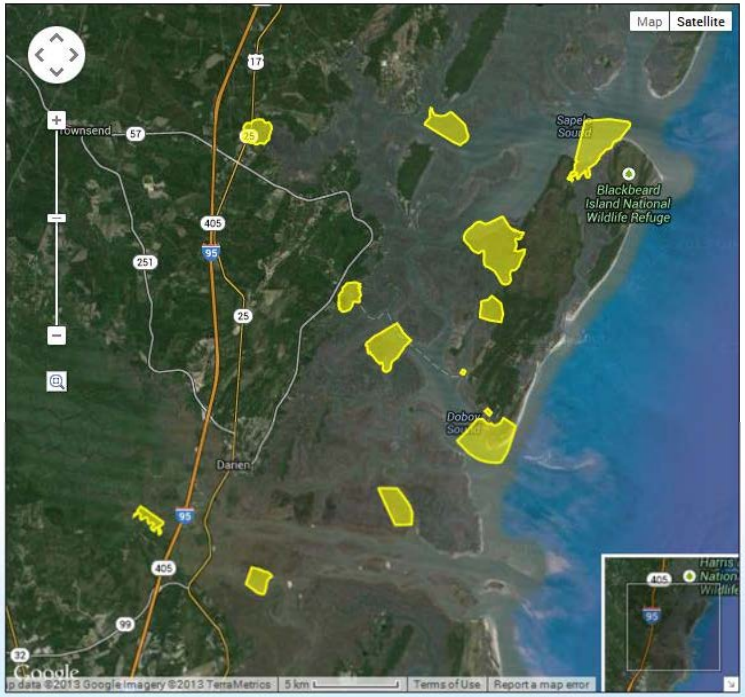
- Domain affiliations add context, standards that can be incorporated
- Proposals provide unifying structure for research – link everything
- Long-term funding model encourages long-term thinking and approaches
- Strong commitment to data management across LTER
  - Peer learning opportunities
  - Leverage expertise, infrastructure through collaboration
  - Network support, resources

Strategies for Data Management

- Standardize to manage diversity and complexity
- Automate to improve efficiency, scalability
- Modularize information systems to accommodate change
- Collaborate to share the load

Standardization

- Geographic terms (site/ location, transect/station, plot, well, mooring,…) and place names
- Project organization terms (roles, member types, study types, project types)
- Identifiers for personnel, data sets, taxa, citations, documents
- Keyword vocabularies
- Data formats, units of measure



Primary Sampling Sites

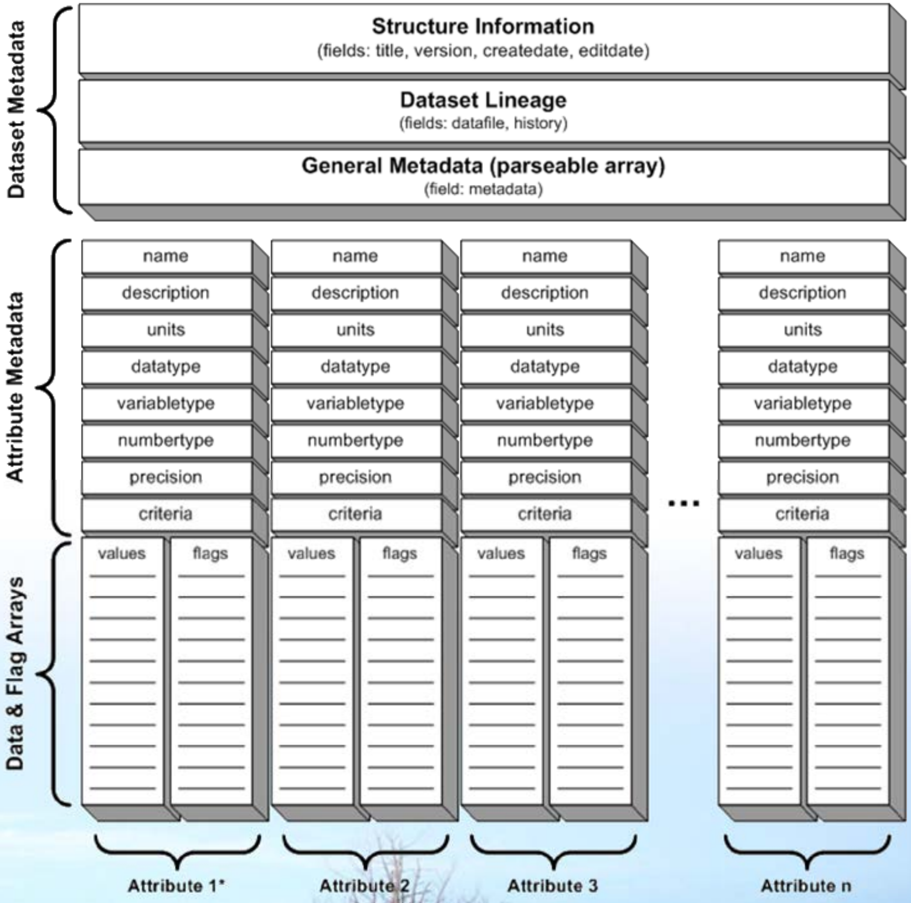
GCE1 (Eulonia)
GCE2 (Four Mile Island)
GCE3 (North Sapelo)
GCE4 (Meridian)
GCE5 (Folly River)
GCE6 (Dean Creek)
GCE7 (Carrs Island)
GCE8 (Alligator Creek)
GCE9 (Rockdedundy Island)
GCE10 (Hunt Camp)

Auxiliary Sites

ML (Marsh Landing)
UGAMI (UGA Marine Institute)
KF (Kenan Field)
ALT-BASIN (Altamaha River Basin
- upriver sites in Altamaha watershed not shown on map)

Standardization

- Tabular data model (GCE Data Structure)
  - Any number of variables
  - Attribute metadata for each variable (name, units, description, type, precision)
  - Structured documentation metadata
  - Processing history (lineage)
  - Q/C rules for every variable
  - Qualifier flags for every value



Dataset Metadata

Structure Information
(fields: title, version, createdate, editdate)

Dataset Lineage
(fields: datafile, history)

General Metadata (parseable array)
(field: metadata)

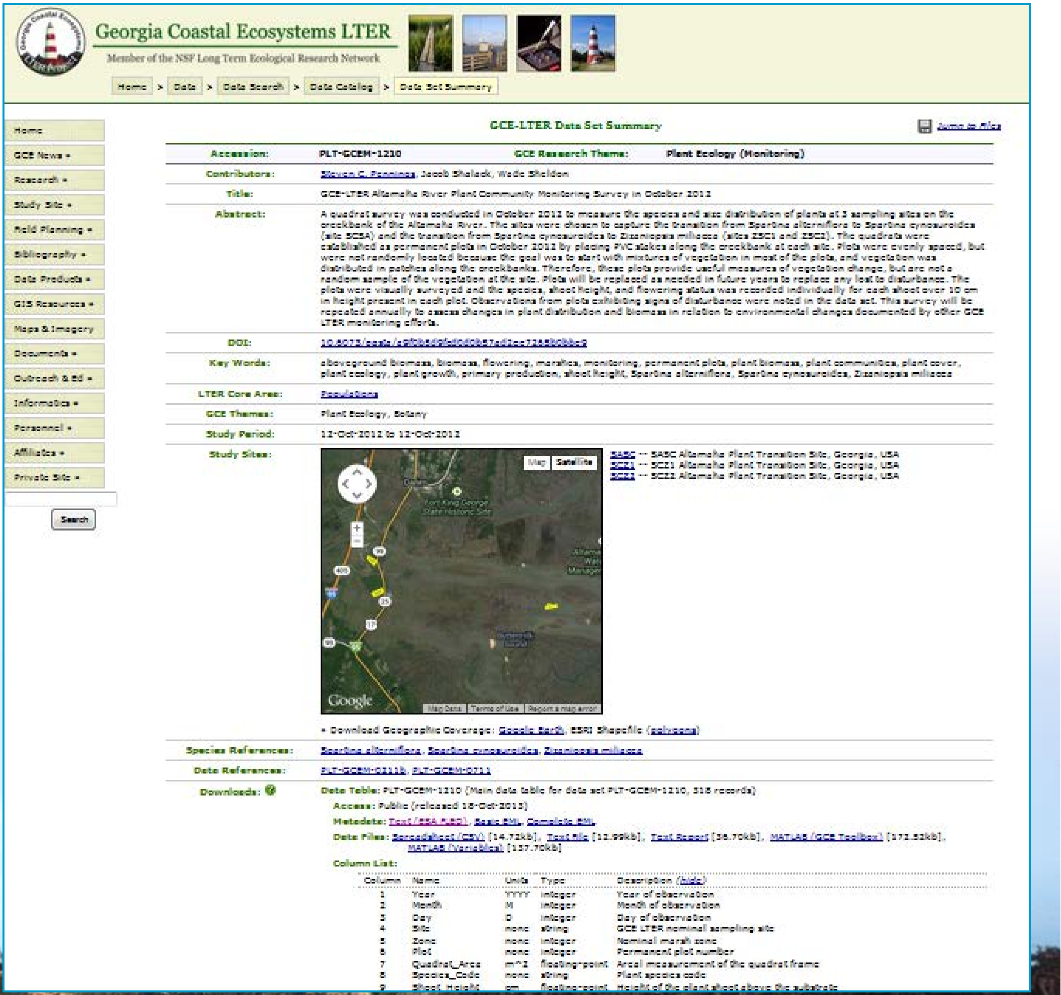
Attribute Metadata

Attribute 1*	Attribute 2	Attribute 3	...	Attribute n
name	name	name		name
description	description	description		description
units	units	units		units
datatype	datatype	datatype		datatype
variabletype	variabletype	variabletype		variabletype
numbertype	numbertype	numbertype		numbertype
precision	precision	precision		precision
criteria	criteria	criteria		criteria
values, flags	values, flags	values, flags		values, flags

Data & Flag Arrays

Automation

- Relational databases store all project information to limit redundancy, support lookups
- Dynamic web pages, services provide dynamic linking, keep everything in sync
- Data management software (GCE Data Toolbox) automates tabular data processing, metadata generation, Q/C, synthesis, harvesting
- Metadata Management System (Metabase) – dynamically generates, versions, publishes data set metadata to manage distribution, minimize maintenance



GCE-LTER Data Set Summary

Accession: PLT-GCEM-1210 **GCE Research Theme:** Plant Ecology (Monitoring)

Contributors: Steven C. Pennings, Jacob Shalack, Wade Sheldon

Title: GCE-LTER Altamaha River Plant Community Monitoring Survey in October 2012

Abstract: A quadrat survey was conducted in October 2012 to measure the species and size distribution of plants at 3 sampling sites on the creekbank of the Altamaha River. The sites were chosen to capture the transition from Spartina alterniflora to Spartina cynosuroides (site SC5A) and the transition from Spartina cynosuroides to Zizaniopsis miliacea (sites ZSC1 and ZSC2). The quadrats were established as permanent plots in October 2012 by placing PVC stakes along the creekbank at each site. Plots were evenly spaced, but were not randomly located because the goal was to start with mixtures of vegetation in most of the plots, and vegetation was distributed in patches along the creekbanks. Therefore, these plots provide useful measures of vegetation change, but are not a random sample of the vegetation at the site. Plots will be replaced as needed in future years to replace any lost to disturbance. The plots were visually surveyed and the species, shoot height, and flowering status was recorded individually for each shoot over 10 cm in height present in each plot. Observations from plots exhibiting signs of disturbance were noted in the data set. This survey will be repeated annually to assess changes in plant distribution and biomass in relation to environmental changes documented by other GCE LTER monitoring efforts.

DOI: 10.6073/pasta/a9f0b8d9fd040b57ad2ec7188b0bbc9

Key Words: aboveground biomass, biomass, flowering, marshes, monitoring, permanent plots, plant biomass, plant communities, plant cover, plant ecology, plant growth, primary production, shoot height, Spartina alterniflora, Spartina cynosuroides, Zizaniopsis miliacea

LTER Core Area: Populations

GCE Themes: Plant Ecology, Botany

Study Period: 12-Oct-2012 to 12-Oct-2012

Study Sites:
SASC -- SASC Altamaha Plant Transition Site, Georgia, USA
SCZ1 -- SCZ1 Altamaha Plant Transition Site, Georgia, USA
SCZ2 -- SCZ2 Altamaha Plant Transition Site, Georgia, USA

▪ Download Geographic Coverage: Google Earth, ESRI Shapefile (polygons)

Species References: Spartina alterniflora, Spartina cynosuroides, Zizaniopsis miliacea

Data References: PLT-GCEM-0211b, PLT-GCEM-0711

Downloads: **Data Table:** PLT-GCEM-1210 (Main data table for data set PLT-GCEM-1210, 318 records)

Access: Public (released 18-Oct-2013)

Metadata: Text (ESA FLED), Basic EML, Complete EML

Data Files: Spreadsheet (CSV) [14.72kb], Text file [12.99kb], Text Report [36.70kb], MATLAB (GCE Toolbox) [172.52kb], MATLAB (Variables) [137.70kb]

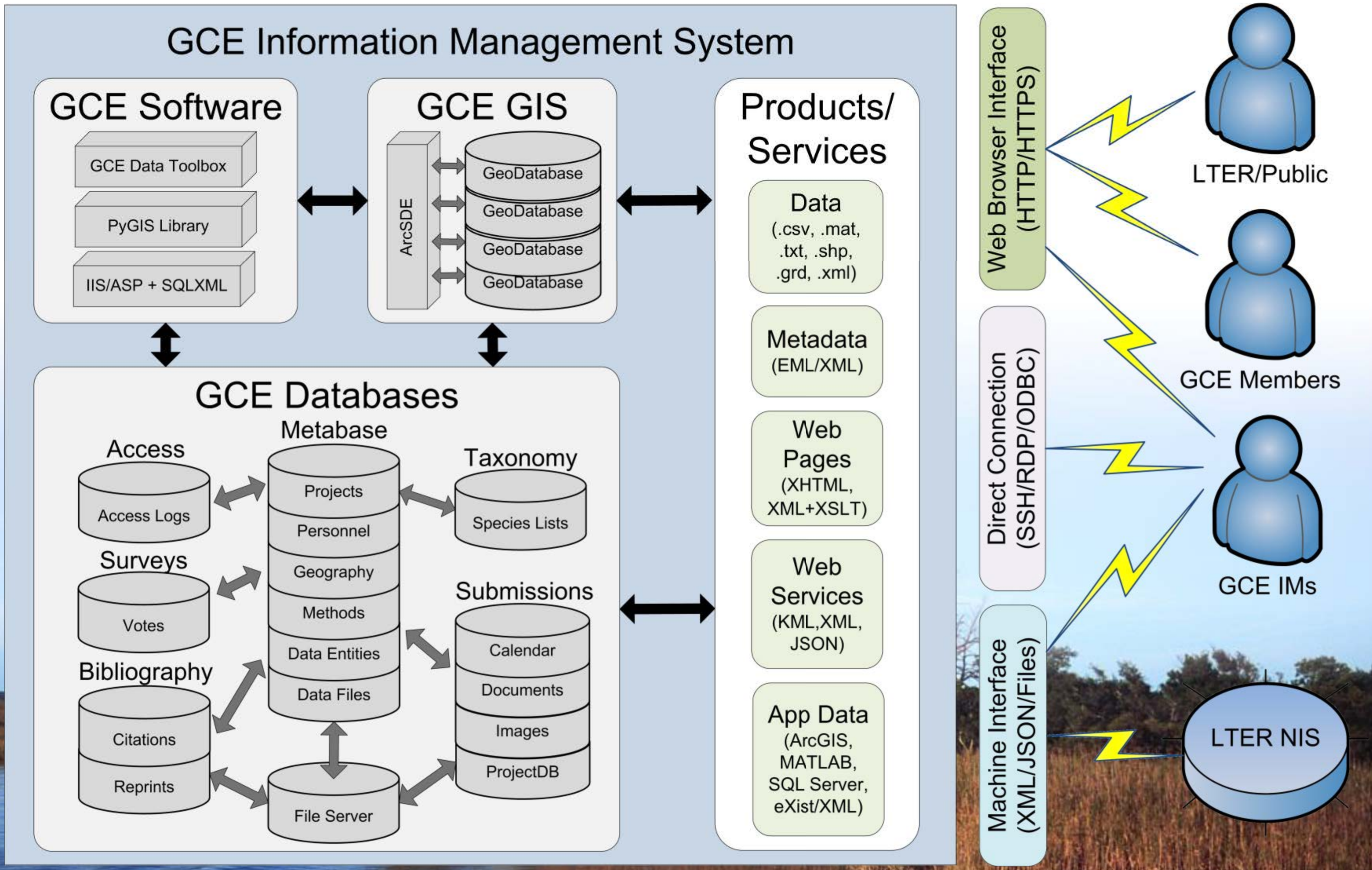
Column List:

Column	Name	Units	Type	Description (hide)
1	Year	YYYY	integer	Year of observation
2	Month	M	integer	Month of observation
3	Day	D	integer	Day of observation
4	Site	none	string	GCE LTER nominal sampling site
5	Zone	none	integer	Nominal marsh zone
6	Plot	none	integer	Permanent plot number
7	Quadrat_Area	m^2	floating-point	Areal measurement of the quadrat frame
8	Species_Code	none	string	Plant species code
9	Shoot_Height	cm	floating-point	Height of the plant shoot above the substrate

(http://gce-lter.marsci.uga.edu/data/PLT-GCEM-1210)

Modularization

- Modularization of information system components, linked by stable identifiers and APIs, permits adaptation over time



Collaboration

- Collaborate broadly inside/outside LTER
  - Closely with 3 other sites (CWT, SBC, MCR)
  - LTER and other informatics working groups
- Collaboration has provided many tangible benefits
  - Access to additional expertise, IT resources
  - Expanded use cases to improve software/database designs
  - Help testing/debugging code
  - Opportunities to standardize approaches when common needs identified
- Collaboration also has intangible benefits
  - Learning through teaching, mentoring others
  - Opportunity to work with others in the same discipline

Tracking Provenance

- Provenance is critical for any long-term, multi-investigator project
  - Instruments, methods, processing can vary over time
  - Personnel contact information changes over time
  - Practices and data systems constantly evolving (information can be lost)
- Standardization and automation key to provenance tracking at GCE
  - Terms and stable identifiers link everything together
  - Canonical databases ensure updates are global
  - Automated metadata generation, publishing keeps info updated even in external repositories
  - Automated capture of metadata, Q/C operations and lineage in the GCE Data Toolbox simplifies managing provenance of tabular data

Lessons Learned

- It’s far easier to standardize up front than harmonize later
- Consistently structuring metadata content and data is critically important
- What format/system you store structured information in (RDBMS, XML, JSON) is less important, and will likely change over time
- The lines between metadata and data get blurrier all the time, so be prepared for change
- The key to getting data from investigators is providing them with a useful service, so design with that in mind (handyman vs tax man)