### 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

Georg



# 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

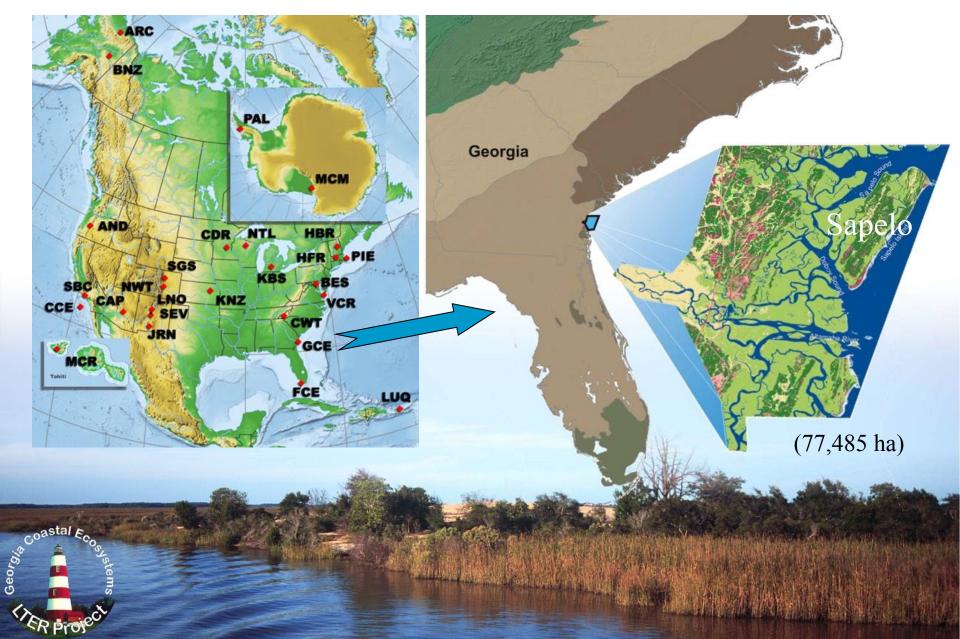
Coastal A

Georg

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



## 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

Georgi

- Genomics analysis
- Archeological data







### 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

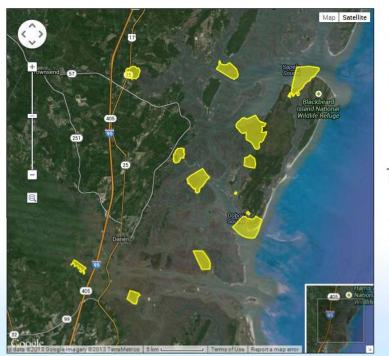
### Standardization

- 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

Coasta/

Georg

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

<u>ML</u> (Marsh Landing) <u>UGAMI</u> (UGA Marine Institute) <u>KF</u> (Kenan Field) <u>ALT-BASIN</u> (Altamaha River Basin

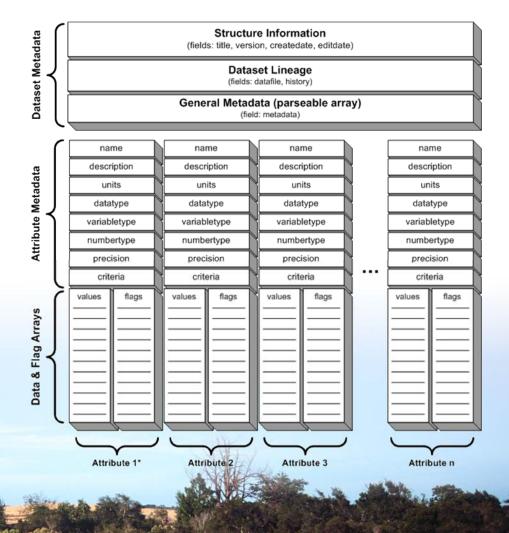
 upriver sites in Altamaha watershed not shown on map)

## 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

oastal A



### Automation

GCE

GIS

Affilia

Prive

- 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

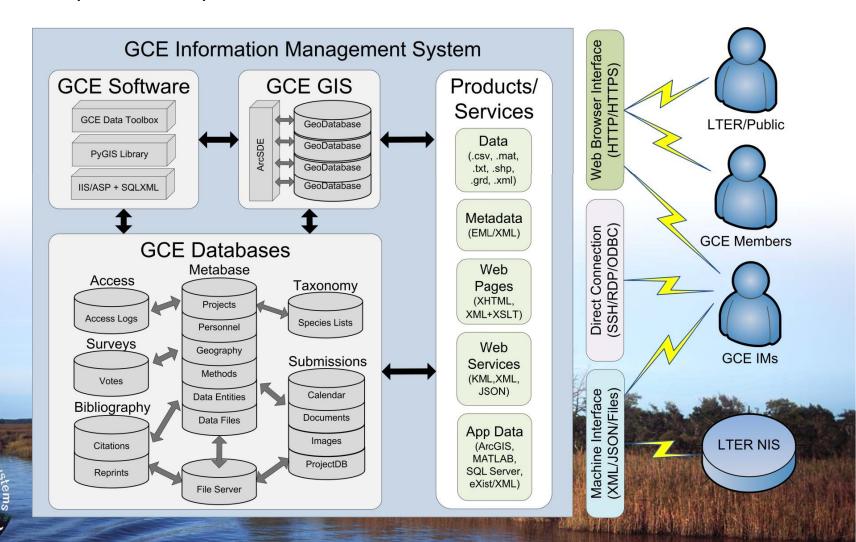
Accession:     PLT-CCEM-1230     GCE Research Thema:     Find Ecology (Membering)       Centributors:     Sizes C. Jacobigs Jord Struct Barlow, Wold Shelden     This:     GCE-USB Allanade Store Flan Community Membering Survey in October 2013       Akstenct:     Advantation Store Stor		F Long Term Ecological I	Research Network	La Set Summery		ð 🌬			
Contributors:     Start CL Pandarg, look Bielder, Nade Bielder,     Dur um       Yile:     Get-UtRA Alexanda Kyer Fan Community Hendering Survey in Otdoer 2012       Abstrett:     Abstrett:       Yr     Abstrett:       Abstrett:     Start CL Pandarg, look Bielder, Niede Bielder, Start Start, Bastret, B					GCE-LT	ER Data Sct S	iummary		🔛 Juma 🕫 🕅
Contributors:     Start CL Pandarg, look Bielder, Nade Bielder,     Dur um       Yile:     Get-UtRA Alexanda Kyer Fan Community Hendering Survey in Otdoer 2012       Abstrett:     Abstrett:       Yr     Abstrett:       Abstrett:     Start CL Pandarg, look Bielder, Niede Bielder, Start Start, Bastret, B		Accession:	PLT-GCEM-1210		GCE	Research They		Flent Ecology (Monitoring)	
11       GET-158 Alamaka River Plant Community Mentoning Burver (in Clobber 2011         A basinesi:       A		Contributors:		Jacob Shalack	0.650				
9       Abstract:       Abstrat:       Abstract: <td< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td>vev in Orie</td><td>Ber 2012</td><td></td></td<>							vev in Orie	Ber 2012	
<ul> <li>exception of the Alfamatia Nova, the site work obtain to applie the set built of the Alfamatia Nova in the Set of the Alfamatia Nova in the Set of the Set o</li></ul>	0								et 3 secolos stes os file
<ul> <li>decidation produce along the procession is. The will be a procession of weighteen at the procession of weightee</li></ul>			ereckbank of the A (site SCSA) and the established as per-	ltamaka River, 1 : transition from manent plots in 0	The sites Sparting Detailor 2	were chosen to cynosuroides to 1012 by placing	capture the o Zizaniopa PVC stakes	c transition from Sparting alterni is miliacca (sites 25C1 and 25C2 along the creekbank at each sit	flora to Spartina cynosuroldos (). The quadrats were b. Piots were evenly spaced, but
in herde present in each piet. Observations from piets callebra piet of subscheme vert noted in the data set. This survey received a multiple barget descended by UTER monitoring offers. POR: 12.511:each (absorbed biologitalization and biomean in elicitants to terminemental danget descended by UTER monitoring offers. Rev Words: callebra Area: Esculation callebra and biomean in elicitants biologitalization and biomean in elicitants biomean in elicitants by UTER monitoring offers. Rev Words: callebra Area: Esculation callebra and biomean in elicitants biomeans, plant callebra and biomeans and callebra and biomeans and plant callebra and biomeans and plant callebra and callebra and and and biomeans and callebra and biomeans and plant callebra and a plant callebra and biomeans and biomeans and callebra and biomeans and callebra and biomeans and callebra and and and and and and and and and an			distributed in patch random sample of	tes along the ore the vegetation a	ckbanks t the site	. Therefore, the . Plots will be re	cac plots pri splated as i	ovide useful measures of vegeta needed in future years to replace	tion change, but are not a e any lost to disturbance. The
D0:       10.001/destals/biblio/doiled1sid/action1sid/biblio/doiled1sid/action2.permanent/piple.glant/biomas.plant communities.plant and tenders, plant doiled1.plant.biomas.plant and tenders.plant doiled1.plant.biomas.plant and tenders.plant doiled1.plant.biomas.plant and tenders.plant.biomas.plant and tenders.plant.biomas.plan			in height present in repeated annually	to assess change	cryations	from plots exhi	biting sign:	a of disturbance were noted in th	e data set. This survey will be
Key Words:     a Bouggound biomaa, feworing, marking, marking, marking, permanet plot, plant bernau, plant demoundes, plant demoundes		DOI:	10.6073/easts/est	obsdafedodobs7	adles72	ashobbeg			
GCE Therms:       Plant Exclopy, Sciny         Study Parind:       12-Och2012 to 12-Och2012         Study Sites:       Image: Study Sites:         Image: Study Sites:       Image: Study Sites: Study Sites: Study Site:		Key Words:	aboveground biom plant coology, plan	ass, biomass, fle Cgrowth, primar	owering, ry produ	marañoa, monió clion, añoct hoig	oring, perm	manent plota, plant biomasa, plan a alterniflora, Spartina eynosuro	it communities, plant cover, lides, Zizanlogsis miliacea
GCT Themes:       Plant Exclopy, Solary         Study Pariod:       12-Oct-2012 to 12-Oct-2012         Study Sites:       Image: Study Sites:         Image: Study Sites:       Image: Study Site: Study Sites: Study Site: S		LTER Core Ares:	Peculations						
Study Sixes:       Study Sixes:         Bitudy Sixes:       Study Sixes:         Study Sixes:       Study Sixes:		GCE Themes:	Plant Boology, Bolany						
m       Section Section (section section secti		Study Period:	12-Oct-2012 to 12-	Oct-2012					
Species References:     Section alternifies, Santon sonzarodos, Santosan milana     Deta References:     Section alternifies, Santon sonzarodos, Santosan milana     Deta References:     Section alternifies, Santon sonzarodos, Santosan milana     Deta References:     Section alternifies, Santosan Ref., Santosan milana     Deta References:     Section alternifies, Santosan Ref., Santosan milana     Deta References:     Section alternifies, Santosan Ref., Santosan milana     Deta References:     Sectional (Internet)     Deter References:     Sectional (Internet)     Sectional (In		Study Sites:	$\bigcirc$	U.			SCZ1 SCI	21 Alternaha Plant Transition Site	c. Georgia, USA
Species References:         Species References:         Species References:         Species References:           Dets References:         Subsciences:         Subsciences:         Subsciences:         Subsciences:           Downloads:         Dets References:         Subsciences:         Subsciences:         Subsciences:           Dets References:         Dets References:         Dets References:         Subsciences:         Subsciences:           Downloads:         Dets References:         Dets References:         Subsciences:         Subsciences:           Dets References:         Column List:         Subscience:         Dets References:         Subscience:         Subscience: <td>_</td> <td></td> <td>and the second second</td> <td>125</td> <td></td> <td>N</td> <td></td> <td></td> <td></td>	_		and the second second	125		N			
Dete References:         Ext-GEEN-OTIL           Downloads:         Dete Telles: FUT-GEEN-IDID (Main data table for data sci FUT-GEEN-IDID, DIB records)           Access:: Full (relaxed 15:-Colorable)         Access:: Full (relaxed 15:-Colorable)           Metedete:: Tau: (FEE AND). Sain FMD, Gamaling EM, Dete Telles: Spraching: (FUT, Fuel ): Fuel Record (SA.FORD).         Metedete:: Tau: (FEE AND). Sain FMD, Gamaling EM, Dete Telles: Spraching: (FUT, Fuel ): Fuel Record (SA.FORD).           Column List:         Column Name         Units         Type           Column Name         Units         Type         Cestration           1         Year         Year         Type         Cestration           2         Dotto Mine         Type         Cestration         Sain Sain Sain Sain Sain Sain Sain Sain	62								
Downloads:         Deter Table: PUT-GCEM-1210 (Main data table for data set PUT-GCEM-1210, S18 records)           Access: Public (released 18-04-2013)         Metdeter: Tat/1614 F.M.D.L. Same SML Gamelate SML           Metdeter: Tat/1614 F.M.D.L. Same SML Gamelate SML         Metdeter: Tat/1614 F.M.D.L. Same SML Gamelate SML           Deter Files: SameSML (Tat/1616 [12.99kb], Tatt Record [36.70kb], Matuae (GCE Tatiker) [172.53kk         Metdeter: Tat/1614 (Marshing) [177.70kb]           Column List:         Column List:         Column List:           Column List:         Column List:         Column List:     <		Sector States and States							
Access: Public (released 18-0dr-1013) Metedete: <u>Test/SEA_Public Para Polis Commists PML</u> Deter Piers: <u>Terrorightest (FSU (PA 7068), Test Stevens</u> (56.7068), <u>MATUAS (GCE Testiken)</u> (172.5388 <u>MATUAS (Vara Strait</u> ) (127.7068) Column Linit Column Linit Column Linit Vara Virris depart (Stevenstein 1 Ward Minister Virris (Stevenstein 2 Mendh, Mi indeger Mend) of observation 3 Dopy D indeger Depy of Departuration 4 Ste none string CCETER cominal sampling ste 5 Zone none indeger Mennical sampling ste									
Dette Piler : <u>Sansadárbaci (CSP)</u> (14.7364), Tast Rig (13.9966), <u>Tast Ross</u> (34.7066), <u>MATLAR (GCE Tealban)</u> (172.5366 <u>NATLAR (Vancellan)</u> (127.7066) Column Linit Column		Access: Fublic (related 15-06-1013) Meteodets ( <u>fant (FSA AUD) Sama PM</u> , <u>Consiste PM</u> , Dete Files: <u>Scradialect (CDN</u> (14.7245), <u>Tant Mar (1</u> .9945), <u>Tant Recent</u> (58.7045), <u>MATLAS (GCS Teolbox)</u> (172.5245), <u>MATLAS (Accessical</u> (127.7045)							
MATLER (Variables)         T37.750b)           Column Linit         Units         Disservation           1         Varia         Virtin         Adaption           2         Mandh         M         Indeger         Mandh         Adaption           3         Day         D         Indeger         Day of observation           4         Site         none string         GCTER meminal sampling site           5         Zone         menservation									
Column Name Units Type Creardon ( <u>Nuits</u> ) 1 Yes (TYT) Indeper Vess of Anna Son 1 Yes (Yes) 2 Yes (Yes) 3 Yes (Yes) 3 Day D Indeper Day Estimation 4 Site none Indeper Reminal sampling site 5 Zon none Indeper Reminal sampling site									
1 Year Y111Y indeper Year of observation 2 Month M indeper Month of observation 3 Dey D indeper Dey of observation 4 Site mone string GCELTER norminal sampling site 5 Zone mone indeper Nerminal marpling site			and the second s						
2 Mendik M indegen Mendikaf afabarrusben 3 Dey D indegen Bay afabarrusben 4 Site name skring GCCL/SR naminal sampling site 5 Zona name indegen Reminal mangikan									
4 Sitz none string GCE LTER nominal sampling sitz 5 Zone none integer Nominal marsh zone			2	Month	M	Integer	Month of a	abservation	
5 Zone none integer Nominal marsh zone									
6 Plat need integer Permanent als sumber				Zone	none	Inleger	Nominal m	narsh sone	
7 Quedrat_Area m <sup>2</sup> finating-point Area in masurement of the quedrat frame									
5 Species_Code none skring Planispecies ode 9 Stoot telefit om fostingrande i telefit of telefit skoet			8	Species_Code		string	Plant spec	des code	

(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

Coastal Econ



## 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

Georg

- 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)