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
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-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

Sapelo (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
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
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
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

(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)