Data Ingestion and Publishing in Long-Term Ecological Research

Corinna Gries, John Porter
Data Sources

- Long-term Monitoring (technicians) (field, lab, analytical machine)
- Remote Sensing, GIS
- Sensors (streaming, wireless, downloaded)
- Short Term Research (grad student, single investigator)
- Images, Documents, Other Binary Data
- Processing and Analytical Code

LTER Data is Diverse and Comes From Many Sources
Data Ingestion

Long-Term Monitoring
- Scripts, workflows, direct entry forms
- Spreadsheets
- GCE Toolbox, R, python workflows

Short-Term Projects
- GIS, Remote Sensing
- LTER Site Information Management
- Local Metadata
- Images, Binary Data
- Code Scripts
- Publications Documents

Sensor Systems
- Filesystem of Archive files, Geoportal Database
- Version Control Systems
- Reference Management System
- Filesystem of Standard or Proprietary Formats
- Data archive
- Metadata

Publications Documents
- Pathways to
- Reference Management System
- Data archive
- Local Metadata
Data Ingestion Challenges

● **Short-Term Projects**
  o Mostly manual process
  o Many logical tables, few data
  o Extensive cleaning due to inconsistent use of codes etc.
  o Data collections methods difficult to capture

● **Long-Term Monitoring**
  o Changing protocols and equipment

● **Sensor Data**
  o Large volume - storage and quality control

● **Binary Data**
  o Maintaining usability of proprietary formats
Tools Developed by LTER

Many custom approaches (macros, shell scripts, statistical program scripts, custom programs, workflow systems, etc.)

GCE Toolbox built on Matlab - host of features, QA/QC and data documentation

DEIMS Drupal Ecological Information Management System
LTER NIS data repository

Provenance Aware Synthesis Tracking Architecture (PASTA)

- Metadata & Data QA/QC
- Versioning
- Audit services
- Persistent ID’s (DOIs)
- Service-Oriented Architecture
LTER NIS Data Portal
Automated generation of statistical programs for accessing and analyzing the data, for R, Matlab, SAS and SPSS
Data Publication Challenges

- Specialized archives (e.g., genetic data, ocean data, arctic data, hydrology data)
- Long-term sustainability in a changing funding/technology environment
- High quality metadata
Thank you to many people

Questions?