Data Ingestion and Publishing in Long-Term Ecological Research

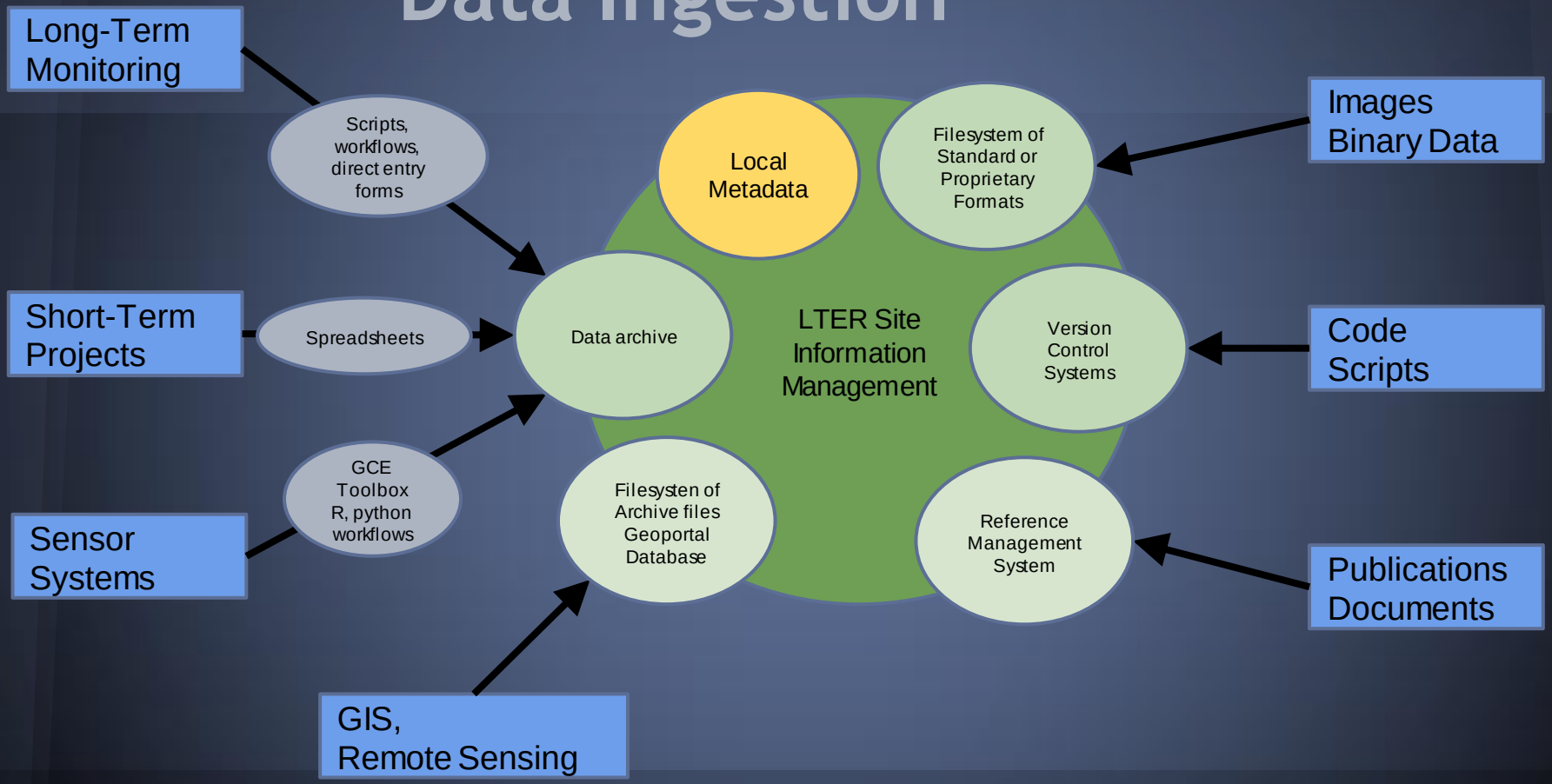
Corinna Gries, John Porter

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



Data Ingestion Challenges

- Short-Term Projects
  - Mostly manual process
  - Many logical tables, few data
  - Extensive cleaning due to inconsistent use of codes etc.
  - Data collections methods difficult to capture
- Long-Term Monitoring
  - Changing protocols and equipment
- Sensor Data
  - Large volume – storage and quality control
- Binary Data
  - Maintaining usability of proprietary formats

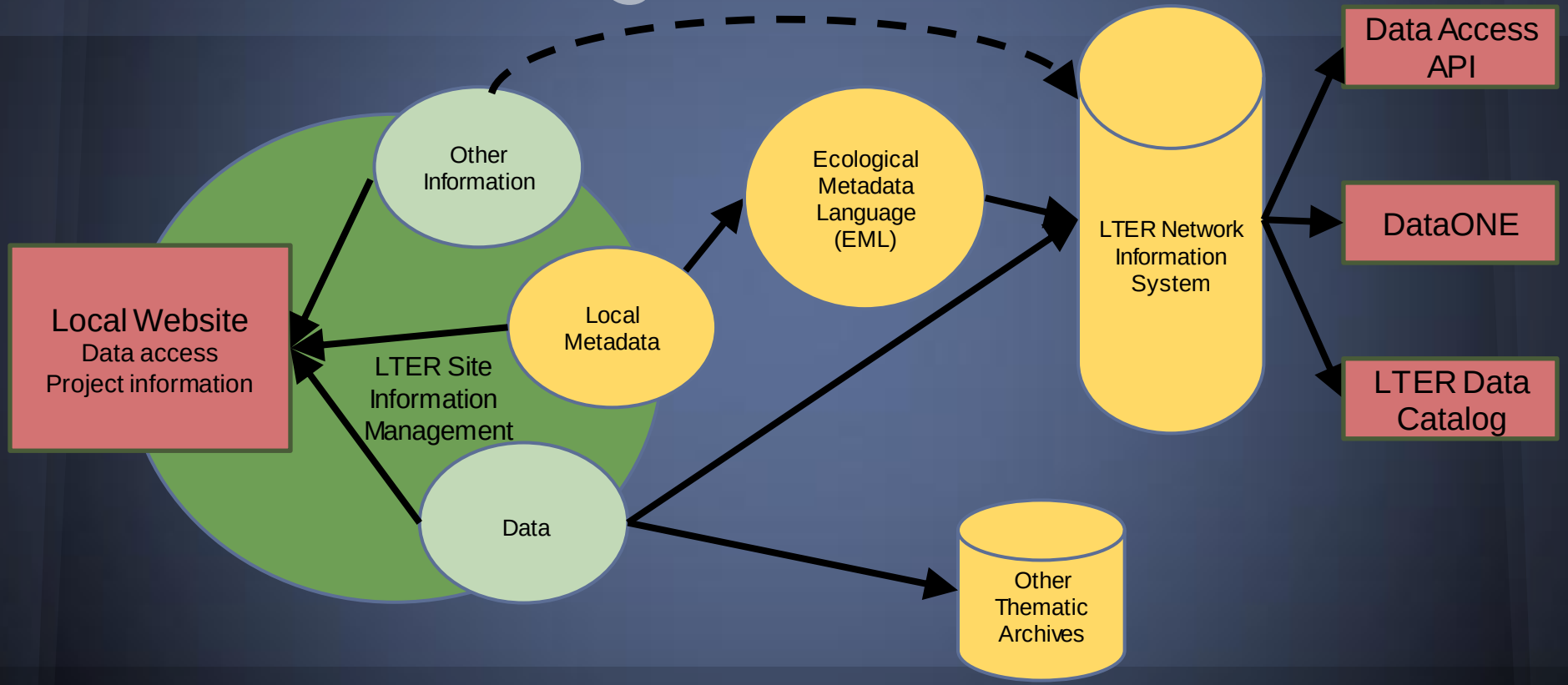
Tools Developed by LTER

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

Data Publishing

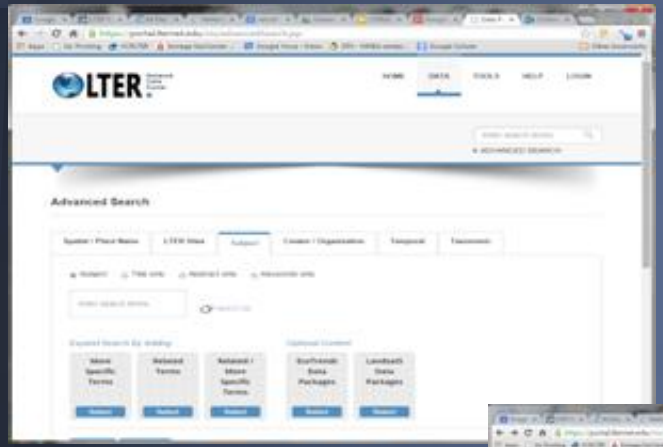


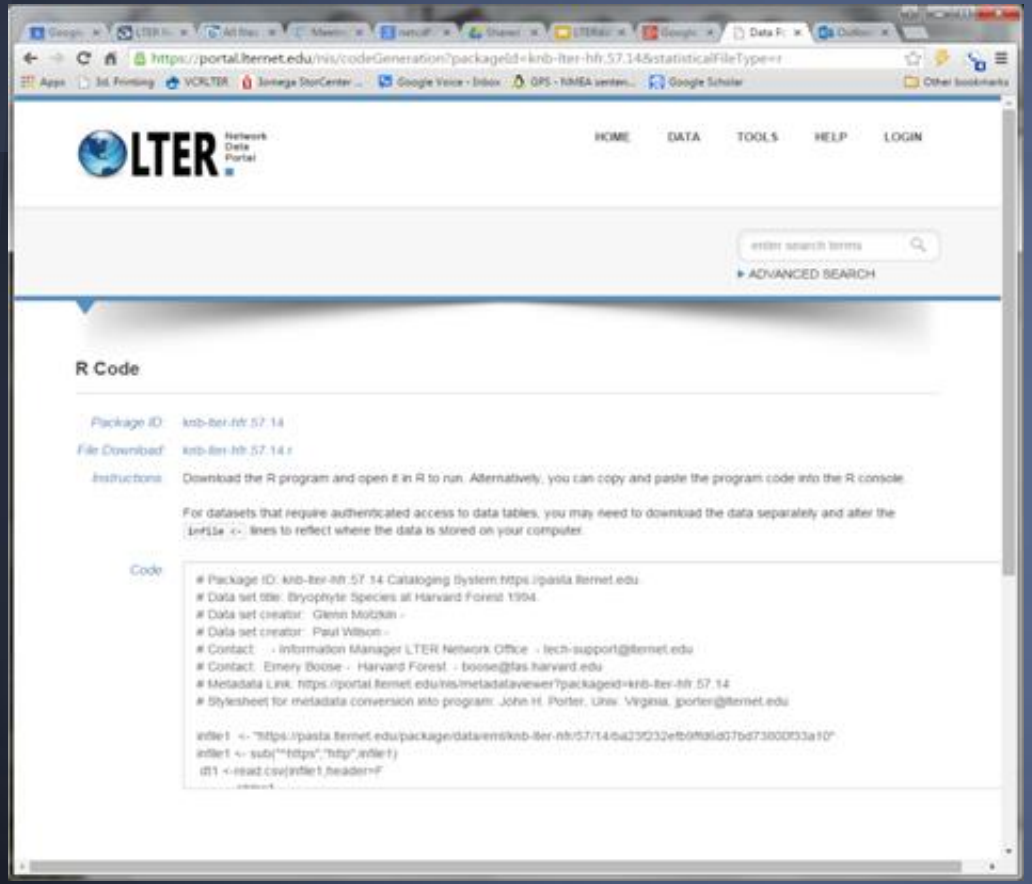
LTER NIS data repository

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 ?