

Data Ingestion and Publishing in Long-Term Ecological Research

Corinna Gries, John Porter

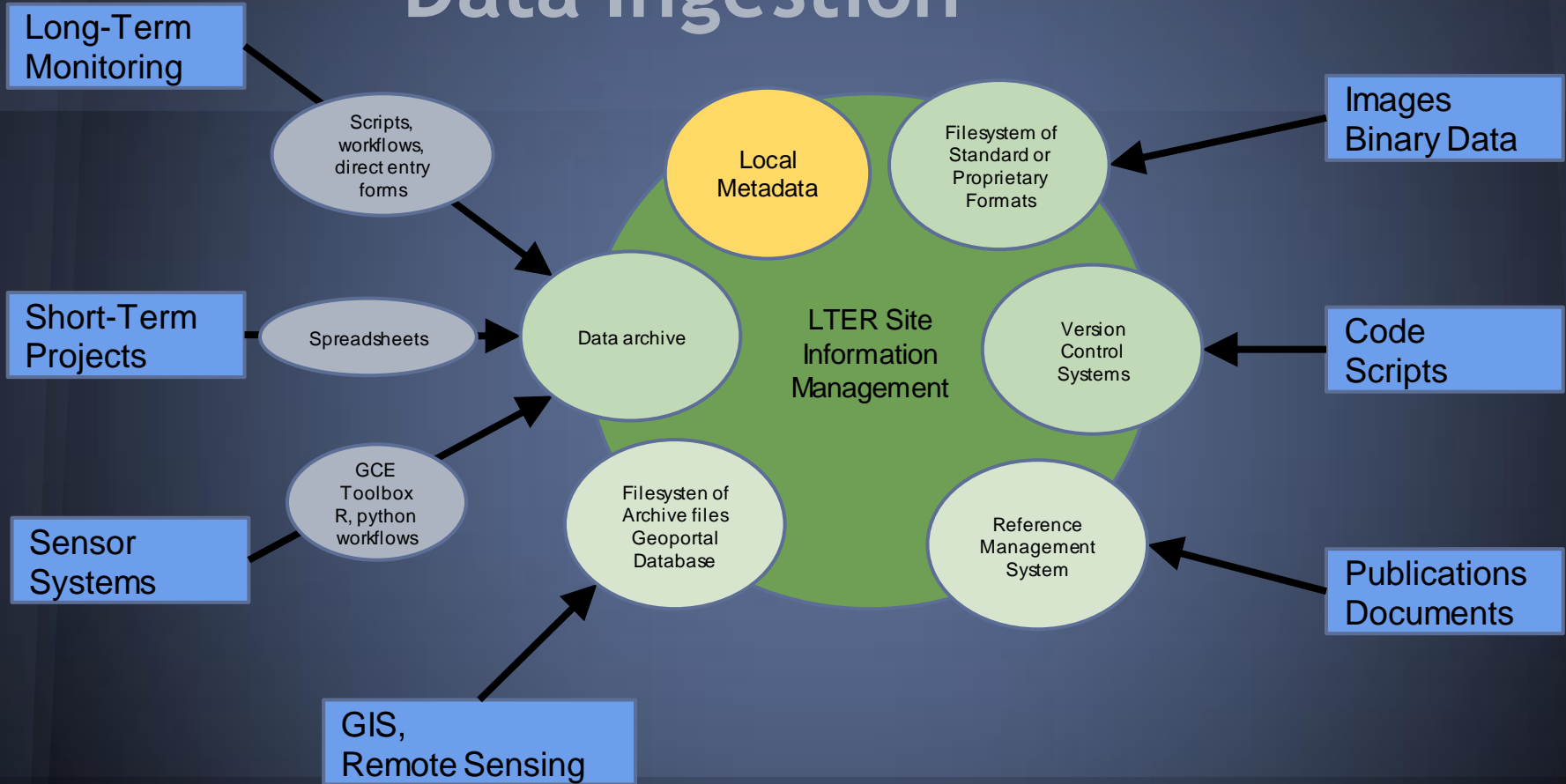


Data Sources

- 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

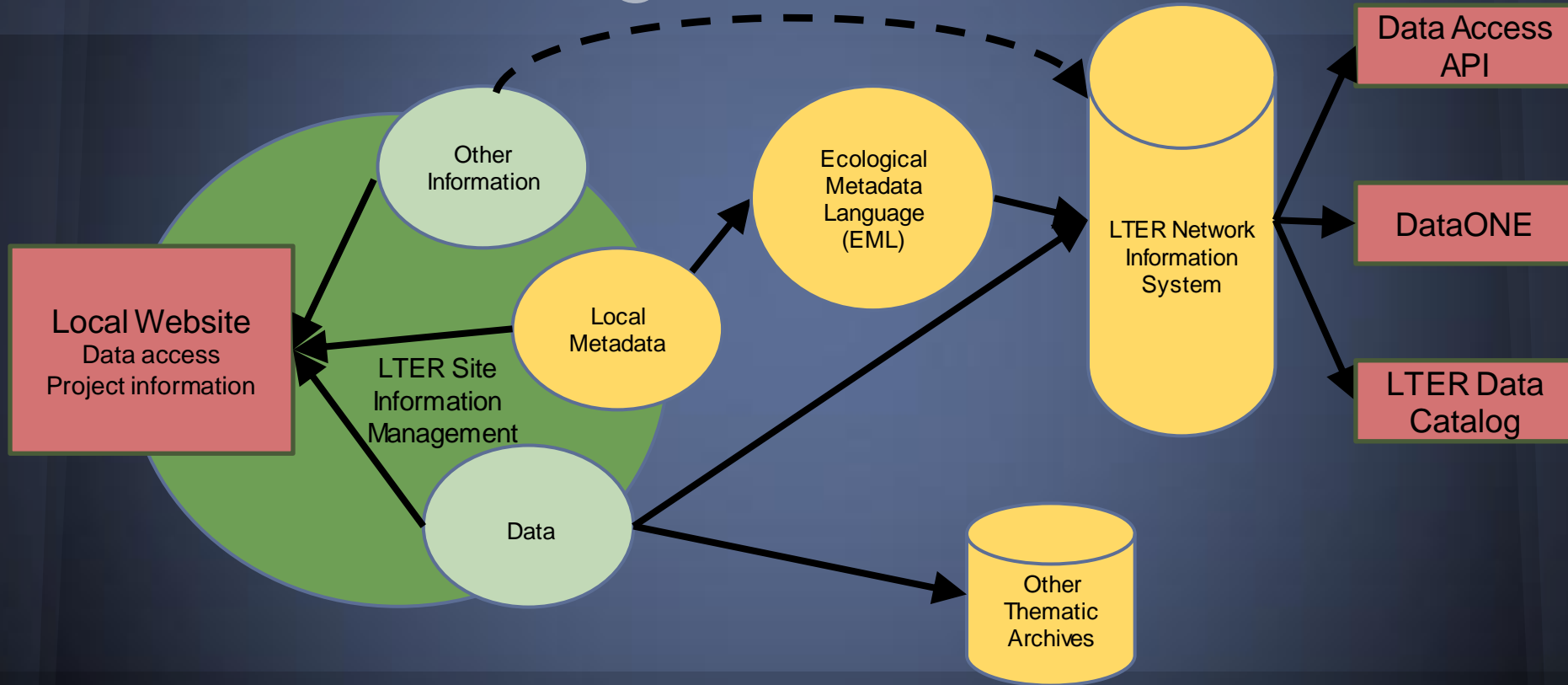
Tools Developed by LTER

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



The screenshot shows a web browser window displaying the LTER Network Data Portal. The URL is <https://portal.lternet.edu/nis/codeGeneration?packageid=knb-lter-hf-57.14&statisticalFileType=r>. The page features a navigation menu with links for HOME, DATA, TOOLS, HELP, and LOGIN. A search bar is located below the menu. The main content area is titled "R Code" and includes the following information:

- Package ID:** knb-lter-hf-57.14
- File Download:** knb-lter-hf-57.14.r
- Instructions:** Download the R program and open it in R to run. Alternatively, you can copy and paste the program code into the R console.

For datasets that require authenticated access to data tables, you may need to download the data separately and alter the `url1` <- lines to reflect where the data is stored on your computer.

Code

```
# Package ID: knb-lter-hf-57.14 Cataloging System: https://pasta.lternet.edu
# Data set title: Bryophyte Species at Harvard Forest 1994
# Data set creator: Glenn Moulton
# Data set creator: Paul Wilson
# Contact: Information Manager LTER Network Office - lter-support@lternet.edu
# Contact: Emery Boose - Harvard Forest - boose@fas.harvard.edu
# Metadata Link: https://portal.lternet.edu/nis/metadataviewer?packageid=knb-lter-hf-57.14
# Stylesheet for metadata conversion into program: John H. Porter, Univ. Virginia, jporter@lternet.edu

url1 <- "https://pasta.lternet.edu/package/data/em/knb-lter-hf-57.14/ba29232eb9ffad07bd7300033a10"
url1 <- sub("https","http",url1)
dfl <- read.csv(url1,header=F)
```

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 ?