Data Curation and Distribution in Support of Cornell University’s Agricultural Ecosystems Program

Gail Steinhart
Research Data & Environmental Sciences Librarian

Brian Lowe
Metadata Programmer

Albert R. Mann Library, Cornell University

DigCCurr - April 19, 2007
Overview

- Motivation
- Strategy
- What we’ve learned so far
Motivation

from Michener et al., 1997
Motivation
Definitions

Curation
The term digital curation is used ... for the actions needed to maintain digital research data and other digital materials over their entire life-cycle and over time for current and future generations of users. Implicit in this definition are the processes of digital archiving and preservation but it also includes all the processes needed for good data creation and management, and the capacity to add value to data to generate new sources of information and knowledge.

- from DCC

We
Academic and research libraries, but not alone
Scientific context

Chesapeake Bay watershed

- Largest US estuary
- Critical fishery, habitat
- Sensitive to nutrient pollution
- Chesapeake Bay Agreement of 2000

Upper Susquehanna River Basin

- Susquehanna is largest US river draining to the Atlantic; largest trib to the bay
- MOU with Chesapeake Bay Program commits NYS to water quality goals of Chesapeake Bay Agreement
Collaborators

Cornell departments and units:
• Animal Science
• Biological and Environmental Engineering
• Crop and Soil Science
• Ecology and Evolutionary Biology
• Horticulture
• Natural Resources
• Mann Library

Other organizations:
• Cornell Cooperative Extension of Chemung County
• Institute of Ecosystem Studies
• Univ. Maryland Center for Environmental Science
• Univ. Nebraska-Lincoln School of Natural Resources
• Upper Susquehanna Coalition

Funding: USDA Cooperative State Research, Education, and Extension Service
Types of data

Observational:
• Atmospheric deposition of N
• Water and soil chemistry
• Hydrologic measurements
• Meteorological data
• Plant tissue chemistry
• Cs-137 in stream sediments

Experimental:
• Effects of willow char amendments to agricultural soils
• Wetland plant species responses to changes in S and P cycling
• Changes in ground water chemistry as a result of chemical amendments
• N leaching in soils under different cropping systems and snow cover manipulations
• Changes in forest and old field chemistry as a result of N fertilization

Simulation models of nutrient and sediment fluxes
Strategy

- Local support for data and metadata preparation
- Local and/or discipline-based “publication” of data, metadata
Strategy

- Local support for data and metadata preparation
- Local and/or discipline-based “publication” of data, metadata
Strategy

- Local support for data and metadata preparation
- Local and/or discipline-based “publication” of data, metadata
“Staging repository”

• Use discipline-specific metadata standards and tools

   >> **Ecological Metadata Language (EML)**
   >> **Morpho**

• Provide a place to share pre-publication data within the group

   >> **Metacat**

• Provide training and recommendations on data and metadata preparation
Ecological Metadata Language: EML

- Developed specifically for ecological data (NCEAS, LTER)
- Modular and extensible XML-based standard
- Accommodates information on methods, geographic coverage, temporal coverage, detailed descriptions of tabular data
- [http://knb.ecoinformatics.org/software/eml/](http://knb.ecoinformatics.org/software/eml/)
- Comes with tools!
Morpho

- Easy to use, platform independent metadata editor.
- Interacts with Metacat: allows users to upload metadata and data; allows users to search, view, and export public data and metadata.
EML record
“Publication” of data

- Deposit in institutional repository
  
  **DSpace**
  
- Submit metadata (and possibly data) to discipline-specific repository
  
  **KNB, other?**
  
- Link from project web portal: [http://www.usaep.mannlib.cornell.edu/](http://www.usaep.mannlib.cornell.edu/)
Test case: Historical data

• Observational data from last 30 years

• Original format: Quattro Pro workbooks with multiple pages

• Various errors (apparent duplicate records, misaligned columns, out of range values)

• Missing or ambiguous information (methods, units, geographic locations)

• Extensible model?
Summary – curation skills

• “Traditional” library and archiving skills (metadata, preservation, interoperability, appraisal and selection)

• Understanding of CI

• Subject area knowledge:
  o Understanding of research practices, tools, and culture (may be discipline-specific)
  o Awareness of standards and tools related to data

• Productive partnerships with researchers (or ability to develop them)
Thank you

Gail Steinhart
Research Data & Environmental Sciences Librarian
Albert R. Mann Library, Cornell University

GSS1@cornell.edu