Collaborate, Automate, Prepare, Prioritize

Creating Metadata for Legacy Research Data

Inna Kouper
Data to Insight Center, IU

Stacy Konkiel
IU Libraries

Jennifer A. Liss
IU Libraries

Juliet L. Hardesty
IU Libraries
Data Management as “Grand Challenge” & Metadata
SEAD is funded by the National Science Foundation under Cooperative Agreement #OCI0940824
SEAD Virtual Archive (SVA)
-- manage sustainability science window to multiple IRs

Active Curation Repository (ACR)
-- metadata harvest
-- annotation
-- web tools

SEAD VIVO
-- social networking
-- links data sets and researchers

IU Scholar Works IR
UIUC IDEALS IR
UMich Deep Blue IR
Investigation

- How can the curation of legacy data be improved by supplying necessary metadata?
- How much time and effort is required to supply domain-specific metadata?
Goals

• Enable discovery of research data
• Communicate experiences with metadata creation for legacy dataset to community
• Begin conversation about metadata practices for legacy data
Methodology

- 20 NCED legacy datasets
- Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata
Methodology

• 4 encoders, each assigned 5 datasets
• Datasets ranged greatly in size and composition
  • 0.01–664 GB
  • 1–140,000 files
Methodology

• Phase I
  Standalone XML files using basic NCED-provided information & “Googleable” facts

• Phase II
  Extensive research re: processes by which datasets were created and used
Findings–Phase I

Metadata creation time Phase I (h:mm)

- Dataset 1
- Dataset 2
- Dataset 3
- Dataset 4
- Dataset 5

- Encoder 1
- Encoder 2
- Encoder 3
- Encoder 4
Findings—Phase I

Successes:

• Supplied many mandatory elements
• Thesauri & Controlled Vocabularies

Challenges:

• Time-intensive startup
• Lacking geospatial information
### General Information

**Owner:** Leslie Hau  
**Abstract:** A 4-meter diameter, 80-cm wide rotating debris flow flume was constructed at the University of California Richmond Field Station for studying large-scale granular flow phenomena. This dataset covers the experiments conducted in 2007 and 2008, where the primary goal was to study rates and mechanisms of bedrock erosion by debris flows. The following data will be posted at the NCED Repository. (Please see the temporary folder for now: https://docs.google.com/folderview?id=0BwYXkKpsltiROHsGRUzjyyGh/edit ) 1. Master list of experiments 2. Force plate data 3. Laser profile data 4. Erosion topography data 5. Videos 6. MATLAB scripts

**Purpose:** scientific research  
**Status:** In work  
**Update frequency:** Unknown

**Access Constraints:** None  
**Usage Constraints:** None

### Publishing Details

**When:** Date: 20070101  
**Published By:** National Center for Earth-surface Dynamics Data Repository  
**MN USA**  
**Available At:** http://repository.nced.umn.edu/browser.php?dataset_id=27

### Keyword Terms

**Subject Keyword Terms**

- **Defined by:** ISO 19115 Topic Categories  
  - geosciencesinformation

- **Defined by:** NASA Global Change Master Directory (GCMD) Science Keywords  
  - EARTH SCIENCE  
  - EROSION  
  - EROSION/SEDIMENTATION  
  - LAND SURFACE

- **Defined by:** None  
  - Bedrock erosion
Findings—Phase II

Successes:

• Enhanced 10 metadata fields

Challenges:

• Accessing and processing the datasets (size, complexity)
Observations

Though the information that we found may enhance opportunities for the discovery of legacy research data, the available information was unlikely to be sufficient to support the tasks of preservation, reproducibility, and re-use.
Observations

• FGDC is insufficient for dealing with legacy research data

• Data curators without domain expertise can be successful in creating some types of metadata

• Structural and administrative metadata is difficult to curate without help of researchers
Proposal: The CAPP Framework

- Subject specialists
- Librarians
- Researchers
- Tool developers

- Choice of metadata standards
- Instructions / manuals
- Workflows / software
- Licensing and contact information

- File format identification
- Provenance
- Native environment
- Entity extraction

- Labor
- Datasets
- Types of metadata
- User needs
Collaborate

- Subject specialists
- Librarians
- Researchers
- Tool developers
Automate

- File format identification
- Provenance
- Native environment
- Entity extraction
Prepare

- Choice of metadata standards
- Licensing and contact information
- Instructions and manuals
- Workflows and software
Prioritize

- Labor
- Datasets
- Types of metadata
- User needs
Future Work

Benchmark:

• Effectiveness of tools and workflows
• Collaborations and relationships
• Domains/interdisciplinarity
Thank you!

Jennifer A. Liss
Metadata/Cataloging Librarian
jaliss@indiana.edu

http://sead-data.net

Download this presentation:
slideshare.net/jenniferaliss/liss-dcmi-2013