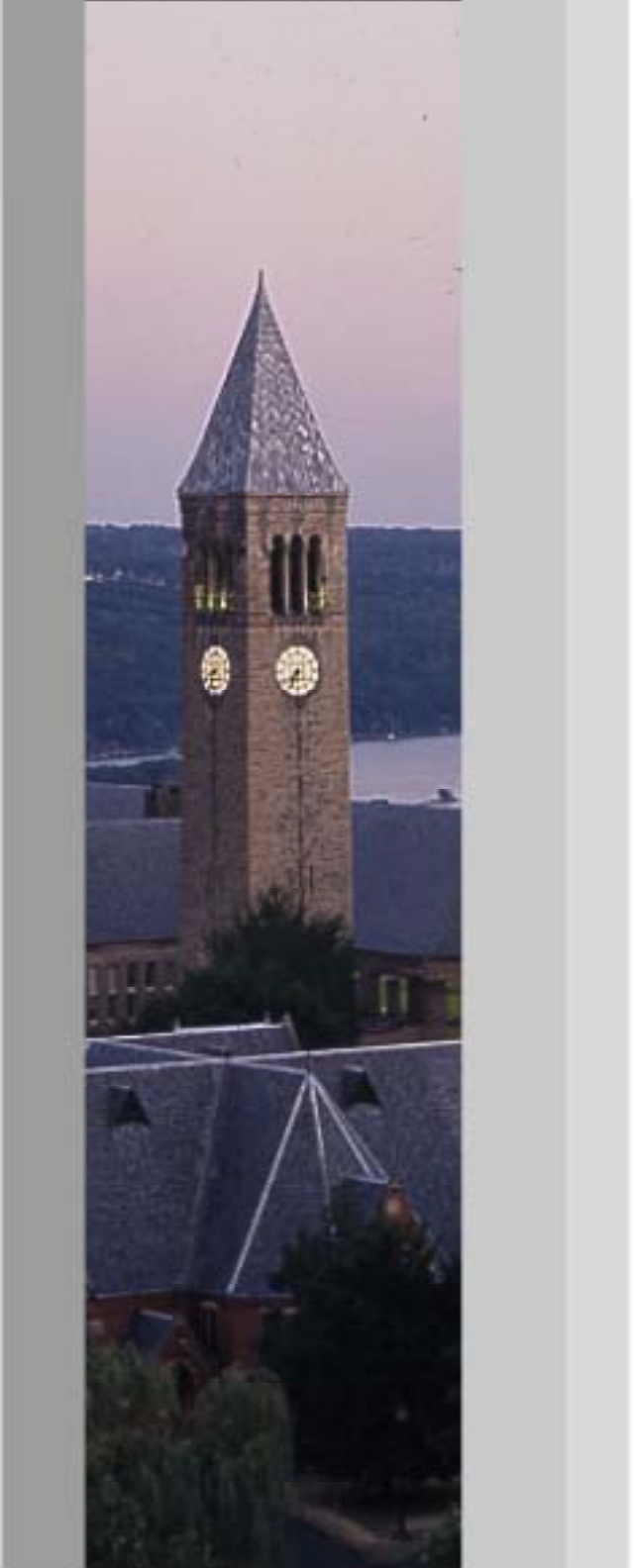

Cornell University Library

Data Curation and Metadata Management: Problem and Promise

Elaine L. Westbrooks
LTF7
May 2, 2008

OVERVIEW

1. Data Curation Defined
2. Description
3. Goals
4. Outcomes
5. Challenges
6. Conclusion

DATA CURATION DEFINED

- **The activity of managing and promoting of the use of data from its point of creation, to ensure it is fit for contemporary purposes and available for discovery and reuse**
  - » Philip Lord, Alison MacDonald, 2003
- **The active & ongoing management of data through its lifecycle of interest & usefulness to scholarship, science, & education….**
  - » Lorcan Dempsey, Blog Posting, February 26, 2007
- **A series of technical, intellectual, and managerial activities in support of stewardship for digital…information assets…**
  - » Oya Rieger, *D-Lb*, July/August 2007
- **The process of maintaining and adding value to a trusted body of digital information for current and future use**
  - » Jane Greenberg, 2007

BLACKOUT ARCHIVE DESCRIPTION

1. Centralized data repository
2. Four levels of data access/use rights
3. Only 300 GB of encrypted data
4. North American Electric Reliability Corp. owns data
5. Current delivery platform: DSpace
6. Next delivery platform: FEDORA (possibly)
7. Stakeholders: Researchers and Power experts
8. Small Science
9. 146 data formats

http://metadata.library.cornell.edu/blackout/

BLACKOUT ARCHIVE GOALS

1. Provide secure online access to data
2. Facilitate policy creation
3. Add value to data
4. Build tools for data users and providers
5. Sustainability

BLACKOUT ARCHIVE: ACCESS

Provide secure online access to data

1. Data present a homeland security threat
2. Data must be secure
3. Data must be easily accessible online
4. Experts must be able to submit data easily

BLACKOUT ARCHIVE: ACCESS LEVELS

Levels of Access for data:

1. Unrestricted data
2. Data restricted to personnel from electricity industry
3. Data restricted to blackout investigators
4. Data restricted by data supplier agreement

BLACKOUT ARCHIVE: POLICY

Facilitate policy creation

1. Archive policies
   a. Collection development
   b. Data curation
   c. Intellectual property
   d. Access and use rights
2. Data contributor policy
3. Citation guidelines
4. Authorization forms

BLACKOUT ARCHIVE: ADDING VALUE

1. Data are organized in one place
2. Metadata is created
3. Expedite the authorization process
4. Maintain quality
5. Increase data's potential for reuse/repurpose
6. Data are archived

BLACKOUT ARCHIVE: TOOLS

1. Extract metadata from our most common files; pdf
2. Freely available Software
3. Users who want to submit a pdf file can click on the file and our tool extracts the metadata in seconds.
4. Problem is that we have over 100 formats

BLACKOUT ARCHIVE: SUSTAINABILITY

1. Economic
   - -What models exist that can be put to use?
   - -What is the role of the library, NSF, users?
2. Technical
   - -Reliable and long-term data preservation
   - -Continuously adapt new technologies
   - -Considering user expectations

BLACKOUT ARCHIVE: OUTCOMES

1. Help stakeholders determine policy
2. Organize data
3. Create metadata
4. Tools to facilitate metadata creation
5. Build relationships with potential data providers
    1. Pacific Northwest National Laboratory
    2. NERC
    3. Power Engineering Faculty

BLACKOUT ARCHIVE: CHALLENGES

1. Sustainability
2. Dedicated Personnel
3. Rights management
4. Assessment
5. Metadata quality

BLACKOUT ARCHIVE: METADATA

- Create metadata where none existed
- Metadata creation must be quick and easy
- Extract metadata from folder structure
- Extract metadata from file headers
    - -Works great for tiffs, pdf, docs, emails
    - -Works poorly with other types of non-textual data

METADATA MANAGEMENT

- Metadata is in xml
- In some cases there is 1:1 relationship to data
- In most cases 1 metadata record is a surrogate for dozens or hundreds of datasets
- Store metadata separately from data
- Extract metadata from files without human review
- File structure of the data is useful metadata
- Metadata creation takes place throughout the process; we don’t try to catch everything at the beginning

CONCLUSION

- Sustainability will always be a challenge
- Metadata management is still elusive
- Libraries provide expert services regarding policy
- Collaboration with experts is critical
- Assessment is critical
- Management of Intellectual Property, security, and privacy are key