Archiving Digital Information:
An Introduction

Linda Zellmer
Head, Geology Library
Indiana University
lzellmer@indiana.edu

Introduction

- Librarian's view on Digital Archiving.
- Need for digital archiving.
- What Earth Science Information should be archived?
- Digital Archiving programs:
  - LOCKSS, CLOCKSS & Portico.
  - Web Archiving (Wayback & ArchiveIt).
  - NDIIPP.
  - Institutional Repositories.

Why is Digital Archiving important?

Fire Damage

Flood Damage, University of Hawaii, October, 2004

Water Damage

"We are living in the golden age of dead media . . . most of them with the working lifespan of a pack of twinkies."

Sterling, 1995

"It is only slightly facetious to say that digital information lasts forever, or five years, whichever comes first."

Rothenberg, 1995

Why Digital Archiving?

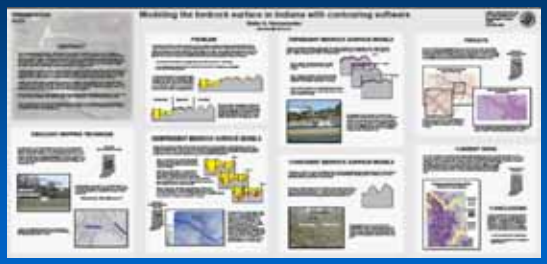
- Disasters happen.
- Storage media changes.
- URLs change.
- Digital information stored on computers and servers may be needed by future users.

What Digital Earth Science Information should be archived?

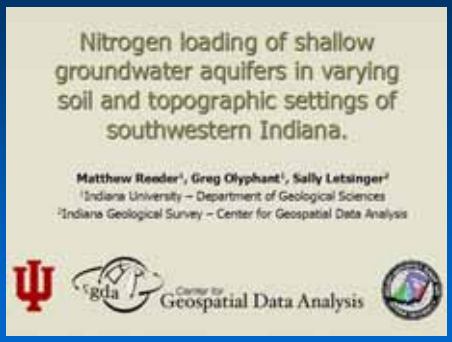
E-Journals

Web Sites

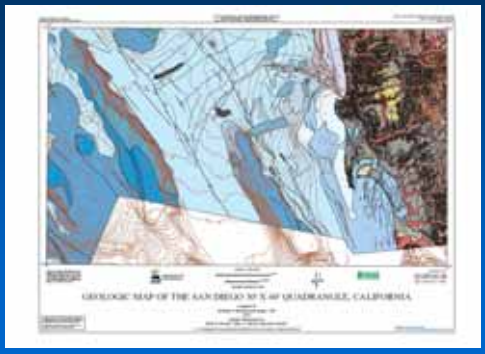
Posters from Meetings



Meeting Presentations



Geospatial Information



Theses & Dissertations

Electronic Journal Archiving

LOCKSS, CLOCKSS & Portico

LOCKSS & CLOCKSS

- E-Journal archiving systems.
- LOCKSS = Lots Of Copies Keeps Stuff Safe.
- CLOCKSS = Controlled LOCKSS.
- Partnership between publishers & libraries to develop a comprehensive, distributed archive to preserve and ensure continuing access to electronic content.

LOCKSS & CLOCKSS

- E-Journal content is collected using a web crawler & stored on PCs in multiple locations.
- Content is regularly compared among LOCKSS locations to identify problems.
- Repaired if needed.
- Content is migrated as format changes.

LOCKSS & CLOCKSS

- When publisher's content is not available, content from LOCKSS computer is made available.
- Participating publishers include Blackwell, Elsevier, Nature, Oxford University Press, Springer, Taylor and Francis & John Wiley & Sons.

Portico

- Fee-based permanent archive of electronic scholarly journals.
- Affiliated with JSTOR.
- Publishers pay an annual fee based on revenue received from journal publishing.
- Libraries pay an annual fee based on their total materials budget.

Portico

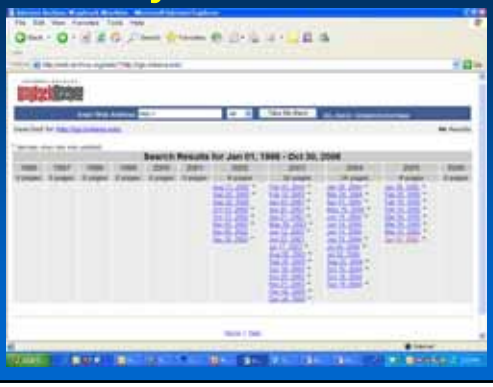
- Archive available if publisher discontinues publication of a title, ceases operations, no longer offers back issues or experiences a catastrophic long term failure of their delivery system.
- Participating publishers include: Annual Reviews, BioOne, Elsevier, Oxford University Press, Taylor & Francis and University of Chicago Press.

Web Site Archiving

Internet Archive

- http://www.archive.org
- A digital library of Web sites developed to provide permanent access to digital collections.
- Contains text, audio, video, software and web pages.
- Wayback Machine can be used to search for older archived content.

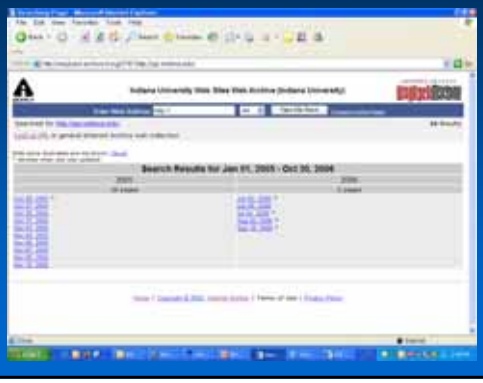
Wayback Machine



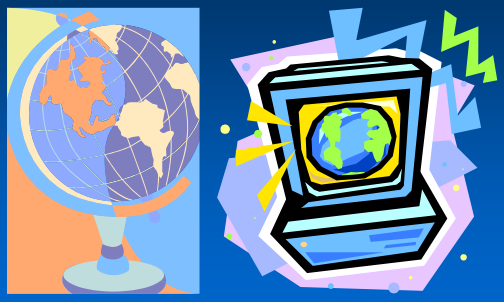
Archive-It

- Subscription service that can be used by organizations to build an archive of their web site ($10,000 per organization).
- Subscribers can identify which web content they want to archive.
- Information available as a separate collection; not searchable in Wayback Machine.

Archive-It



Geospatial Data Archiving



Spatial data has the lifespan of a Mayfly, which hatches, breeds & dies within 24 hours.

Zellmer, 2003

Digital Spatial Archiving

- Some history:
  - Historical Census Tract Boundary data has been recreated by the University of Minnesota's Population Center with a $5,000,000 NSF grant.
  - Some early Landsat Data is no longer available because NASA forgot to refresh the tapes.
  - GIS data from a federally funded study of the Snake River in Wyoming was lost when a GIS workstation was replaced.

Digital Spatial Archiving

- Digital spatial data has been lost over time.
- SDTS no longer recommended format for archiving spatial data.
- Historical Data Working Group of FGDC is working with Open Geospatial Consortium to develop archival standards for GIS Data.
- Format needs to preserve topology & data relationships.

Digital Spatial Archiving

- North Carolina Geospatial Data Archiving project.
- Funded by National Digital Information Infrastructure and Preservation Program of the Library of Congress.
- Researching approaches to digital spatial data archiving.
- Final results not yet available.

Institutional Repositories

Institutional Repositories

- Digital collections that capture and preserve the intellectual output of a single institution, such as a university or a group of users.
- IR information is openly accessible to all.
- IR Information can be found using search engines such as Google.
- Collections will be migrated as formats change.

Institutional Repository Collections

- **Articles & preprints**
- **Technical reports**
- **Working papers**
- **Conference papers (posters, PowerPoints)**
- **Electronic theses & dissertations**
- **Datasets: statistical, geospatial, matlab, etc.**
- **Images: visual, scientific, etc.**
- **Audio files**
- **Video files**
- **Learning objects**
- **Reformatted digital library collections**

Institutional Repositories

- Publishers attitudes towards institutional repositories differ.
- Most earth science publishers (AGU, AAPG, & GSA) do not allow materials to be placed in Institutional Repositories.
- Major commercial publishers allow posting of article preprints in Institutional Repositories.

Institutional Repositories

Institutional Repositories

Institutional Repositories

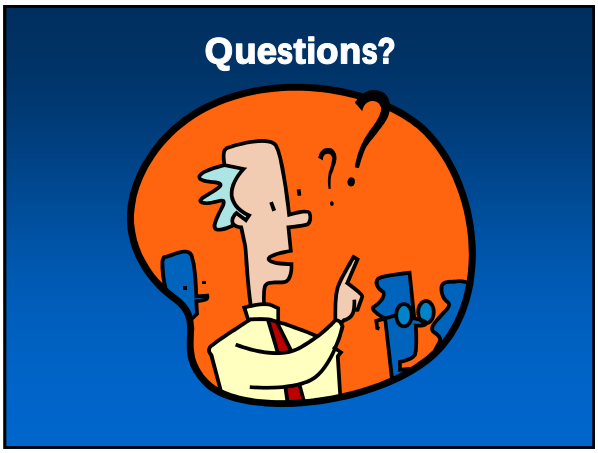
Role of Institutional Repositories

- Collect information developed in the organization, institution or community.
- Provide open access to information in multiple formats.
- Preserve information for future users.

"It is only slightly facetious to say that digital information lasts forever, or five years, whichever comes first."

Rothenberg, 1995

The End


Questions?