Day-to-Day Digital Preservation: A Case Study

GSLIS Data Curation Institute
June 4, 2008

Tim Donohue
IDEALS Technical Lead

Illinois Digital Environment for Access to Learning and Scholarship

Outline

- Intro to IDEALS
- How we got started...
- Our preservation strategies
- Current / future data curation work
- Next steps...

What is IDEALS?

Digital repository for the scholarship and research of the faculty, students, and staff of the University of Illinois at Urbana-Champaign.

- Dissemination
- Persistent Access
- Preservation

http://ideals.uiuc.edu/

A joint initiative between the University Library and CITES with support from the Office of the Provost.

What type of materials?

Kenneth M. Cuno

THE ORIGINS OF PRIVATE OWNERSHIP
OF LAND IN EGYPT: A REAPPRAISAL

Long-term vole demographic data files

Lowell Getz - Department of Animal Biology - University of Illinois at Urbana Champaign - May 2004 - Vitae

Data files of

1. 25 years of monthly live-trapping of Microtus ochrogaster and M. pennsylvanicus in alfalfa, bluegrass, and tallgrass prairie,
2. 63 months of twice-weekly 2-day live-trapping of M. ochrogaster in alfafa.

Explanation

Collaborative Research in Semiconductors

September 6, 2007

Larry W. Sumney
President and CEO
Larry.Sumney@src.org

idential Material

LIBRARY TRENDS

Winter 2004 52(3) 373–670

addition, there are fi
rina brevicauda, for e

ACDIS
Occasional
Paper

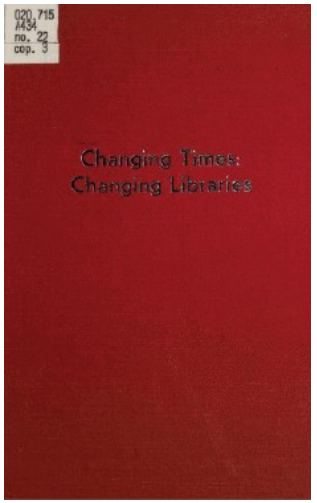
Beyond Precision: Issues of Morality and Decision Making in Minimizing Collateral Casualties

Lt. Col. Dwight A. Rathjen
National Defense Fellow, United States Air Force
Program in Arms Control, Disarmament, and International Security
University of Illinois at Urbana-Champaign

The Philosophy of Information

Ken Herold

Issue Editor



Also audio and video

IDEALS goals….

- Help increase access to published and unpublished research
- Help increase the impact to published and unpublished research
- Provide a persistent, permanent URL for citing research
- Preserve research for long term access and use

The IDEALS "service model"

- Not *just* a repository…
- *Set of services* to collect, disseminate, and provides persistent and reliable access to the research and scholarship of faculty, staff, and students at the University of Illinois at Urbana-Champaign.

Other services we offer…

- Consultation on copyright issues
- Access restrictions / Embargo of items
- Statistics on number of downloads (working on departmental monthly reports…)
- Pilot service to deposit research into PubMed Central and ‘disciplinary’ repositories

IDEALS: the beginning

In the beginning: Promises proceed us

- Can we really commit to preserving everything?
- What does it really mean to preserve this stuff?
- What kind of staff expertise do we need?
- What kind of resources do we need?
- What kind of technical infrastructure do we need?

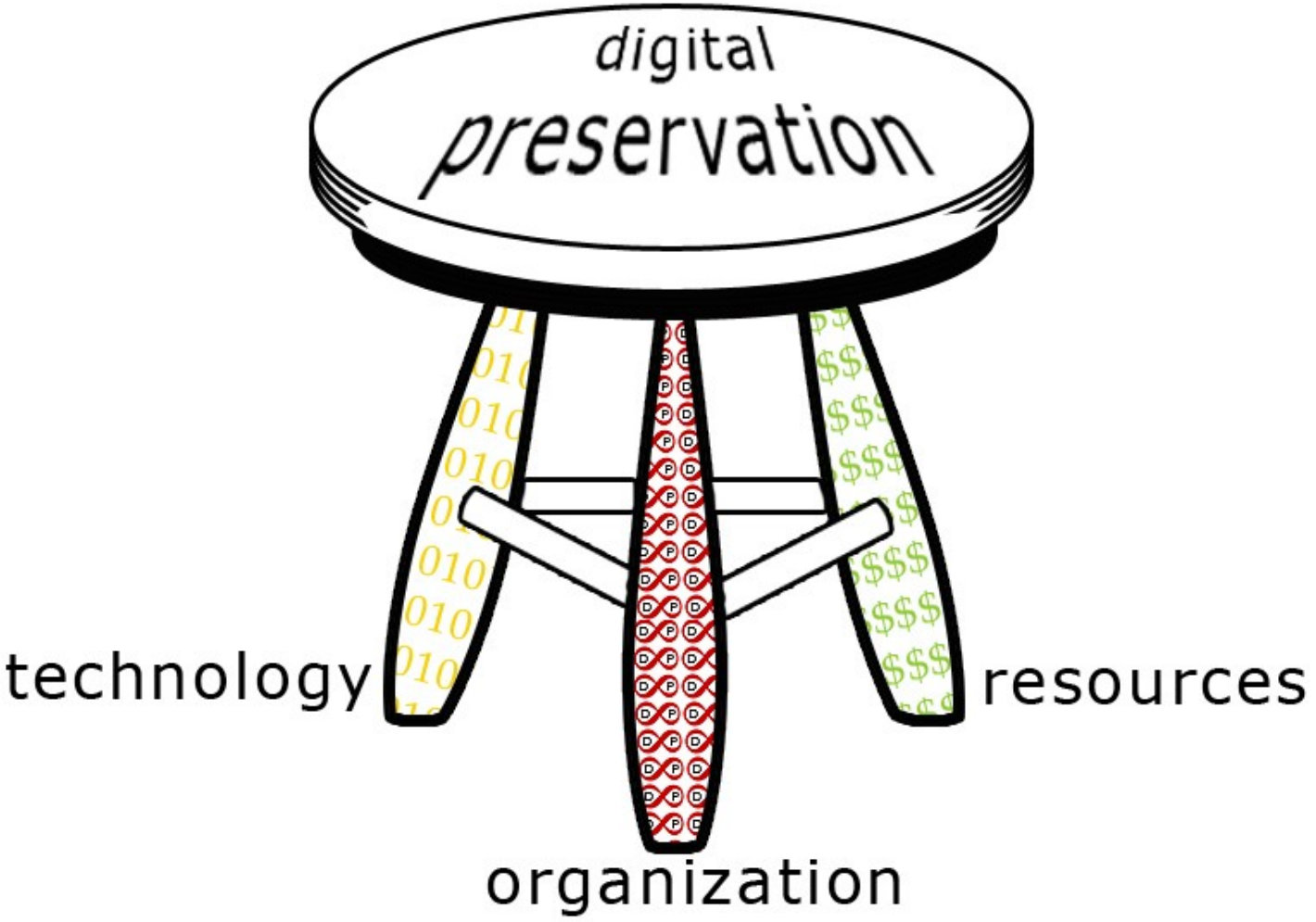
Getting our act together

- Got our Preservation Librarian involved

- Training and self education
  - Cornell's Digital Preservation Management Workshop and Online Tutorial

    http://www.icpsr.umich.edu/dpm/
  - Understanding Open Archival Information System (OAIS) conceptual model
  - Trustworthy Repositories Audit & Certification (TRAC)

The Digital Preservation Platform



Borrowed from the ICPSR Digital Preservation Tutorial:
http://www.icpsr.umich.edu/dpm/

Preservation Takeaways:

- Be explicit about what you will do and what you won't do.
- You don't have to preserve everything if you say you aren't.
- Digital preservation management is not just about the technology.

Photo borrowed from:
http://flickr.com/photos/santos/

Getting our act together, cont.

Backup tapes stored next to the server!

Not Really Our Server Room!

Photo borrowed from: http://www.flickr.com/photos/sylvar/

Looking forward to production: Digital Preservation White Paper

http://hdl.handle.net/2142/135

- Laid out for the Library and CITES administration what supporting a digital preservation management program would mean:
  - **Commitment on the part of both organizations**
  - **Resources** in terms of funding and staff are specifically allocated
  - Processes, policies, and the institutional commitment are **documented** and **as transparent as possible**.
  - The technical infrastructure is developed using **community standards**.
  - Commitment of resources for **planning** and community standards building.

IDEALS Preservation Policy: Operating Principles

Adherence To:

- OAIS Reference Model
- Community standards for preserving digital content
- Hardware, software, and storage media best practices.
- Intellectual property, copyright, and ownership rights of all content

Commitment To:

- Interoperable, scalable digital archive
- Clear, openly documented policies & procedures
- Archival requirements for provenance, custody, authenticity, integrity

Goal: A Certified, Trustworthy Repository

What resources do we need?

- Funding
  - Currently from the Office of the Provost
- Designated staff
  - Built into our job descriptions

Technology infrastructure

- Move from Library to CITES
  - Better environment
  - Better security
  - Distributes support for the tech infrastructure

Risks and Challenges

- Technological Change
- Sustainability
- Partnership between the University Library and CITES
- Identifying an Exit Strategy

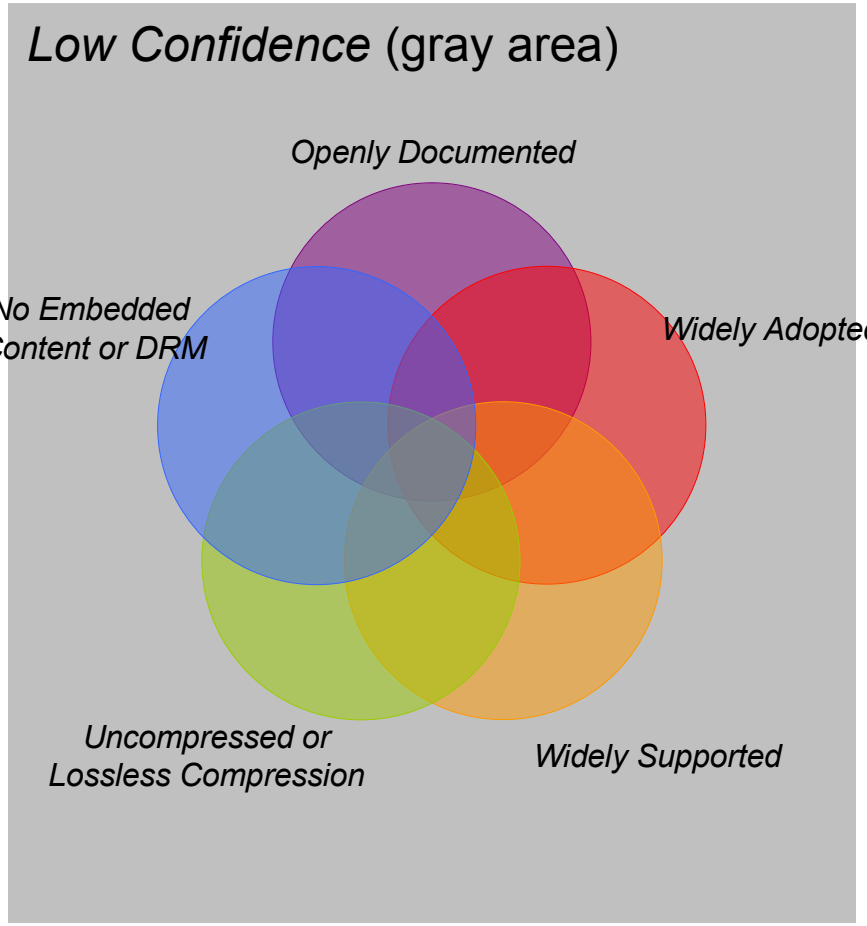
How IDEALS supports data (files)

What have others done?

- Michigan's Deep Blue – format support policy
  - http://deepblue.lib.umich.edu/about/
- Florida Digital Archive – format "action plans"
  - http://www.fcla.edu/digitalArchive/
- LC: Sustainability of Digital Formats
  - http://www.digitalpreservation.gov/formats/
- Australian Partnership for Sustainable Repositories (APSR)
  - http://www.apsr.edu.au/

Digital Preservation Support

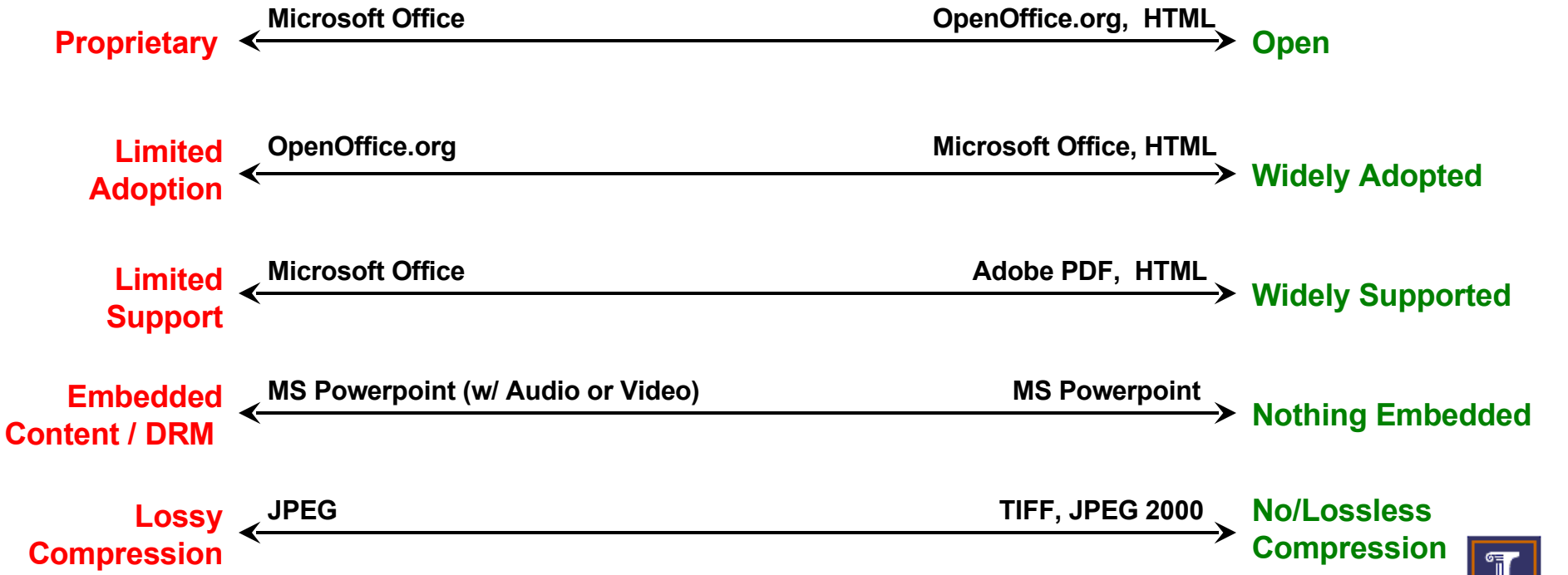
- Format-based Categories of Support
  - *High Confidence*
    - Full Support (including migration)
  - *Medium Confidence*
    - No migration *promised*
  - *Low Confidence*
    - "Bit-level" support only



(size ≠ weight)

Format Support Matrix

- Compilation of “known” formats
- Concentration on textual formats



Format Recommendations

Textual

- ⬆ CSV, Text, PDF/A, XML Open Document Format
- ⇔ RTF, MS Office, PDF, HTML

Images

- ⬆ TIFF, JPEG 2000
- ⇔ GIF, JPEG, PNG

Data Concentration

Audio

- ⬆ AIFF, WAVE, Ogg Vorbis, FLAC
- ⇔ AAC, MP3, Real, WMA

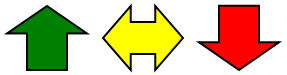
Video

- ⬆ AVI, Motion JPEG 2000
- ⇔ MP2, MP4, Quicktime, WMV

⬆ *High Confidence / Preference*

⇔ *Medium Confidence / Preference*

What we are doing

- Basic Activities (All Items: )
  - Regular Virus Scans, Checksum verification
  - Nightly off-campus backups
  - Refresh storage media
  - Preservation Metadata (extremely minimal)
    - Format, checksum, file size, etc.
  - Permanent Identifiers (Handles)
  - *Always* keep the original file(s)
  - Monitoring and reassessment of formats
    - Very minimal/infrequent for 

What we are doing

- Intermediate Activities (⇔)
  - Automated nightly “access copies” generated for major formats
  - When possible, attempt to migrate formats to preserve ***content*** and ***style*** (hopefully)
    - No promises that ***functionality*** will be preserved

Examples:

MS Excel → CSV (*possible functionality loss*)

MS Word → PDF (*possible style / font loss*)

What we are doing

- Full Support Activities ()
  - When necessary, migrate document to successive format.
  - Attempt to preserve ***content***, ***style*** and ***functionality***

  Example:

  OpenOffice.org 2.x → OpenOffice.org 3.x

What we are NOT doing

- Checking every file for content problems
  - character encodings, DRM, embedded content
- Verifying ALL automated migrations are "successful"
- Checking validity of format (e.g. JHOVE)
- Removing/modifying/replacing original file
  - Exceptions: viruses found or OCR necessary

Making data available in IDEALS

Data in IDEALS: still early days

- Mostly ‘simplistic’ data sets
- Data in spreadsheets, text, XML formats
  - i.e. “familiar” formats
- Problem: capturing relationships (between datasets, procedures, papers, etc.) in DSpace

Vole Demographic Data

Lowell Getz (Dept of Animal Biology)

- 25 years of data
- Data in 20 Excel files
- HTML Explanations
- PDF Manuscripts
- HTML “Sitemap” organizes data/files

Long-term vole demographic data files

Lowell Getz - Department of Animal Biology - University of Illinois at Urbana Champaign - May 2004 - Vitae

Data files of

1. 25 years of monthly live-trapping of Microtus ochrogaster and M. pennsylvanicus in alfalfa, bluegrass, and tallgrass prairie,
2. 63 months of twice-weekly 2-day live-trapping of M. ochrogaster in alfalfa.

Explanation

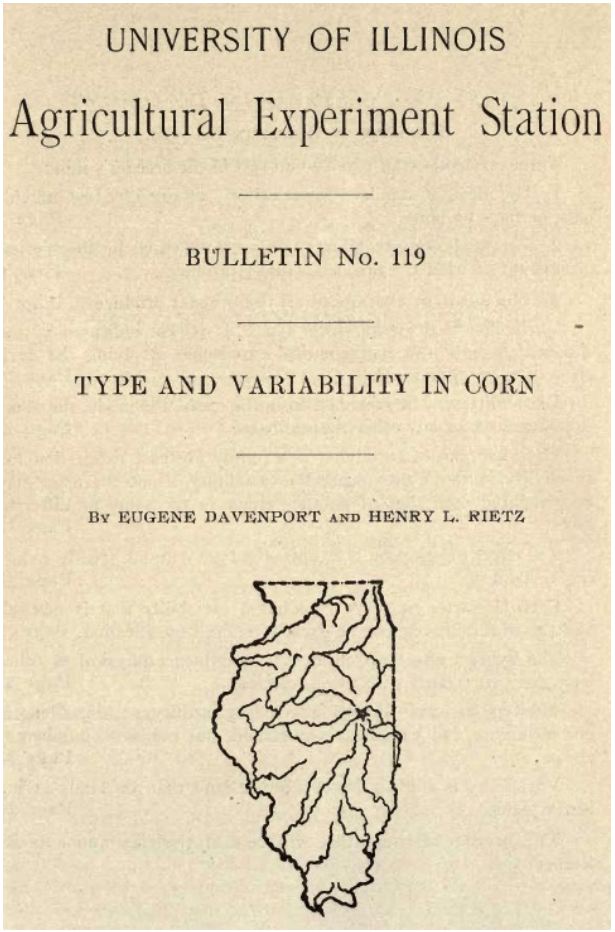
There are two primary sets of long-term data files in this folder. One set includes data obtained for the prairie vole, Microtus ochrogaster, and the meadow vole, M. pennsylvanicus, in three habitats (alfalfa, bluegrass, and tallgrass prairie) by trapping at monthly intervals for 25 years. The other data are for the prairie vole, from a study in which the population was trapped twice weekly for 63 months. In addition, there are files giving monthly estimates of population density of the short-tailed shrew, Blarina brevicauda, for each of the three habitats.

http://hdl.handle.net/2142/161

Illinois long-term selection exp. for oil & protein in corn

Department of Crop Sciences

- Data since 1896 (ongoing)
- SAS Statistical System files (text)
- ReadMe describes experiments
- Tech Reports in PDF
- Collection description organizes data/files


UNIVERSITY OF ILLINOIS

Agricultural Experiment Station

BULLETIN No. 119

TYPE AND VARIABILITY IN CORN

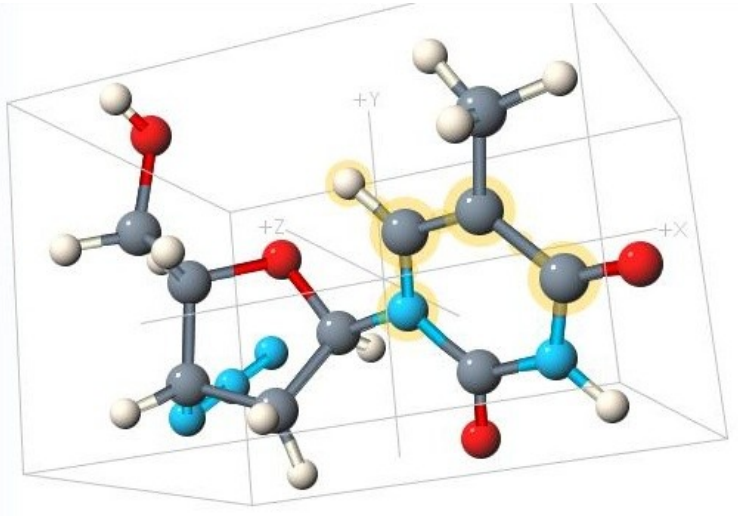
By EUGENE DAVENPORT and HENRY L. RIETZ

http://hdl.handle.net/2142/3524

Soon: Crystallography Data

Scott Wilson, School of Chemical Sciences

- Processed data in CIF (Crystallographic Info File)
- Transform to CML (Chemical Markup Lang.)
- Original unprocessed data kept on server in Clark X-Ray Facility (size concerns)
- Data available to SPECTRa Search tools…



Borrowed from Jmol website:
http://jmol.sourceforge.net/

SPECTRa http://www.lib.cam.ac.uk/spectra/

Future: Morrow Plots Data

- Oldest continuous agricultural research fields in USA (est. 1876), 2nd in world
- National Historical Landmark
- Data ongoing
- Likely require digitization of lab notebooks, etc.

Borrowed from Agronomy Day 2001 website:
http://agronomyday.cropsci.uiuc.edu/2001/morrow-plots/

Gaps – What we are NOT doing

- Making data "useable" directly from IDEALS
- Making data itself "searchable" (besides metadata)

Photo borrowed from:
http://www.flickr.com/photos/cseesze

- Providing unique "visualizations" of data
  - Some will be coming, for crystallography data

Sustainability issues…

- How do we preserve *more* than just files?
- How do we keep data understandable?
- Disk space concerns… versus access

Photo borrowed from: http://www.flickr.com/photos/columna/

A more “ideal” future

- When possible, finer grained access to data
- Better ways to show relationships between data and results/papers
- Begin to develop models for talking to faculty about their data

Photo borrowed from:
http://www.flickr.com/photos/dsevilla/

For More Information

http://ideals.uiuc.edu/

Sarah Shreeves
IDEALS Coordinator
sshreeve@uiuc.edu

Tim Donohue
Technical Lead
tdonohue@uiuc.edu

Policies / Documentation:

http://services.ideals.uiuc.edu/wiki/