SAYING WHAT WE DO—DOING WHAT WE SAY:

Preservation Issues (Metadata And Otherwise) In Institutional Repositories

Sarah L. Shreeves
University of Illinois at Urbana-Champaign
(with many thanks to Tim Donohue)

Intellectual Access to Preservation Metadata IG
ALA Annual Conference—July 12, 2009

WARNING!

Metadata will not be a huge part of this talk mostly because, well, most IRs don't do a good job at preservation metadata (or descriptive metadata for that matter).

More on that later...

Why do we start IRs?

- Centralize access to material produced at institution
- Create environment for preservation and permanent access to material
- Provide open access to content
- Advance a new scholarly communication model

Rieh, SY et al. 2008. "Perceptions and Experiences of Staff.." *Library Trends* 57 (2)

Why do we start IRs?

"an exploration or an experiment"

"don't have a clear notion of what it will become… [we're] asking [people on campus] to help us define what it can do for them…"

"a trend we should explore"

Preservation Challenge for IRs

can receive

pdf, doc, xsl, html, xml, txt, jpg, tiff, jp2, csv, rtf, avi, mp3, ppt, wav, ogg, png, gif, ram, odt...

from

faculty, staff, students

with

little to no knowledge of how materials were produced or their context

or answers to questions like

DRM? Embedded files? Lossy compression? Macros?

Regular back ups = digital preservation

"Not many interviewees were interested in digital preservation issues"

Confident in the long term sustainability of IRs

Interviewees were "far less coherent when discussing digital preservation."

"Those that were [interested] consistently emphasized that IR staff should know what they are promising."

TRAC compliance is part of the digital preservation program

From MIRACLE study at Univ. of Michigan

Why this study in contrasts?

Our software and technical infrastructure just does preservation ….

It's too hard to get our software and technical infrastructure to do that…

Preservation is something we can do later….

It's too hard period. ..We can't deal with data sets! We can't deal with audio and video! We can't deal with complex objects! We can't deal with petabytes!

No staff, resources, training, expertise….

In short, IR managers have been so distracted by access and ingest issues that very little attention has been given to date to the problem of how promises to preserve this material will be honored.

Building an IR without making plans for technological, organizational, and resource allocation is like building a house on sand.

McGovern and McKay. 2008. Leveraging short term opportunities... . Library Trends 57 (2)

Deep breath!

Browse

IDEALS
Titles
Authors
Subjects
Date
Communities

My Account

My Exports
Login to My IDEALS (U of I users only)
Non-Illinois Login

Information

Help
About
Contact Us

IDEALS Home

Welcome to IDEALS

See the top 10 downloads for 2008!

IDEALS collects, disseminates, and provides persistent and reliable access to the research and scholarship of faculty, staff, and students at the University of Illinois at Urbana-Champaign. Faculty, staff, and graduate students can deposit their research and scholarship - unpublished and, in many cases, published - directly into IDEALS. Departments can use IDEALS to distribute their working papers, technical reports, or other research material. Contact Sarah Shreeves, IDEALS Coordinator, for more information.

Recent Additions

[2009-07-11] **Hypergraph-Based Combinatorial Optimization of Matrix-Vector Multiplication**
Wolf, Michael M. (2009-07-11)
PDF (1MB)

Top Downloads this Month

1. Textron: Fostering Continuous Improvement in a Changing Business Context [total: 264]
2. A Statewide Collection Map: Results from the Statewide Assessment of Monographs and Electronic Resources in Illinois [total: 224]
3. Estrategias para Motivar el Aprendizaje Colaborativo en Cursos a Distancia [total: 197]
4. Does Fitness Bring People Together? [total: 168]
5. Libraries, People, and Change: A Research Forum on Digital Libraries (Papers presented at the Allerton Park Institute held October 27-29, 1996) [total: 137]
6. The dream chasers [total: 108]

Promises, Promises

"create a reliable and easy to use repository service to preserve, manage, and provide persistent and widespread access to the digital scholarship faculty and students now produce... "

- 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? (Dspace was mostly already chosen...)

Getting our act together

1. Starting talking to our Preservation Librarian!

2. Training and self education

3. Assessment of where we were and where we needed to go

Takeaways

- "Preservation" needs to be unpacked.
- Not about the technology.
- Explicitness is key.
- You don't have to preserve everything to the fullest extent if you say you aren't.

From Dorothea Salo. 2009. Institutional repositories for the digital arts and Humanities. Humanities Digital Curation Institute. Champaign IL. May 2009. http://www.slideshare.net/cavlec/digital-preservation-and-institutional-repositories

Getting our act together pt 2

- Secured **explicit** administrative support and commitment for digital preservation management program in DEALS. http://hdl.handle.net/2142/135
- Developed high level preservation policy: http://hdl.handle.net/2142/2383
- Developed actionable procedures and policies that can be reassessed and changed as needed
- Began next stage of identifying gaps, like… .

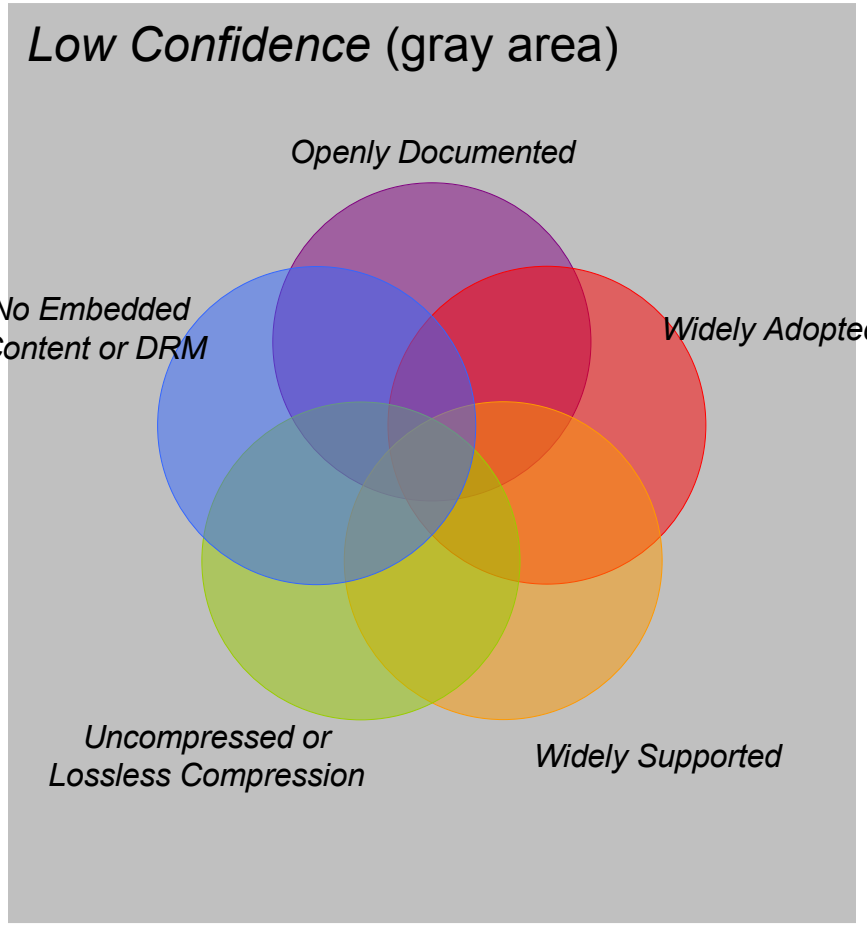
Getting our act together pt 2

Backup tapes stored next to the server!

Not Really Our Server Room!

Digital Preservation Support

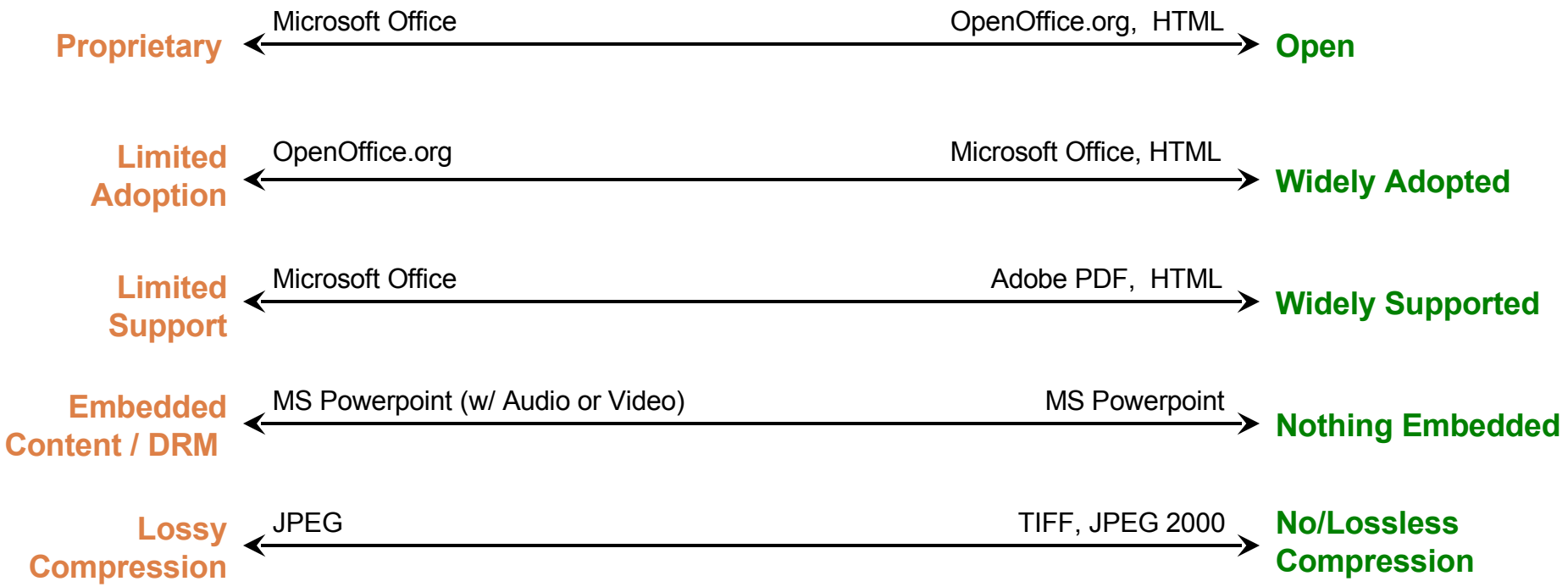
- Format-based Categories of Support
  - *High Confidence*
    - Full Support (including migration)
  - *Medium Confidence*
    - No migration *promised*
  - *Low Confidence*
    - "Bit-level" support only



(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

Audio

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

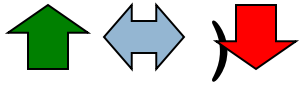
Video

- ⬆ 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
  - a Nightly off-campus backups
  - u Refresh storage media
  - m Preservation Metadata (minimal)
    - Format, checksum, file size, etc.
  - , Permanent Identifiers (Handles)
  - f *Always* keep the original document
  - l Monitoring and reassessment of formats
    - Very minimal/infrequent for 

What we are doing

- Intermediate Activities ( )
  - Additional monitoring, more frequent reassessment
  - When possible, attempt to migrate formats to preserve ***content*** and ***style*** (hopefully)
    - No promises that ***functionality*** will be preserved
    - (e.g.) Powerpoint → PDF (*possible functionality loss*)
    - (e.g.) PDF 1.4 → PDF/A (*possible style loss*)

What we are doing

- Full Support Activities ()
  - Additional monitoring, more frequent reassessment
  - When necessary, migrate document to successive format.
  - Attempt to preserve *content*, *style* and *functionality*
    - (e.g.) PDF/A → successor to PDF/A

About that metadata… .

We automatically collect:

- type of format (but this is not verified)
- size of file
- provenance information (who deposited it and when; automatic conversion activities; and SOME changes that occur later in a file life)
- checksum

If we make manual changes our procedure is to manually add information to provenance information.

Our First Problem…

- Character issues in Word (and PDF)
- Found by chance
- Consultation with submitter
- Originally Wordperfect
- Re-submitted as RTF

AWhite and Nerdy@:
Computers, Race, and the Nerd Stereotype

Lori Kendall

Previous research on nerds has analyzed the relationship of this stereotypical identity to issues of race, gender, and computer expertise. For instance, in an earlier article, I argue that narratives such as that presented in the popular movie, Revenge of the Nerds, depict the incorporation of the previously marginalized nerd identity into closer alliance with hegemonic masculinity, demonstrating the increasing legitimacy of expertise in computers as a form of masculine prowess. However, I also suggest that "the continued negativity of the nerd stereotype reveals a persistent uneasiness with computer use and computer users…" (280) In a similar analysis, Ron Eglash analyzes images of nerds as white and male by default, yet "hardly a portrait of white male superiority" (50). He explores possibilities for reversal, analyzing images of black nerds in popular culture, and attempts by black and women to subvert the nerd stereotype. However, ultimately he notes that "nerd is still used in the pejorative sense; its routes to science and technology access are still guarded by the unmarked signifiers of whiteness and male gender" (60).

Both of these articles point to contradictions in nerd identity that allow it to "both maintain normative boundaries of power and offer sites for intervention" (Eglash 49). It is logical to expect that the tension inherent in those contradictions would resolve over time, as computers become more ubiquitous in society. We've been through several cycles of developments in computer- and internet-related technologies, including the phenomenal wave of internet start-ups in the 1990s (just prior to my previous article) and the subsequent "dot-com bust" of 2000 (just prior to Eglash's). In the U.S., information technologies are increasingly part of most people's lives. As the household presence of computers becomes no more extraordinary than that of other consumer electronics such as televisions and microwave ovens, one might expect that the nerd stereotype would fade from view, an anachronism from an earlier age, reflecting now-defunct uneasiness with the then-new computer technologies.

Big Gaps!

- We aren't checking the validity of formats
- We collect pretty minimal metadata
- We're not checking every file for problems
- We don't check every automated conversion

BUT

- We do explicitly acknowledge these gaps.

Some questions...

- What's the right balance in IRs?
- Is transparency an issue?
- Are some materials more deserving of 'full' preservation than others in our IRs?

Contact Information

Sarah Shreeves

Coordinator, IDEALS

http://www.ideals.uiuc.edu/

217-244-3877

sshreeve@illinois.edu