DIGITAL PRESERVATION IN THE WILD

Tim Donohue
Research Programmer - IDEALS
University of Illinois
(with many thanks to Sarah Shreeves)
It all starts with a small story

a dream of sorts
IDEALS: the “dream”

“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…”
IDEALS: the "dream"

“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…”

BUT…

What’s it mean to **preserve** this stuff?

Can we preserve **everything**?

What kind of **infrastructure**?

What kind of **expertise** do we need?

What kind of **resources**?
IDEALS: the initial reality

Not Really Our Server Room!

Backup tapes stored next to the server!
1. Brought in Preservation Librarian

2. Training and self education

3. Assessment of where we were and where we needed to go
Open Archival Information System (OAIS) Model

http://public.ccsds.org/publications/archive/650x0b1.pdf

Image borrowed from the ICPSR Digital Preservation Tutorial:
http://www.icpsr.umich.edu/dpm/
The Foundation II

TRAC (Trustworthy Repositories Audit & Certification)

http://www.crl.edu/PDF/trac.pdf

- Organizational Infr.
- Digital Object Mgmt
- Technical Infr. & Security
http://www.slideshare.net/cavlec/digital-preservation-and-institutional-repositories
“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.
The 5 Stages of Preservation

- Denial / Ignorance
- Anger / Fear
- Bargaining
- Depression
- Acceptance & Hope

Based on the Kübler-Ross five stages of grief:
http://en.wikipedia.org/wiki/K%C3%B6bler-Ross_model
Denial / Ignorance

Google™ backups**

** - *This service is entirely fictitious*

Again, Not Really Our Server Room!
Anger / Fear

Data Loss

Obsolescence
Bargaining

Please, just let me get *this data* migrated elsewhere
Depression

This is too hard.

How can we ever preserve everything?

We don’t have the resources for this.

Why even try?
Acceptance & Hope

We can take small steps to...

- Preserve some things locally
- Develop policies (say what you do)
- Enact policies via procedures (do what you say)
- Work with others on best practices to preserve the rest

"In the end, we will conserve only what we love, we will love only what we understand, we will understand only what we are taught."
—Baba Dioum

YOUR EXISTENCE GIVES ME HOPE
The Principles of Preservation

(1) Say what you do...

(2) Do what you say...

IDEALS - Saying what we do

- Secured explicit administrative support and commitment for digital preservation management program in IDEALS. 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 & documenting gaps
Format-based, “Categories of Support”

- **High Confidence**
  - Full Support

- **Medium Confidence**
  - No migration promised

- **Low Confidence**
  - “Bit-level” support only

IDEALS Preservation Support Policy

https://services.ideals.uiuc.edu/wiki/bin/view/IDEALS/PreservationSupportPolicy
### IDEALS Format Support Matrix

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

<table>
<thead>
<tr>
<th>Proprietary</th>
<th>Microsoft Office</th>
<th>OpenOffice.org, HTML</th>
<th>Open</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited Adoption</td>
<td>OpenOffice.org</td>
<td>Microsoft Office, HTML</td>
<td>Widely Adopted</td>
</tr>
<tr>
<td>Limited Support</td>
<td>Microsoft Office</td>
<td>Adobe PDF, HTML</td>
<td>Widely Supported</td>
</tr>
<tr>
<td>Embedded Content / DRM</td>
<td>MS Powerpoint (w/ Audio or Video)</td>
<td>MS Powerpoint</td>
<td>Nothing Embedded</td>
</tr>
<tr>
<td>Lossy Compression</td>
<td>JPEG</td>
<td>TIFF, JPEG 2000</td>
<td>No/Lossless Compression</td>
</tr>
</tbody>
</table>
IDEALS Format Recommendations

**Textual**
- High Confidence / Preference: CSV, Text, PDF/A, XML, Open Document Format
- Medium Confidence / Preference: RTF, MS Office, PDF, HTML

**Images**
- High Confidence / Preference: TIFF, JPEG 2000
- Medium Confidence / Preference: GIF, JPEG, PNG

**Audio**
- High Confidence / Preference: AIFF, WAVE, Ogg Vorbis
- Medium Confidence / Preference: AAC, MP3, Real, WMA

**Video**
- High Confidence / Preference: AVI, Motion JPEG 2000
- Medium Confidence / Preference: MP2, MP4, Quicktime, WMV

For more information, visit:
https://services.ideals.uiuc.edu/wiki/bin/view/IDEALS/FormatRecommendations
Basic Activities (All Items: ▶️ ▶️ ▶️)
- Regular Virus Scans, Checksum verification
- Nightly off-campus backups
- Refresh storage media
- Preservation Metadata (minimal)
  - Format, checksum, file size, etc.
- Permanent Identifiers (Handles)
- Always keep the original document
- Monitoring and reassessment of formats
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)
IDEALS – Doing what we say

- 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
Our First Preservation Problem...

- Character issues in Word (and PDF)
- Found by chance
- Consultation with submitter
- Caused by conversion to Word (from Wordperfect)
- Resubmitted 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. In a similar analysis, Ron Eglash analyzes images of nerds as white and male by default, yet hardly a portrait of white male superiority. 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.

Both of these articles point to contradictions in nerd identity that allow us to both maintain normative boundaries of power and offer sites for intervention. It is logical to expect that the tension inherent in these 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, anachronism from an earlier age, reflecting now-defunct uneasiness with the then-new computer technologies.
We Acknowledge our Gaps

- Not checking format validity (yet)
- Minimal metadata collection
- Not checking files for problems (besides viruses)
- Not checking every automated conversion
Back to that “dream”?

“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…”

**Total Items 2007-09**

![Graph showing Total Items 2007-09]

**Downloads per month**

![Graph showing Downloads per month]

Total Items: 11,500
Total Downloads: 870,000+
Credits

- Slide 2 (Book image): http://www.flickr.com/photos/riot/100006656/
- Slide 3 (MS office): http://www.flickr.com/photos/niallkennedy/374272762/
- Slide 5/13 (Server room): http://www.flickr.com/photos/sylvar/31436961/
- Slide 6 (Get your act…): http://www.flickr.com/photos/dreamsjung/3595425744/
- Slide 11 (I know this…): http://www.flickr.com/photos/ali_kat_xx/1373989245/
- Slide 13 (Disk backups): http://www.flickr.com/photos/tonyaustin/2355186770/
- Slide 14 (Disk lost): http://www.flickr.com/photos/mag3737/2415681602/
- Slide 14 (Broken cd): http://www.flickr.com/photos/rickheath/72041533/
- Slide 15 (VHS to DVD): http://www.flickr.com/photos/28910181@N05/3085532220/
- Slide 15 (Mac transfer): http://www.flickr.com/photos/cyprien/6173244/
- Slide 16 (Depression): http://www.flickr.com/photos/dbarefoot/2652496167/
- Slide 17 (Hope): http://www.flickr.com/photos/livenature/259458056/
- Slide 17 (Dioum Quote): http://www.flickr.com/photos/wallyg/469808717/
- Slide 27 (Gaps): http://www.flickr.com/photos/aduki/2416528101/
Contact Info

Tim Donohue
University of Illinois
tdonohue@illinois.edu
http://www.ideals.uiuc.edu/
http://www.ideals.uiuc.edu/wiki/

This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License