Persistence, *got yours?* Preserving Scholarship.

Victoria Reich 

Stanford University Libraries

Katherine Skinner EDUCOPIA INSTITUTE

March 11, 2014

Persistence?

"We are made to persist.
that's how we find out who we are."
Tobias Wolff - Professor, Stanford University

Victoria

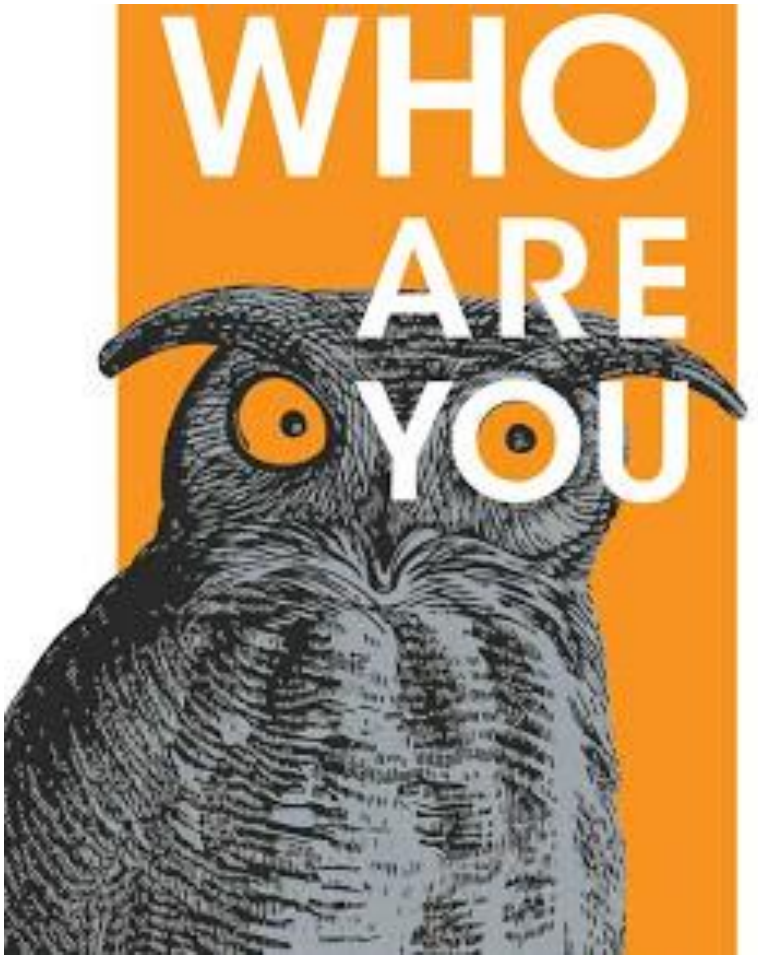
http://www.lockss.org/contact-us/victoria-reich/

Katherine

http://educopia.org/staff/katherine-skinner

Welcome!

We would like to learn more about you!



http://itsaboutartanddesign.blogspot.com/2012/10/who-are-you-vintage-owl-poster.html
This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License

Your Workplace

- Library
- Publisher
- Aggregator
- Other - please specify in Questions box

Pick One

Your Workplace – Annual Library Collection Budget?

- $25+ million
- $15-25 million
- $6- 15 million
- Under $6 million
- Don’t know/Not a library

Pick One

Your Workplace – Annual Publisher Revenues

- >$50 million
- $10-50 million
- $5-10 million
- under $5 million
- Don't know/Not a publisher

Pick One

Your DP Experience

- Novice – new to concepts & principles
- Intermediate – familiar with concepts & principles
- Advanced – have hands on experience
- Expert – community leader

Pick One

What Is The Scholarly Record?

Scholarship defined:

“Knowledge resulting from study and research”

Its record:

Journals, books, newspapers, blogs, government documents, maps, videos, websites, music, art….

What's The Point?
Digital Preservation

- Access
- Use & Services

Is there really a problem?

How do we know?

- "Atlas of Digital Damages"
- PACER and public court documents
- Cyberwar on Estonia
- Link-rot and Thomson Reuters' Web of Science
- General digital loss stories (and why they're quiet)

Most Content Is Web-based

- Books, journals, web pages, databases, government documents, newspapers, blogs.
- Web size estimate: 1200 Petabytes
- Internet Archive (Wayback) size estimate: 9 Petabytes
  - A Petabyte is 1000 Terabytes https://en.wikipedia.org/wiki/Petabyte

I am most concerned

- About persistent access to electronic
  - Toll (paid) access books and serials
  - Open access books and serials

Pick One

I am most concerned

- About persistent access to electronic
  - Government documents
  - News
  - Databases
  - Digitized content in my IR
  - Web sites

Multiple Answer

Biggest Threat to Content?

- Format obsolescence
- Media, hardware and software obsolescence/failure
- Economic failure
- Natural disaster
- Humans
  - Operator error
  - Insider attack
  - Overwriting

Pick One

Economic, Social Challenges

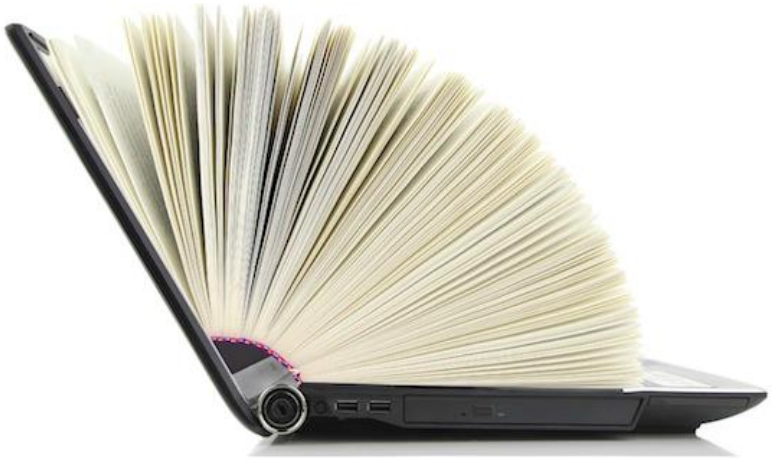
- Digital preservation is a hard sell
  - "now" and not the "future"
  - Post-cancellation access
    - "what we pay for" not the "freely available"
- Unresolved questions
  - Who bears responsibility?
  - Who holds IP/rights?
  - How ensure coverage while minimizing duplication?
- Sustainability

E-only/E-first Publications

- Journals
- E-books
- News
- Datasets

Image Source: http://dianesavino.d3photogroup.com/

SO—let's look more closely at the content set that has received the most attention…



JOURNALS…

Image source: http://library.polytechnic.wa.edu.au/ebooks-and-ejournals

How Well Are We Doing?

- Attempts to measure probability that journals are preserved
  - 2010 ARL median research library receives ~80K serials
  - Keepers Registry reports 21K preserved; 10.5K in progress
  - Other sources confirm 40% to 50% of the content is under some kind of “treatment”

WAY TOO OPTIMISIC!!

50% Is Not Risk Adjusted

- Librarians' primary efforts have focused on post cancellation access, not on preserving the record of scholarship
- What's preserved is the output of the big/expensive publishers NOT the small/affordable (or OA) publishers

50% Is Not Adjusted For Difficulty

- Format and structure
  - Common formats; least likely to become obsolete
  - Homogenous structure (PDF with ONIX) spit out of publishing systems
- Rights
  - Big publisher negotiations = tons of content
  - Little publisher negotiations = little bit of content

50% Looks Backwards

- Only looking at traditional media
- Consider contemporary scholarship forms
  - Workflows, source code repositories, social media
- These new forms often
  - Lack well established and robust business models
  - Are large dynamic corpus
    - Difficult to preserve
  - Operate in an unclear legal framework
    - What does copyright mean for a work that's made up from dynamic "mash-up" from around the web

50% Eclipses Source Citation Rot

"According to a 2014 study conducted at Harvard Law School, 'more than 70% of the URLs within the *Harvard Law Review* and other journals, and 50% of the URLs within United States Supreme Court opinions, do not link to the originally cited information.'"

-Jill Lepore, "The Cobweb." *New Yorker* January 26, 2015

http://www.newyorker.com/magazine/2015/01/26/cobweb

The Other 50+%?

- The community is preserving much less than ½ journal content now
- The community is struggling to preserve even that much

And then, there's that other content…

- Primary sources
  - News
  - Data
  - eRecords
  - Government information
  - Personal archives of major figures
  - Ebooks
  - Films, music, tv, netflix, cultural productions
  - Social media
  - Web-based content (e.g., human rights violations, refugee emergencies, natural disasters, etc.)

Reality:

The rate of loss to future researchers from "never preserved" will vastly exceed that from all other causes.

What To Do?

Photo: Jim Merithew/Wired.com

Build Collections Thoughtfully

- Take stock of your collections
  - identify those that are unique and those that are jointly held
- Advocate for preservation of your collections
  - Know what options are available
  - Match options to collection types
- Insist on open source solutions and transparency

Share Expertise, An Example

- LOCKSS Program TRAC/ISO16363 audit
  - Equaled previous highest score; first ever perfect score for technology
  - Documents.clockss.org
    - all documents on which auditors based assessment
  - Linked from dshr.blog.org
    - Announcement certification
    - Process description
    - Lessons learned
    - Demos you can run for yourself

Think Globally Act Locally

- Standards and practices
  - Too few resources to develop tools, standards, practices applicable to only our community
    - OAIS
    - Metadata (PREMIS, METS)
  - Almost all content is web published; what tools do Google, Facebook and Apple use?

Preservation Stories

- ETD Preservation
  - http://educopia.org/research/electronic-theses-and-dissertations
- Newspaper Preservation
  - http://educopia.org/research/chronicles
- Digital Forensics
  - http://www.bitcurator.net/

Preservation In Action

- Archive-It https://www.archive-it.org/
- Chronopolis http://libraries.ucsd.edu/chronopolis/
- CLOCKSS Archive www.clockss.org
- DuraCloud Preservation Plus www.duracloud.org/
- HathiTrust http://www.hathitrust.org/
- Internet Archive https://archive.org/web/
- LOCKSS Program www.lockss.org
- MetaArchive Cooperative www.metaarchive.org/
- Portico www.portico.org
- StoryTracker http://storytracker.pastpages.org/en/latest/#

Organizations

- IIPC: International Internet Preservation Consortium
  - http://netpreserve.org/
- DPC: Digital Preservation Coalition
  - http://www.dpconline.org/
- OPF: Open Preservation Foundation
  - http://openpreservation.org/
- NDSA: National Digital Stewardship Alliance
  - http://www.digitalpreservation.gov/ndsa/
- DPOE: Digital Preservation Outreach and Education
  - http://www.digitalpreservation.gov/education/

Thank you