Watching Our Backs

Community verification of digital preservation systems

John Mark Ockerbloom

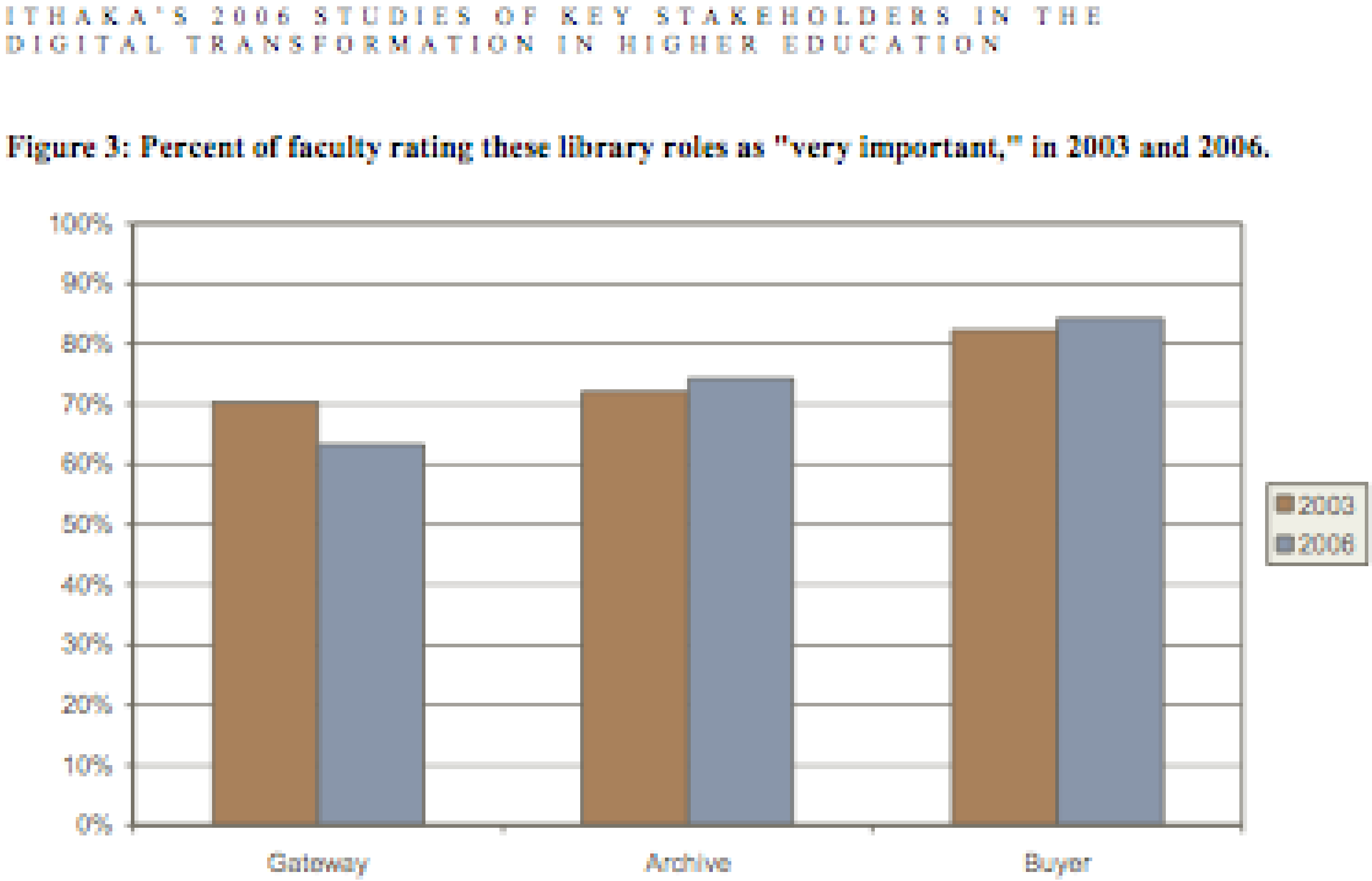
Digital Library Federation Fall Forum

November 14, 2008

Key ideas of this talk

- For preservation: “Trust, but verify”
- Client usage is important part of verification
  - Example: LOCKSS verification at Penn
- Tests can be planned and carried out for many types of outcomes, systems
- Shared verification efforts sustain shared preservation efforts

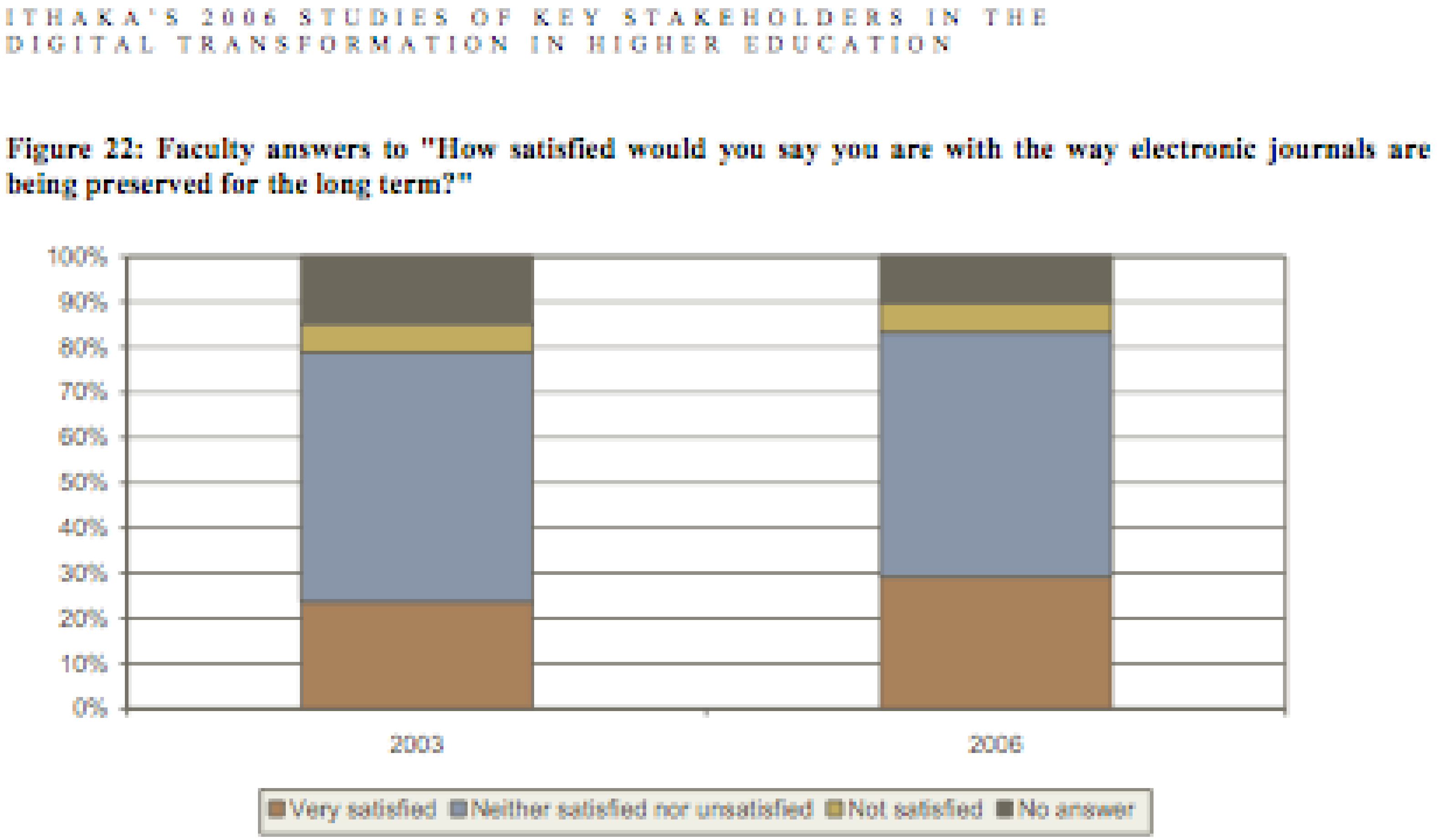
Preservation is valued



ITHAKA'S 2006 STUDIES OF KEY STAKEHOLDERS IN THE DIGITAL TRANSFORMATION IN HIGHER EDUCATION

Figure 3: Percent of faculty rating these library roles as "very important," in 2003 and 2006.

But assurance lags



ITHAKA'S 2006 STUDIES OF KEY STAKEHOLDERS IN THE DIGITAL TRANSFORMATION IN HIGHER EDUCATION

Figure 22: Faculty answers to "How satisfied would you say you are with the way electronic journals are being preserved for the long term?"

What are we investing in, electronically?

- Electronic materials: > 40% of ARL materials budgets
  - (2005-2006 figures; some libraries reported > 50%)
- Electronic preservation: much smaller investments
  - Local preservation largely special rather than general collections
  - "Preservation in place" delegates preservation to publishers by default
  - Preservation consortia for libraries developing
    - Portico, LOCKSS, Hathi Trust, Preserv…
- Questions library directors have:
  - What are we buying?
  - How much will it cost us? (Not just now, but also in future)
  - How do we know they'll give us what we need when we need it?
    - Especially when preservation copy not the usage copy
  - What might go wrong?
  - What happens when things go wrong?
- Early reassurances can avert future nasty surprises

Centralized audits

Benefits:

- Can lower redundancy and costs (by outsourcing to experts)
- Can thoroughly vet policy and management (via things like OCLC's audit checklist)
- "Trusted broker" can evaluate sensitive data (confidential content, finances, etc…)

Limitations:

- Ultimate test of preservation is usage, not audit
- Auditors will not interact with preservation systems in same manners, extents, as actual clients
- Finding, funding appropriate auditor may be problematic

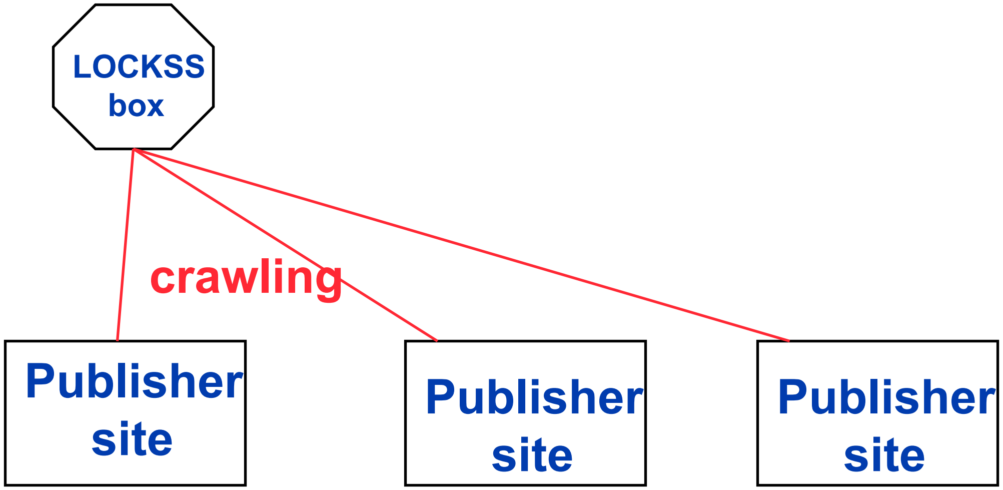
Distributed client auditing

- Measure, record, share externally visible preservation outcomes
  - Through normal usage, and through controlled experiments
- Testing by clients for clients
- Costs can be spread out among clients, targeted and scaled according to client concerns
- Clients need to have appropriate rights to do test
  - Dark archives need to open up appropriate access both for testing and recovery
  - Client testers need to be able to share results (at least among selves)
- Different types of archives may call for different types of tests
  - E.g. centralized third-party archives like Portico vs. distributed self-maintained archives like LOCKSS

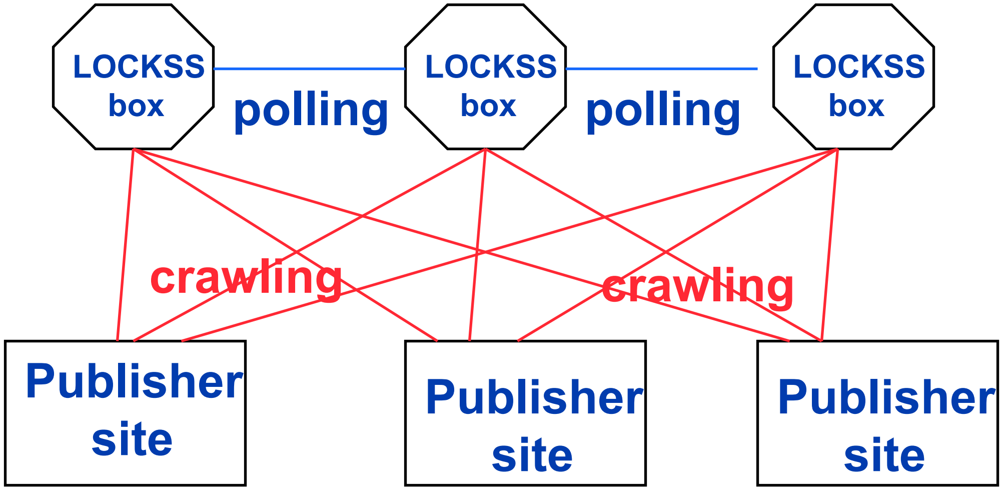
An example: LOCKSS

- Install LOCKSS box(es) to preserve journals, other static content you subscribe to
  - If a crawling plugin exists, and the publisher has okayed LOCKSS crawling
  - Content cached on archive disk(s), noted in manifest file
  - Content periodically checked against peers
- If publisher content lost:
  - LOCKSS box takes over delivery
- If locally cached content lost:
  - LOCKSS box "self-repairs" from peers that it's checked with in the past

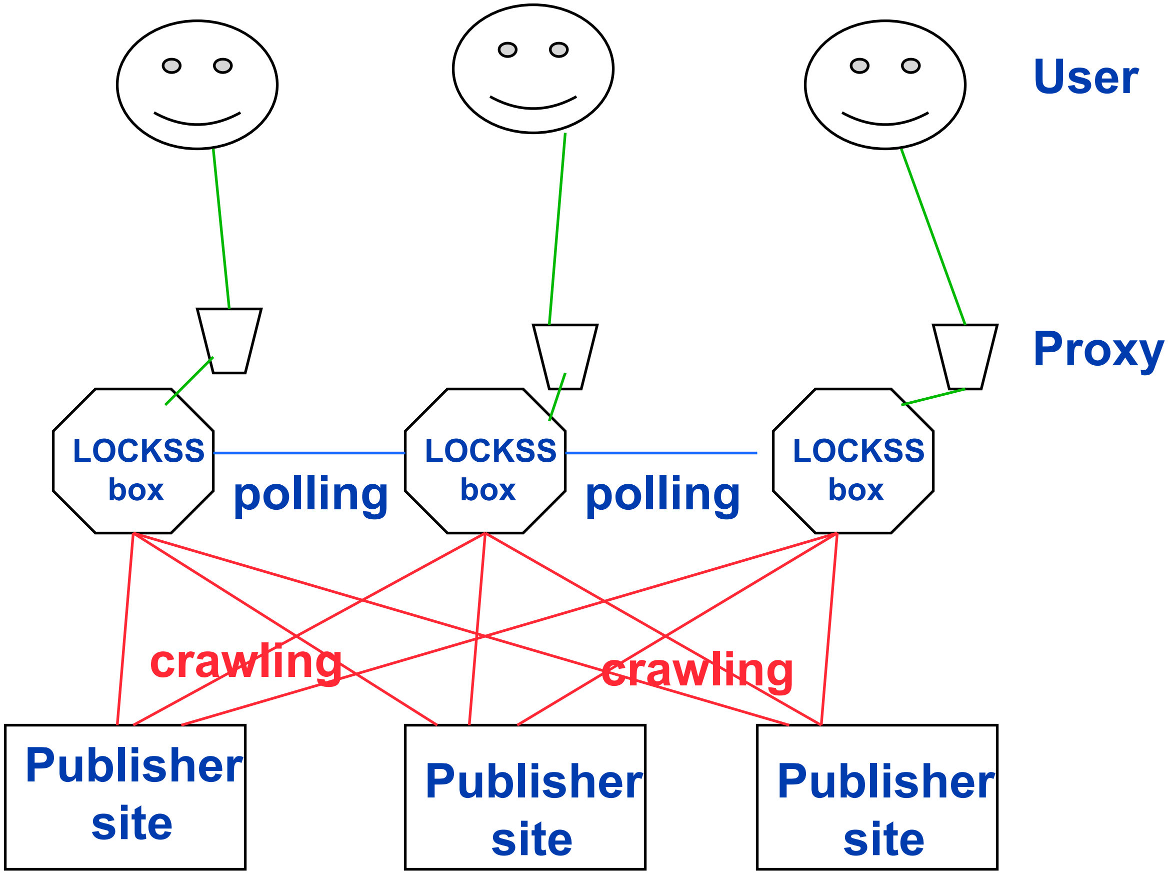
How LOCKSS works



How LOCKSS works



How LOCKSS works



First failure test: Spring 2007

- 80 GB disk filled up
- We backed up our manifest file, then replaced our archive disk with an empty disk
- Most of the archive self-repaired, but not all
  - Most reconstructed from crawls, but not all material was still crawlable (expected, due to publisher site and subscription status changes)
  - Some reconstructed from polling, but some didn’t, or did so unacceptably slowly (not expected; apparently due to protocol changes around the time of the failure test)
- LOCKSS worked with us to expedite repairs, and updated protocol to avoid problem in future
- We planned for another failure test

Support for testing by preservation system crucial

- Diagnostic tools
  - Overall summaries of crawl and poll status
  - Drill down to individual archival units
- Controls
  - Could decide which archival units to include on box
  - (Might also be useful to have controls for running test scenarios, as with certain programming practices; pulling disks a little drastic)
- LOCKSS staff willing to work with me and respond to my concerns
  - Thanks especially to Tom Lipkis
- Preservation systems need to give enough information, control to let users easily detect when things go wrong, diagnose causes
  - Trust for trust

Second failure test: Summer 2008

- 250 GB disk filled up; we did another empty disk swap
- Recovery crawls unexpectedly slow
  - And seemed to be oddly reported
  - (and poll-based recovery wouldn't happen until crawls had been tried or archival units marked as "discontinued")
- Problems included
  - Runaway recursive crawls
  - Misleading crawl summaries
  - Publisher bottlenecks (crawls still not done as of now)
  - Poll recovery still too slow in some cases (file by file)
- Subsequent daemon releases designed to alleviate many of these problems
- Larger scale made many of these problems manifest

What can be tested by clients

- Operation: See if recovery, access, etc., work as expected under controlled or live conditions
  - E.g. failure test, proxy test, versioning/migration tests…
- Coverage: See if titles and volumes are present with the coverage and currency we expect
  - E. g. title and volume content scans against library holdings or pub. list
- Fidelity: See if contents are what we expect them to be (and in expected formats)
  - E.g. file and metadata sampling with visual cross-check, JHOVE validation; manifest checking if applicable
- Policy: See if repository meets its obligations to libraries
  - E.g. check reports based on OCLC's *Trustworthy Repositories Audit and Certification* checklist; see if checklist items need to be added or expanded
- Multi-category: E.g. post-cancellation replacement tests

Investments needed for tests

- Staff expertise and focus
  - Know how the systems work, commit to oversight
    - » Penn “Lockss/Portico group” supported by admins
  - Know what outcomes to expect, behaviors to watch
- Staff time for testing and reporting
  - Devise experiments / measurements
  - Conduct tests, monitor progress
  - Share results with appropriate audiences
  - LOCKSS test time: a few minutes a week to monitor recovery, a few hours total to write up summaries and questions for LOCKSS staff
- In some cases, special equipment / environment
  - For LOCKSS test, using the production box not a good idea if scale high, or cache in active use
  - But LOCKSS boxes are commodity items

Efficient, effective community auditing

- Check with archives/projects to see what formal audits have occurred or are planned
  - E.g. from the OCLC checklist
  - And see the reports (if you're paying an organization to audit, they should let you see the reports, if not full data)
- Plan simple tests for cases of concern not covered in formal audit
  - E.g. failure test, proxy usage, migration assessment
- Share results with community
  - Useful to have well-known location/index of such results
- Work with other coalition members to make sure bases are covered, redundancy minimized
  - Can be a fairly lightweight process, using existing organizations (e.g. CRL, NERL) or user/customer groups for Portico, LOCKSS…
  - Can also involve collaboration to automate more complex tests, monitoring

Moving testing into the community: CRL

- Did audits of Portico, planning more (along with AP, UMI Dissertations, Hathi Trust, other groups)
- Level, focus of audits based on interests of members (who are funding the audit expenses)
- Acting as "trusted broker" (for things like financial reviews)
- Convened small group to consider, plan community auditing
  - Including Penn, CDL, Chicago, Dartmouth, CDL, TRLN…
- Conducted survey of usage and concerns of LOCKSS, Portio users
- Confluence space used to share some results, reports
- Interested in knowing more, participating?
  - See http://www.crl.edu/
  - Or Contact Bernie Reilly (reilly@crl.edu)

Other community focuses?

- Purchasing groups: influence and funding
  - publishers to use preservation backup
  - preservation systems to be adequately tested
  - fund crucial audit, development activities?
- Research support: planning and development
  - Where is testing most effective? What tools, infrastructure can be built to enhance verification and quality assurance?
- Shared knowledge resources: coordinate testing plans and results
  - Simple options: Wiki
  - More complex: Registry
  - Piggyback on WorldCat/union catalog/global metadata network?
    - We’ve done it for digitization and rights info, why not preservation info as well?

Conclusion: Preservation with our eyes open

- We must verify that our digital archiving systems work
  - We've invested huge amounts in these electronic materials
  - Diagnose problems before they bite us, improve the systems
- We clients have the resources and expertise to do this
  - Can verify outcomes, not just inputs and practices
  - Can find important results not found in centralized auditing
- We can coordinate to magnify the effectiveness and efficiency of our verification
  - Through consortial organization, shared knowledge resources, influence on publishers and preservation organizations
- First steps: Harness the will to find the way…
- Thanks!
  - My contact address: ockerblo@pobox.upenn.edu
  - Slides: http://works.bepress.com/john_mark_ockerbloom/