Lots of LOCKSS Keeping Stuff Safe: The Future of the LOCKSS Program

Nicholas Taylor (@nullhandle)
Program Manager for LOCKSS and Web Archiving
Stanford University Libraries

CNI Fall 2016 Membership Meeting
12 December 2016
why more LOCKSS?

- mature, community-validated technology
- research-based + built to a specific threat model
- web-centric preservation for web-centric scholarship
- community-centric preservation for collective challenges + opportunities
- robust, distributed digital preservation

“Cologne Love Padlocks” by orkomedix under CC BY-NC-SA 2.0
Program History
inception

• a serials librarian + a computer scientist
• print journals → Web
• conserve library’s role as preserver
  • collect from publishers’ websites
  • preserve w/ cheap, distributed, library-managed hardware
  • disseminate when unavailable from publisher

Chris Dobson: "From Bright Idea to Beta Test: The Story of LOCKSS"
philosophy + focus

• lots of copies keep stuff safe
• preservation is an active community effort
• lots of communities keep stuff safe
• enable communities to preserve + access their scholarly record
present day

- financially self-sustaining
- tens of networks
- hundreds of institutions
- all types of content
looking forward

• organizational changes
• software evolution
• LOCKSS networks
• distributed digital preservation
Organizational Changes
David + Vicky
personal introduction

• 10 years in research libraries:
  • Stanford University Libraries (2013 – present)
  • Library of Congress (2010 – 2013)
  • U.S. Supreme Court (2007 – 2010)

• professional background:
  • web archives
  • digital library services
  • library technology

• what I care about:
  • scalability + sustainability of PLNs, CLOCKSS
  • mainstreaming LOCKSS for digital preservation
  • building collaborative technical communities
SUL Web Archiving

- end-to-end service:
  - collect
  - preserve
  - make accessible
  - make discoverable

- integrate w/ collection development

- use cases:
  - scholarly inputs/outputs
  - institutional legacy/compliance
  - government information
LOCKSS + DLSS administrativa

- LOCKSS integrating w/ SUL Digital Library Systems & Services (DLSS)
- led by Tom Cramer, Director & Associate University Librarian
- LOCKSS + SUL Web Archiving, under Nicholas Taylor
LOCKSS + DLSS synergies

• realize operational efficiencies
• adopt, drive shared engineering best practices
• promote API-oriented architectures
• streamline repository → PLN data hand-offs
• contribute upstream to shared tools
• broaden, diversify community outreach
new functionality

- supported by Mellon Foundation grant
- ingest/harvest
  - form-filling
  - AJAX
- dissemination
  - Memento
  - Shibboleth
- preservation
  - polling performance
new architecture

• existing functionality
• discrete components as web services
• incorporate external software

*San Francisco Oakland Bay Bridge, East Spans New and Old* by Shanan under CC BY-NC 2.0
web services imperative

1. “All teams will henceforth expose their data and functionality through service interfaces.”
2. “Teams must communicate with each other through these interfaces.”
3. “There will be no other form of interprocess communication allowed: no direct linking, no direct reads of another team's data store, no shared-memory model, no back-doors whatsoever.”
4. “All service interfaces, without exception, must be designed from the ground up to be externalizable. That is to say, the team must plan and design to be able to expose the interface to developers in the outside world.”
5. “Anyone who doesn't do this will be fired.”

Steve Yegge: “Stevey's Google Platforms Rant”
risk of large projects

small projects (< $1 million)  large projects (> $10 million)

- successful (on time, on budget): 4%  10%
- challenged (late, over budget, lacking functionality): 20%  38%
- failed (cancelled, or delivered and never used): 76%  52%

Based on an 8-year survey of 50,000 software projects by the Standish Group.

Standish Group: “Chaos Manifesto 2013: Think Big, Act Small”
why re-architect LOCKSS?

• reduce support + operations costs
  • leverage web-scale open-source software
  • align w/ web archiving mainstream
• de-silo components + enable external integration
  • metadata extraction
  • archive access via DOI + OpenURL
  • polling + repair protocol
• prepare to evolve w/ the Web
  • web services architecture as flexible foundation
integration opportunities

• polling + repair
  • repository replication layer
  • other distributed digital preservation systems

• access
  • Dockerized full-text search for web archives
  • DOI + OpenURL access to web archives

• metadata extraction
aligning with web archiving

Web ARChive (WARC) format

compatible technologies

• Heritrix
• OpenWayback
• WarcBase
• Web Archiving Proxy
web archiving system APIs (WASAPI)

National Digital Platform Projects funded in August 2015

Systems Interoperability and Collaborative Development for Web Archiving (LG-71-15-0174-15): The Internet Archive, working with partner organizations University of North Texas, Rutgers University, and Stanford University Library will undertake a two-year research project to explore techniques that can expand national web archiving capacity in several areas.
leveraging community components
development progress

• access WARC-stored content via:
  • DOI
  • OpenURL
  • URL
  • Solr full-text search

• web services:
  • metadata extraction
  • metadata database
product roadmap

• 2017
  • Docker-ize components
  • web harvest framework
  • polling + repair web service
  • release to PLNs

• 2018
  • IP address + Shibboleth access via OpenWayback
  • OpenWayback format negotiation framework
  • full-text search web service
  • release to GLN
LOCKSS Networks

“Railroad Wye Switch” by Noel Hankamer under CC BY-NC-SA 2.0
Controlled LOCKSS (CLOCKSS)

• what is it?
  • library/publisher partnership
  • preserve the scholarly record
  • 12 globally-distributed nodes
  • dark until no longer accessible
  • triggered content world-accessible

• looking forward
  • expand capacity
  • increase pursuit of long tail
  • champion standards to simplify archiving (e.g., Signposting)
Private LOCKSS Networks (PLNs)

• what are they?
  • community of interest
  • jointly designate content
  • run distributed nodes
  • establish governance
  • preservation via diverse technologies, institutions, networks

• looking forward
  • create documentation
  • enable self-setup
  • support community collaboration
  • preserve web archives
national networks

• what are they?
  • in-country preservation
  • local stewardship
  • perpetual access
  • non-consumptive use

• looking forward
  • more networks
  • preserving national long-tail content
Distributed Preservation
distributed preservation landscape

• better understanding of role of distributed dark archives
• next logical step beyond mature local preservation
• appealing option for those w/o mature local preservation
a greater role for LOCKSS?

• bolster existing efforts
• undergird PLN service providers
• mainstream distributed digital preservation
LOCKSS for web archiving

• growth in web archiving
• centralization in web archiving
• native WARC support
• logical complement for web archive preservation
reliance on service provider

NDSA: “2016 NDSA Web Archiving Survey”
flat data transfer trend

NDSA: "2016 NDSA Web Archiving Survey"
Recap
vision

• better ensure the preservation of web archives
• LOCKSS team more actively engaged in community-supported development efforts
• communities enabled to more easily contribute to LOCKSS software, or run it w/o our help
• a longer tail of institutions able to capitalize on distributed digital preservation
• LOCKSS components applied in contexts other than LOCKSS networks
Questions