



**LOTS OF COPIES KEEP STUFF SAFE**

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

Nicholas Taylor ([@nullhandle](#))

Program Manager for [LOCKSS](#) and

[Web Archiving](#)

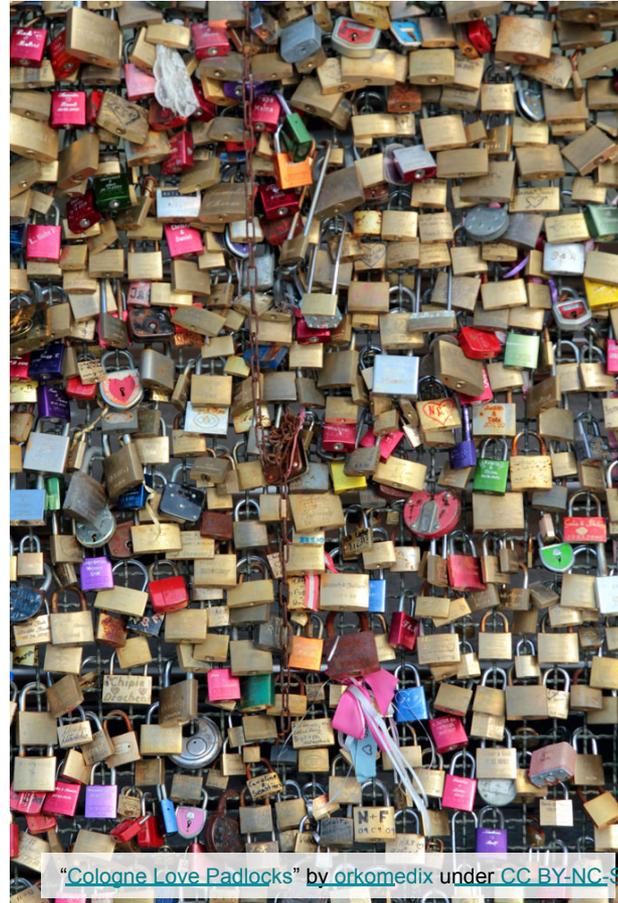
[Stanford University Libraries](#)

[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



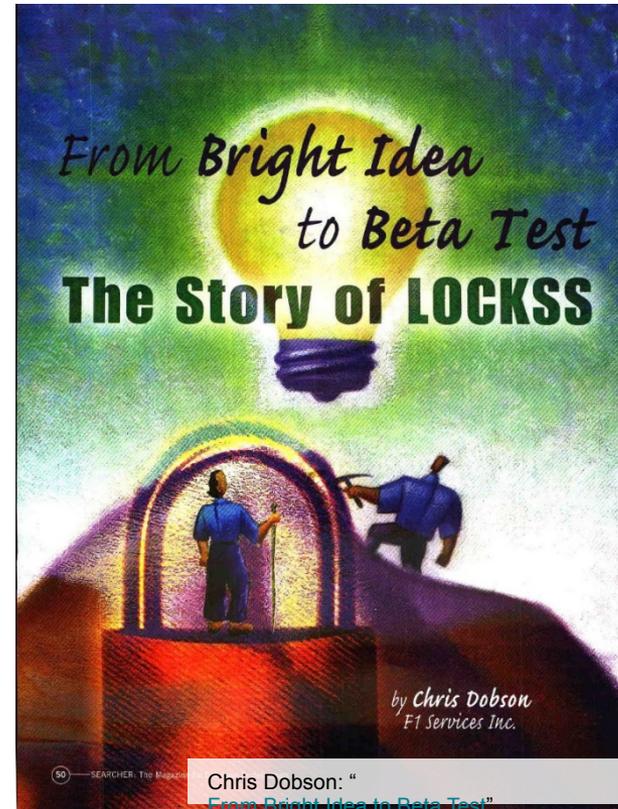
"Cologne Love Padlocks" by orkomedix under CC BY-NC-SA 2.0



# Program History

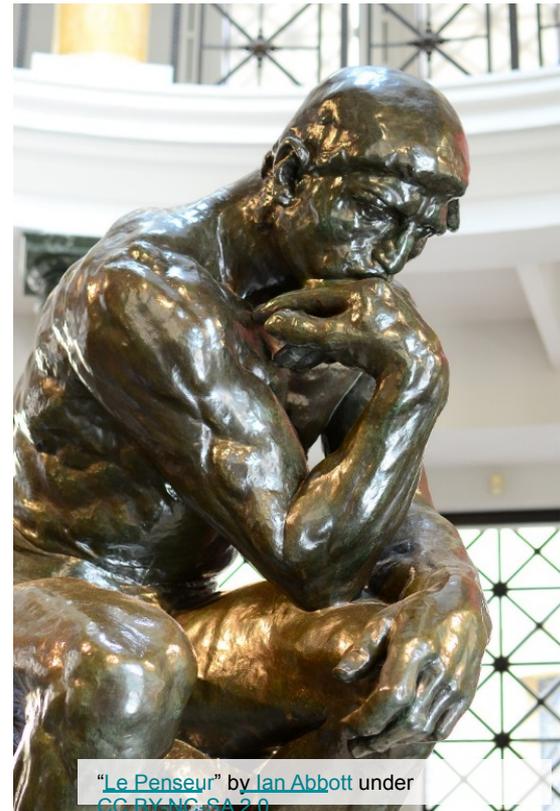
# inception

- 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



# philosophy + focus

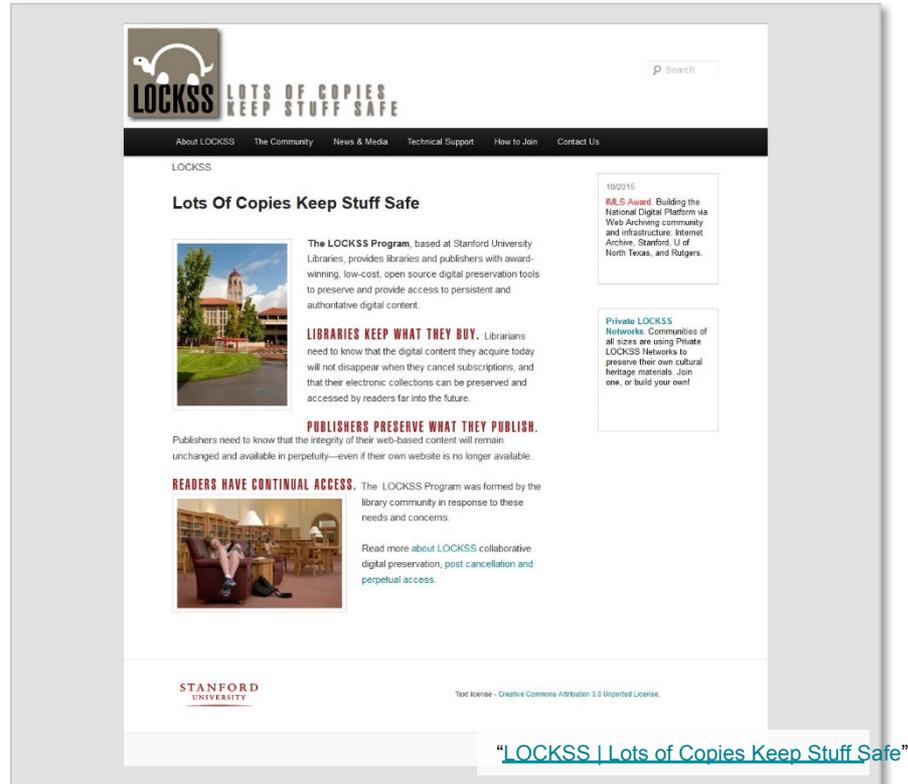
- lots of **copies** keep stuff safe
- preservation is an **active** community effort
- lots of **communities** keep stuff safe
- enable communities to preserve + access **their scholarly record**



["Le Penseur"](#) by [Jan Abbott](#) under [CC BY-NC-SA 2.0](#)

# present day

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



The screenshot displays the LOCKSS website interface. At the top left is the LOCKSS logo, which features a stylized white bird with its wings spread, perched on a black horizontal bar. To the right of the logo is the tagline "LOTS OF COPIES KEEP STUFF SAFE" in a bold, black, sans-serif font. Below the logo and tagline is a navigation menu with links for "About LOCKSS", "The Community", "News & Media", "Technical Support", "How to Join", and "Contact Us". A search bar is located in the top right corner.

The main content area is titled "LOCKSS" and features the article "Lots Of Copies Keep Stuff Safe". The article includes three main sections, each with a small image and a brief description:

- LIBRARIES KEEP WHAT THEY BUY.** Accompanied by an image of a large, multi-story brick building with a central tower. The text states: "Librarians need to know that the digital content they acquire today will not disappear when they cancel subscriptions, and that their electronic collections can be preserved and accessed by readers far into the future."
- PUBLISHERS PRESERVE WHAT THEY PUBLISH.** Accompanied by an image of a person sitting in a chair reading a book in a library. The text states: "Publishers need to know that the integrity of their web-based content will remain unchanged and available in perpetuity—even if their own website is no longer available."
- READERS HAVE CONTINUAL ACCESS.** Accompanied by an image of a person sitting in a chair reading a book in a library. The text states: "The LOCKSS Program was formed by the library community in response to these needs and concerns."

On the right side of the article, there are two smaller text boxes:

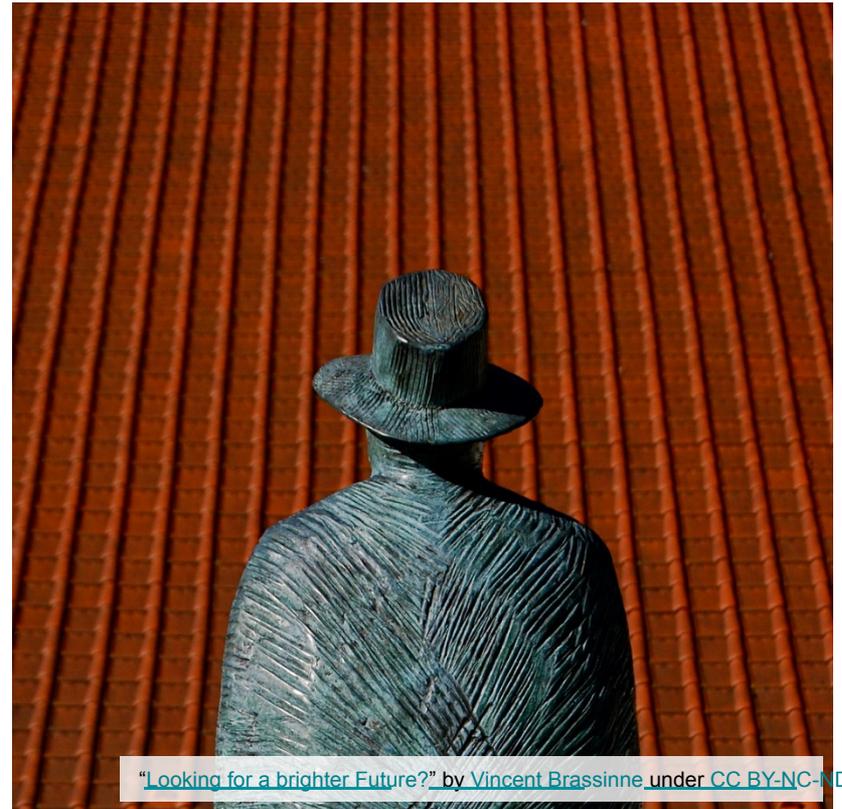
- 10/20/15** **MLS Award:** Building the National Digital Platform via Web Archiving community and infrastructure. Internet Archive, Stanford, U of North Texas, and Rutgers.
- Private LOCKSS Networks:** Communities of all sizes are using Private LOCKSS Networks to preserve their own cultural heritage materials. Join one, or build your own!

At the bottom of the page, the Stanford University logo is visible on the left, and the text "Text license - Creative Commons Attribution 3.0 Unported License" is on the right. A blue link at the bottom right of the screenshot reads "LOCKSS | Lots of Copies Keep Stuff Safe".



# looking forward

- organizational changes
- software evolution
- LOCKSS networks
- distributed digital preservation



["Looking for a brighter Future?"](#) by Vincent Brassinne under [CC BY-NC-ND 2.0](#)

A photograph of two LEGO minifigures on a dark gravel track. The minifigure on the left is wearing a white shirt, red pants, and a large brown wig. It is handing a white baton to the minifigure on the right, which is wearing a white shirt, red pants, and a red and orange wig. The background shows a green field and trees under a clear sky.

# Organizational Changes

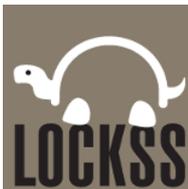
# David + Vicky



[American Library Association: "Victoria Reich and David S.H. Rosenthal"](#)

# 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



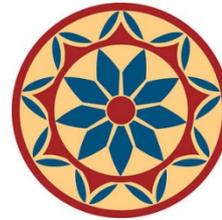
# LOCKSS + DLSS administrativa

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

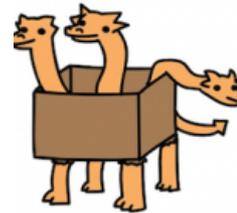


# LOCKSS + DLSS synergies

- 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



blacklight



Fedora™

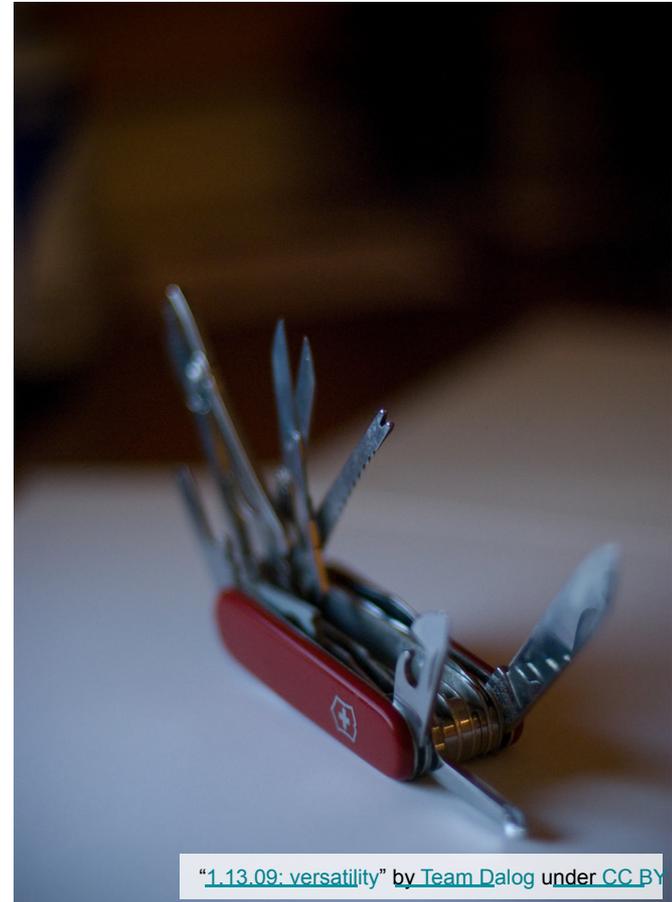




# Software Evolution

# 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

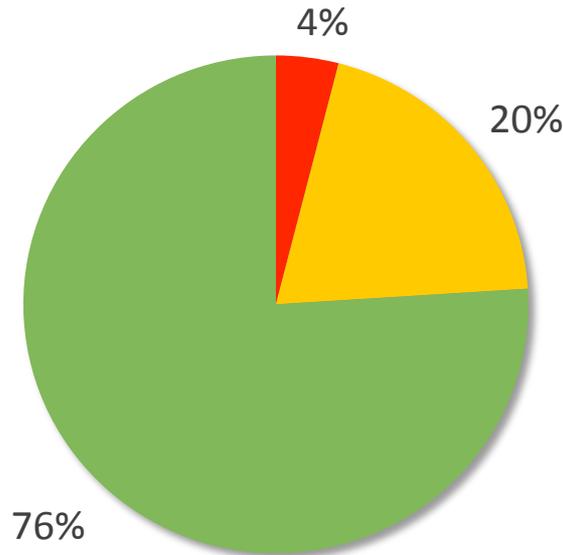
1. “All teams will henceforth **expose their data and functionality through service interfaces.**”
2. “Teams **must communicate** with each other **through these interfaces.**”
3. “There will be **no other form of interprocess communication allowed**: no direct linking, no direct reads of another team's 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”](#)

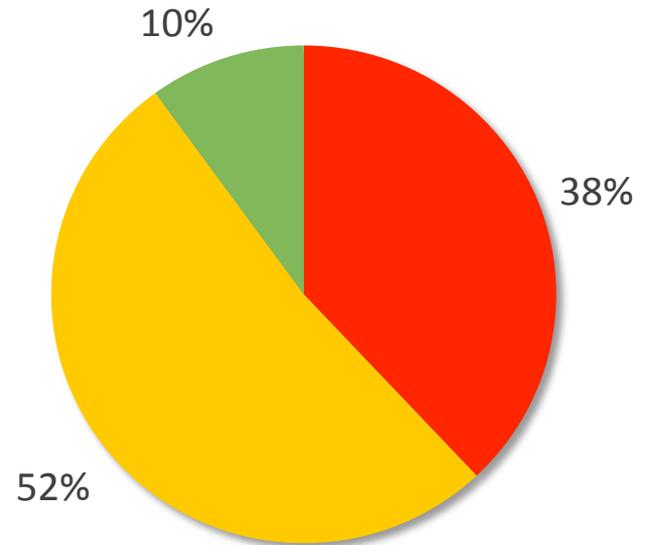


# risk of large projects

small projects (< \$1 million)



large projects (> \$10 million)



**successful**  
(on time,  
on budget)

**challenged**  
(late, over budget,  
lacking functionality)

**failed** (cancelled,  
or delivered  
and never used)

Based on an 8-year survey of 50,000 software projects by the [Standish Group](#).

[Standish Group: "Chaos Manifesto 2013: Think Big, Act Small"](#)



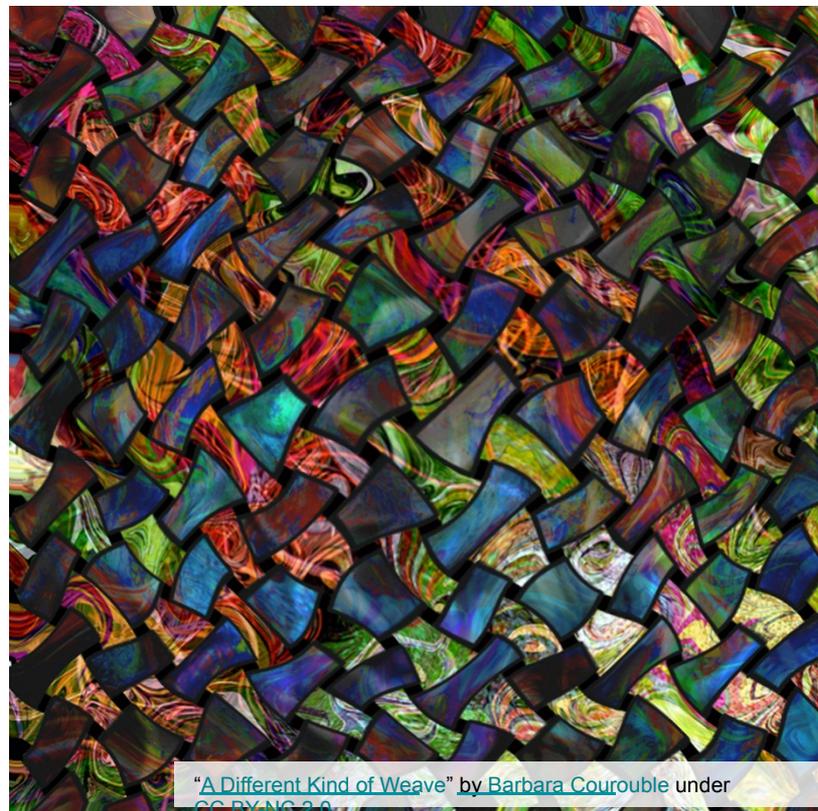
# why re-architect LOCKSS?

- 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

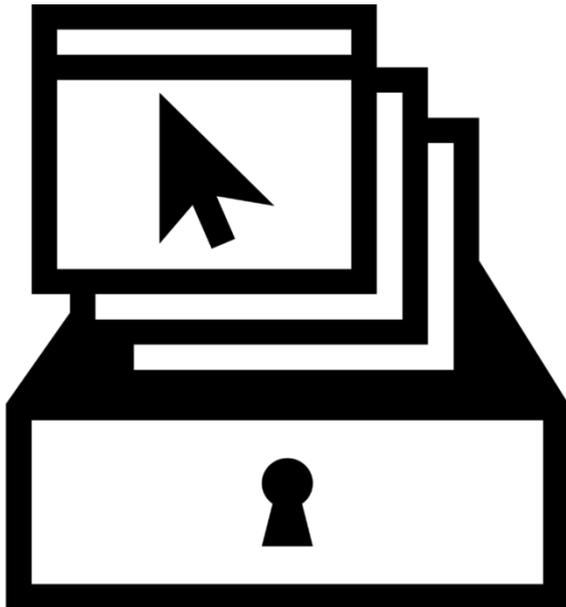
- polling + repair
  - repository replication layer
  - other distributed digital preservation systems
- access
  - Dockerized full-text search for web archives
  - DOI + OpenURL access to web archives
- metadata extraction



["A Different Kind of Weave"](#) by [Barbara Courouble](#) under [CC BY-NC 2.0](#)

# aligning with web archiving

**Web ARChive (WARC)  
format**



**compatible technologies**

- Heritrix
- OpenWayback
- WarcBase
- Web Archiving Proxy

# web archiving system APIs (WASAPI)

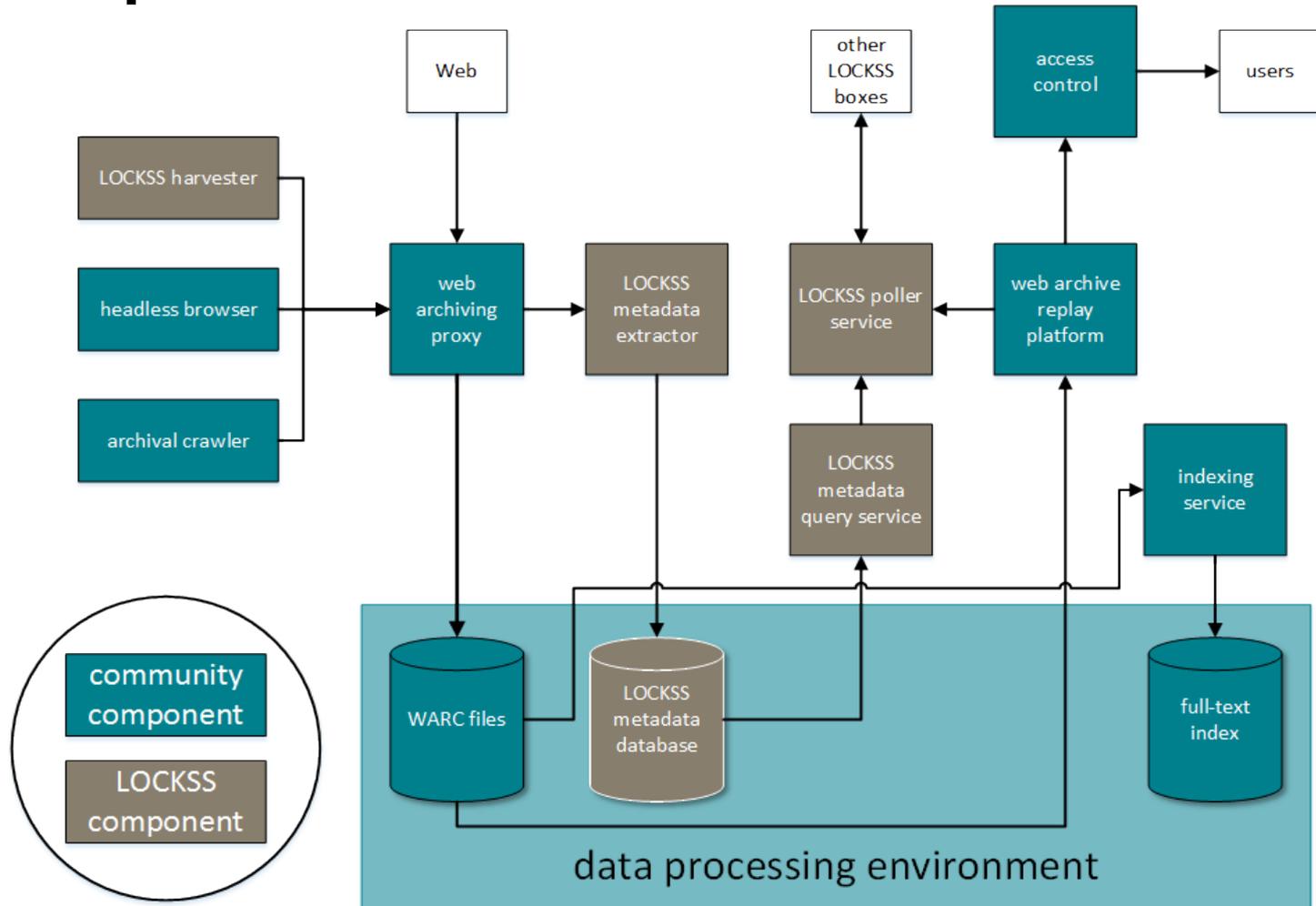
## National Digital Platform Projects funded in August 2015

### 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



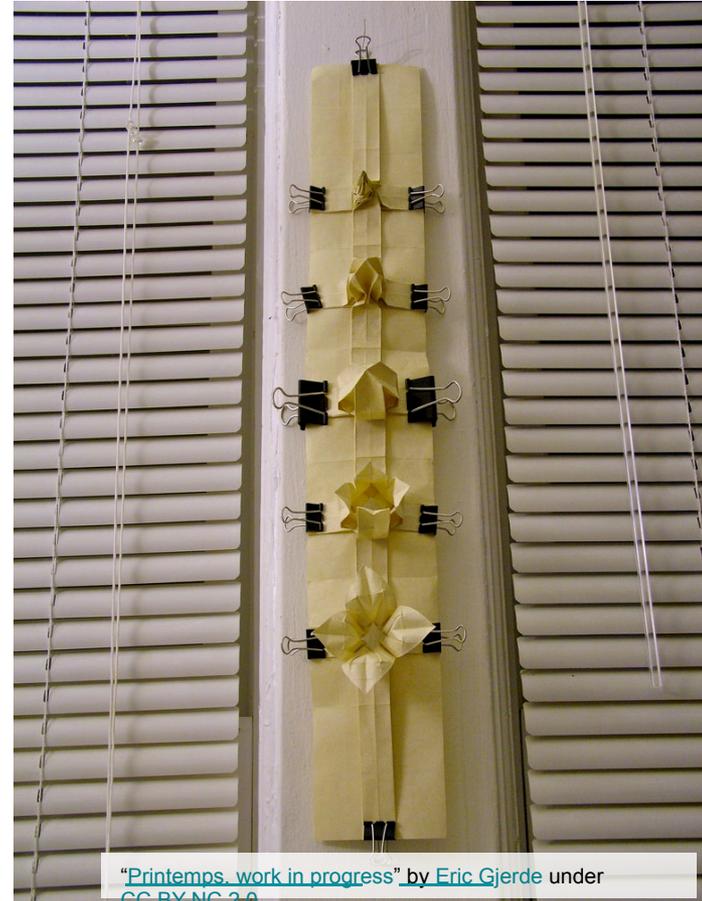
# development progress

- access WARC-stored content via:
  - DOI
  - OpenURL
  - URL
  - Solr full-text search
- web services:
  - metadata extraction
  - metadata database

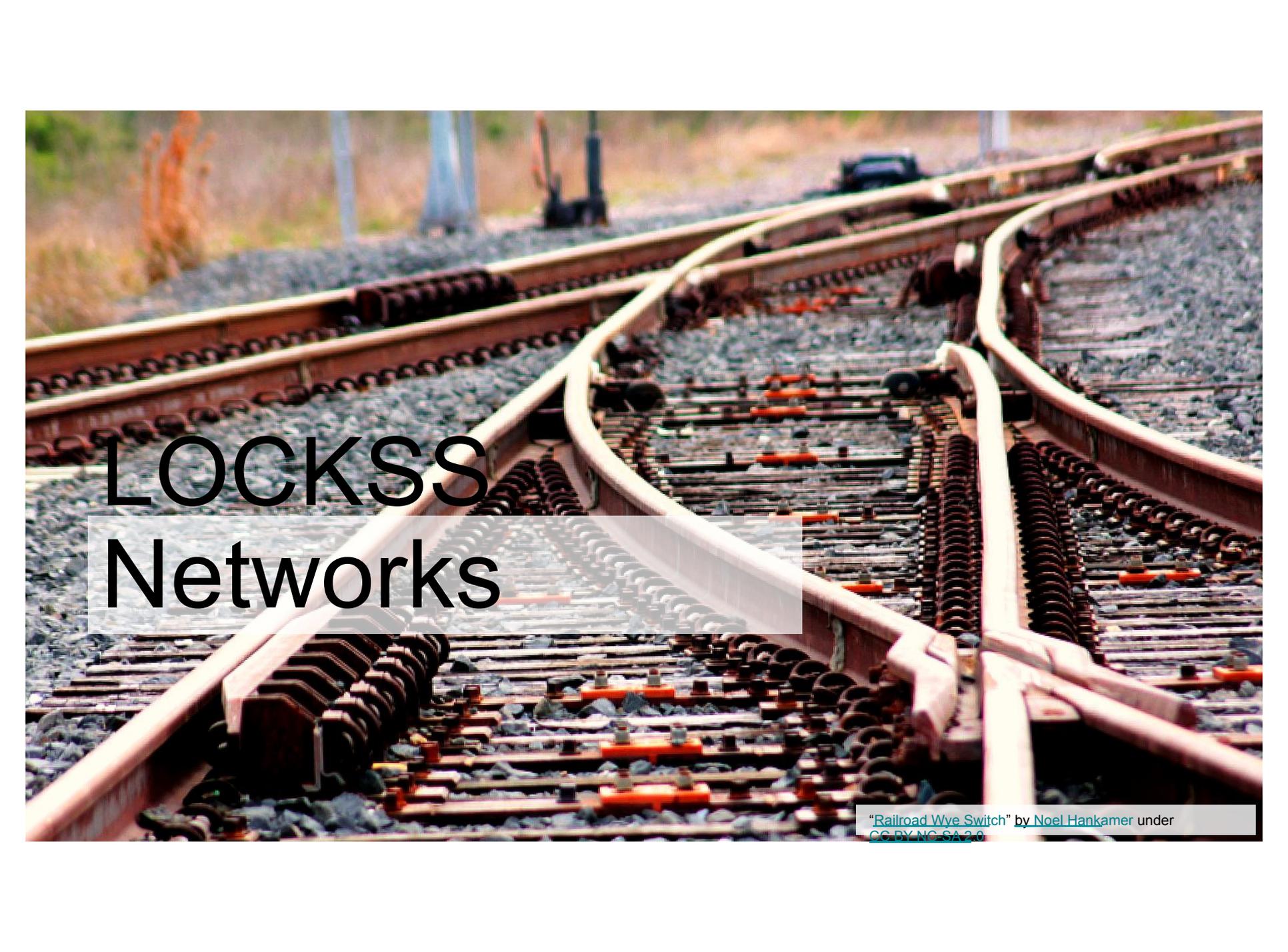


# product roadmap

- 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



["Printemps, work in progress"](#) by [Eric Gjerde](#) under [CC BY-NC 2.0](#)



# 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](#))



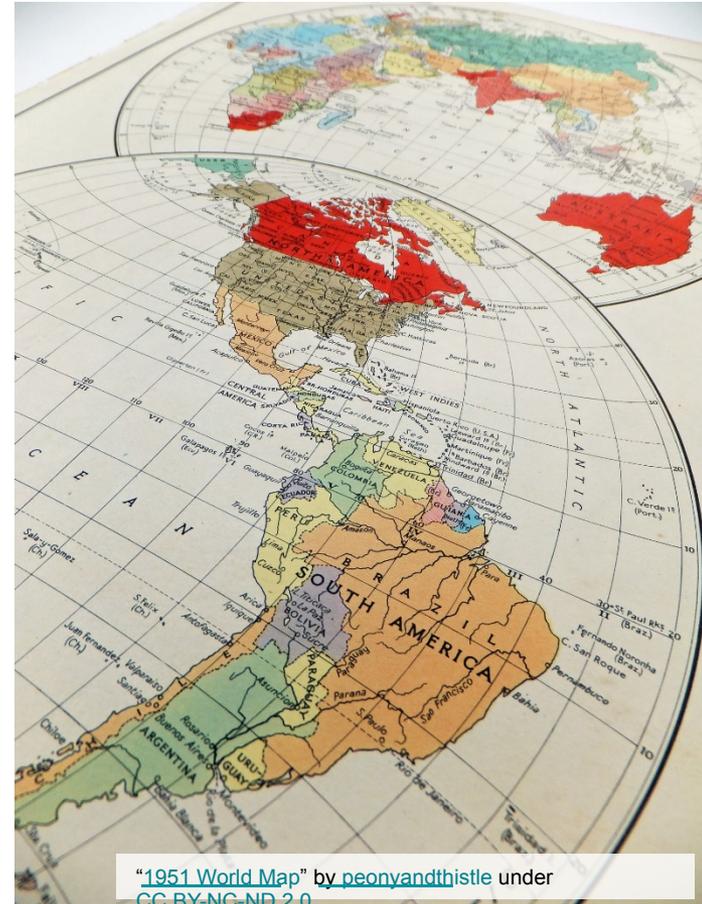
# Private LOCKSS Networks (PLNs)

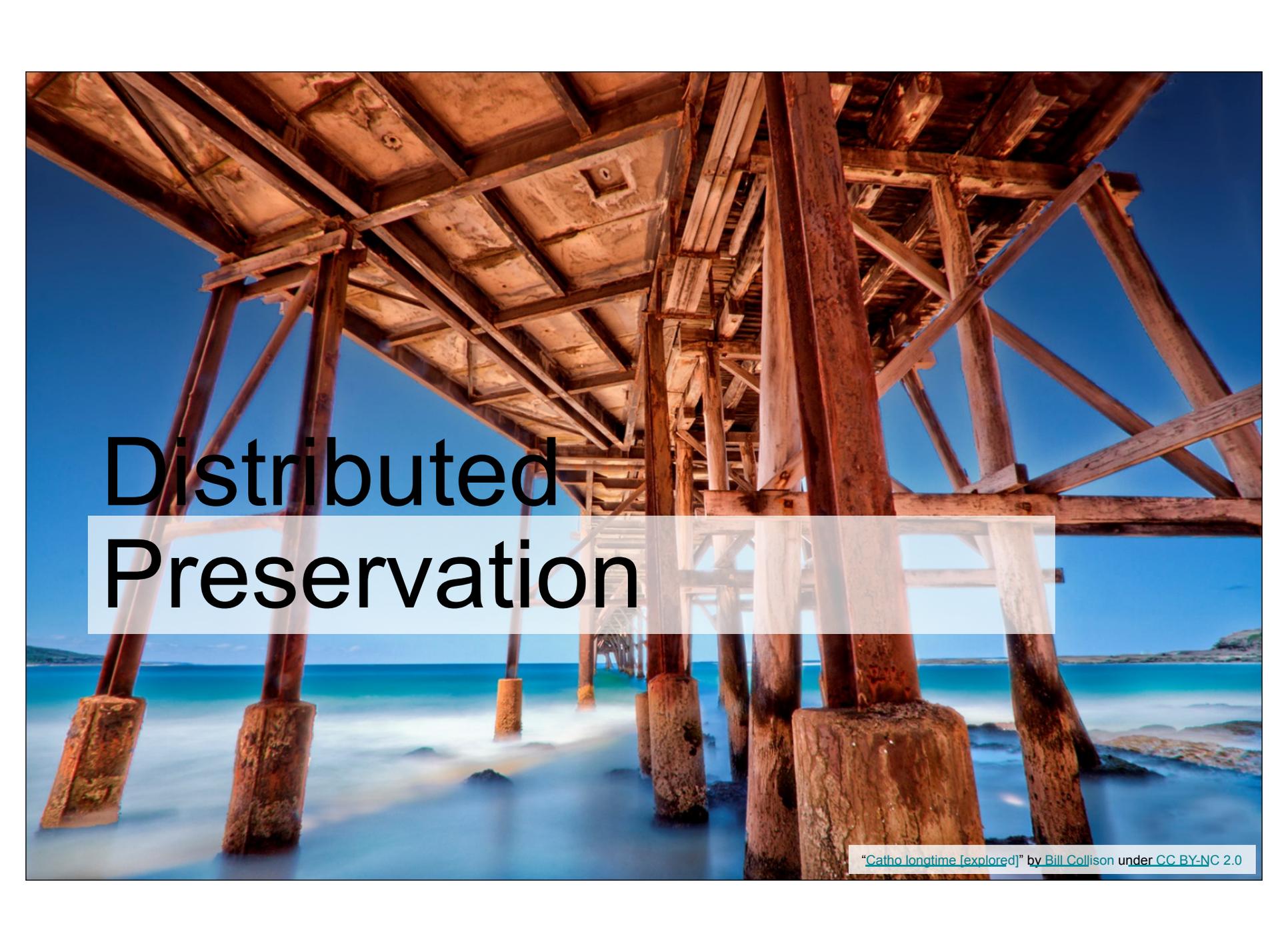
- what are they?
  - community of interest
  - jointly designate content
  - run distributed nodes
  - establish governance
  - preservation via diverse technologies, institutions, networks
- looking forward
  - create **documentation**
  - enable **self-setup**
  - support **community collaboration**
  - 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**



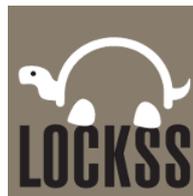


# Distributed Preservation

"Catho longtime [explored]" by Bill Collison under CC BY-NC 2.0

# 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



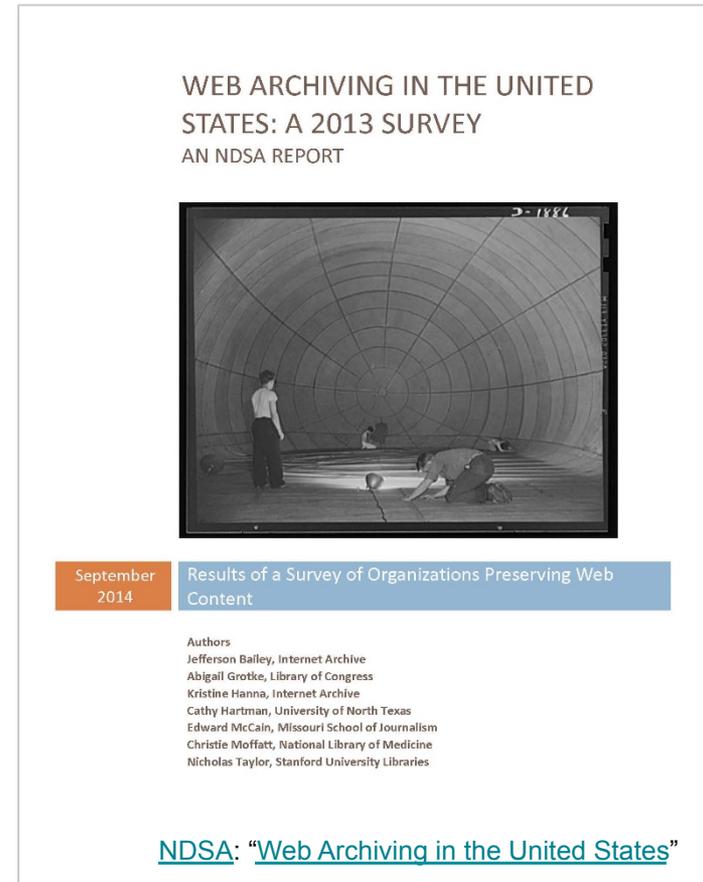
# a greater role for LOCKSS?

- bolster existing efforts
- undergird PLN service providers
- mainstream distributed digital preservation

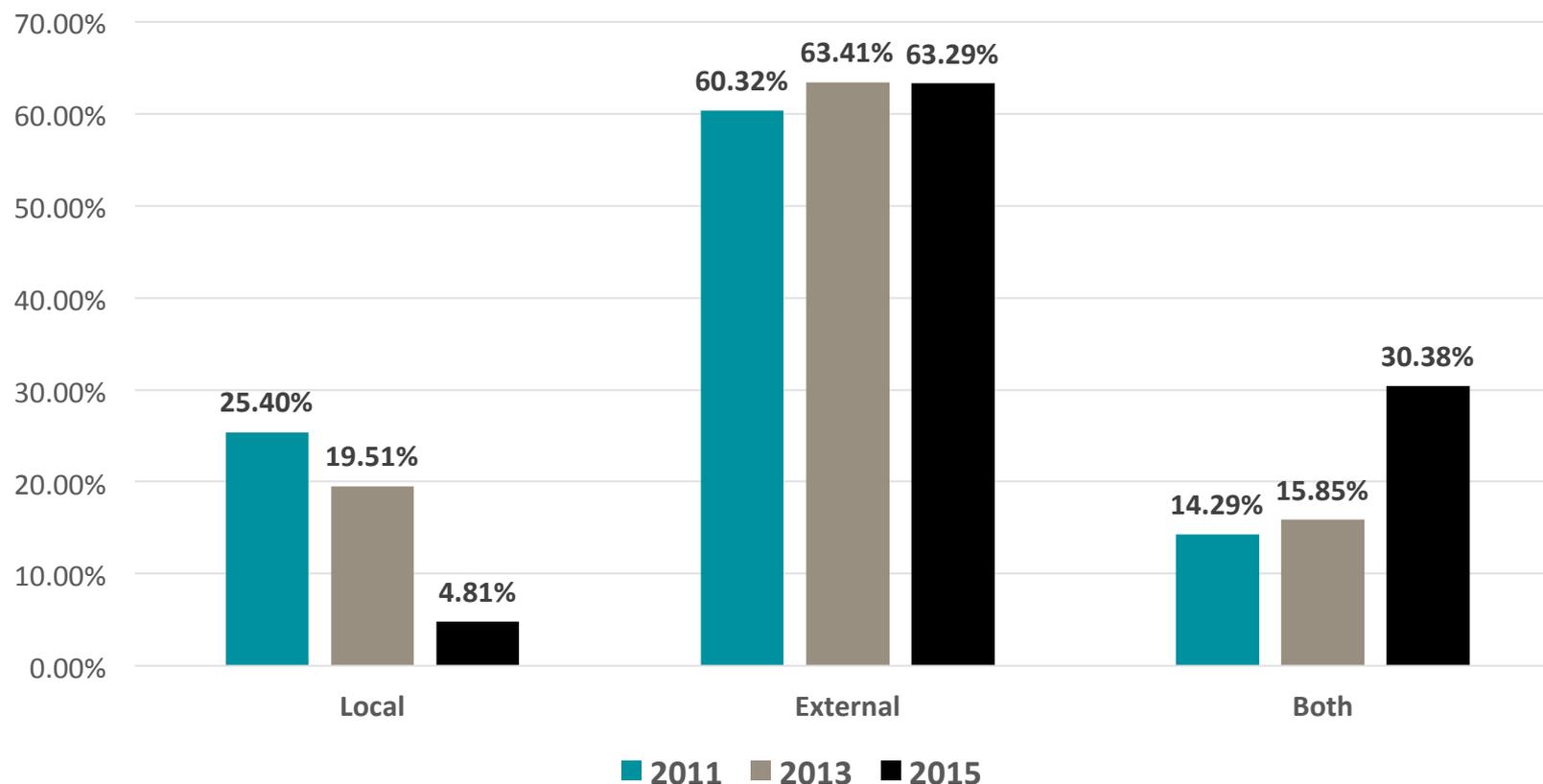


# LOCKSS for web archiving

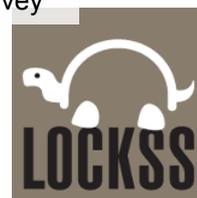
- growth in web archiving
- centralization in web archiving
- native WARC support
- logical complement for web archive preservation



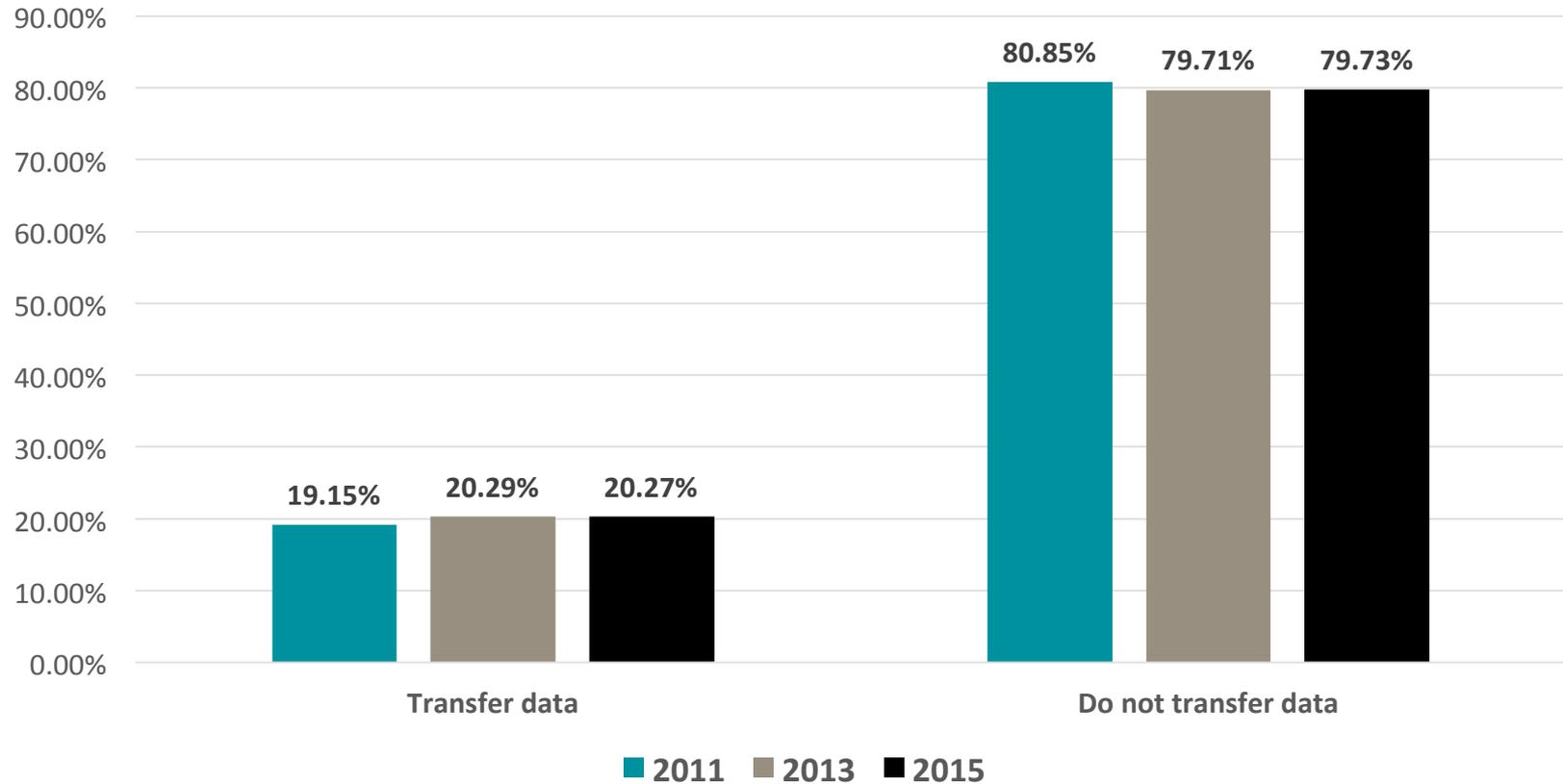
# reliance on service provider



[NDSA](#): "2016 NDSA Web Archiving Survey"



# flat data transfer trend



[NDSA](#): "2016 NDSA Web Archiving Survey"





Recap

p

# 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**



A large radio telescope dish is silhouetted against a sunset sky. The dish is on the left side of the frame, and the sun is setting on the right horizon, creating a bright glow. The sky is filled with wispy clouds. The word "Questions" is written in a large, black, serif font across the middle of the dish.

Questions

A white rectangular box with rounded corners is positioned over the lower part of the telescope dish. Inside the box is a large, black, serif question mark.

?