



LOTS OF COPIES KEEP STUFF SAFE

Lots More LOCKSS for Web Archiving: Boons from the LOCKSS Software Re-Architecture

Nicholas Taylor ([@nullhandle](#))

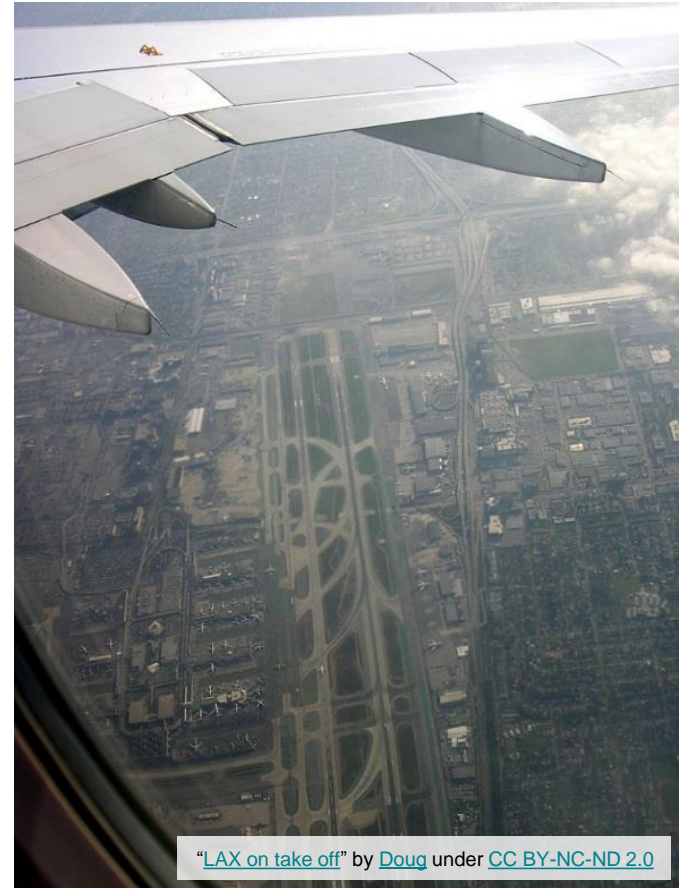
Program Manager, [LOCKSS](#) and [Web Archiving](#)
[Stanford University Libraries](#)

[Web Archiving Week](#)

15 June 2017

overview

- LOCKSS background
- software re-architecture
- software components
- roadmap



"LAX on take off" by [Doug](#) under [CC BY-NC-ND 2.0](#)

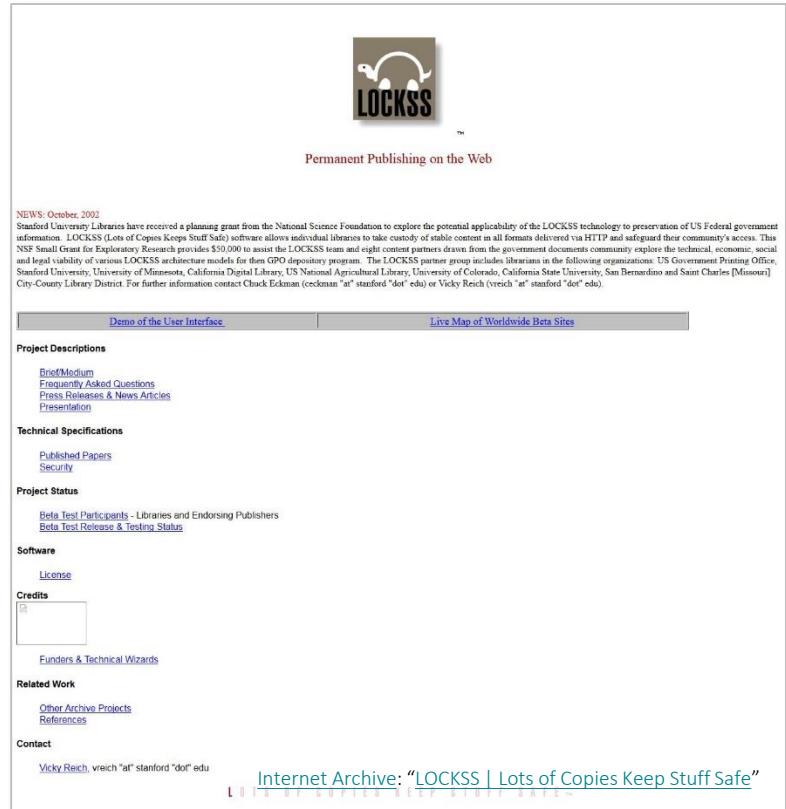


LOCKSS Background

"Padlocks1" by [Domiriel](#) under [CC BY-NC 2.0](#)

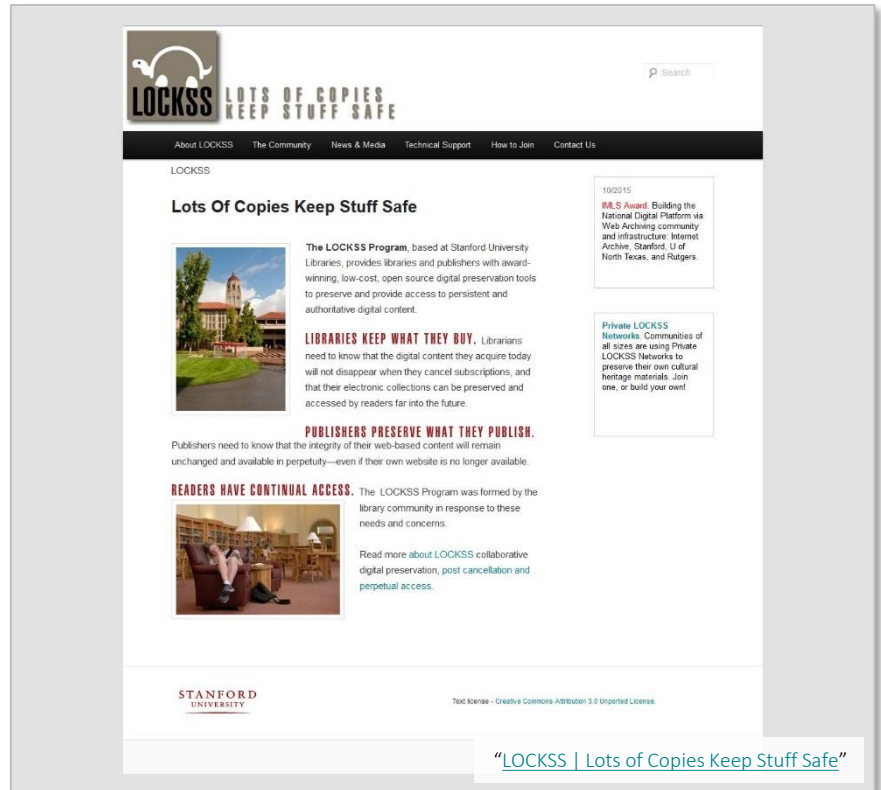
beginnings

- 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



present day

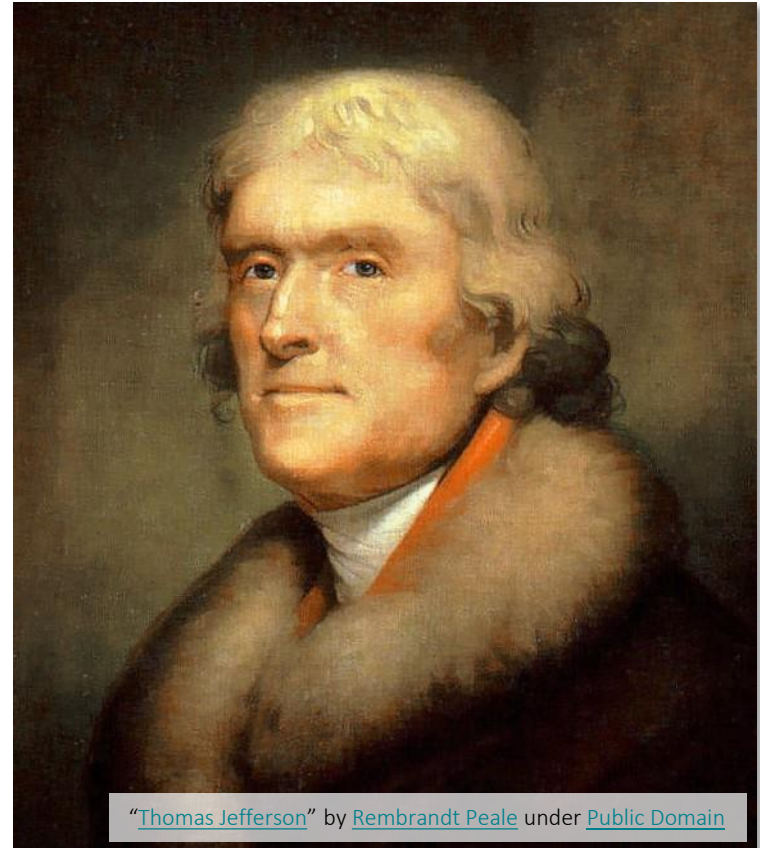
- financially self-sustaining
- tens of networks
- hundreds of institutions
- all types of content



lots of copies

“The lost cannot be recovered; but let us save what remains: not by vaults and locks which fence them from the public eye and use, in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.”

National Archives: [“from Thomas Jefferson to Ebenezer Hazard, 18 February 1791”](#)



“Thomas Jefferson” by Rembrandt Peale under [Public Domain](#)

decentralized copies

- no monopoly on copy-making
- independent, de-correlated copies
- no central point of failure or vulnerability
- local custody, self-determination



"Domino's" by [david pacey](#) under [CC BY 2.0](#)

articulated threat model


- long-term **bit integrity** is a hard problem
- more (correlated) copies **doesn't necessarily** keep stuff safe
- don't underestimate:
 - people making **mistakes**
 - **attacks** on information
 - organizational **failure**



community-validated

- built upon peer-reviewed research
- successfully operating for almost 20 years
- CRL TRAC assessment of CLOCKSS
 - overall score matching previous best
 - only perfect technology score awarded to date



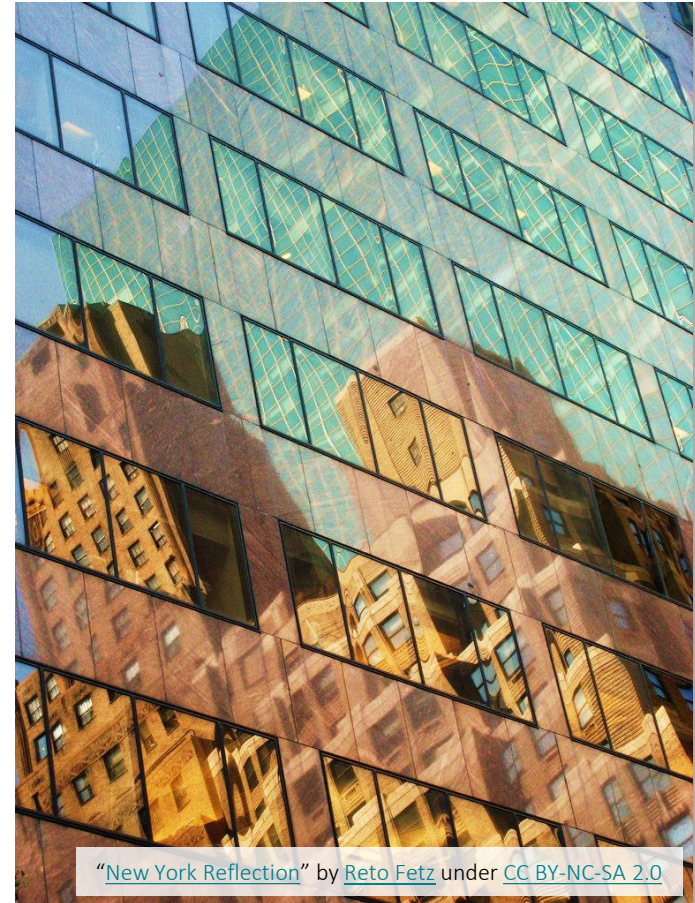


Software Re-Architecture

"The two bridges" by [Frank Schulenburg](#) under [BY-SA 2.0](#)

why re-architect LOCKSS?

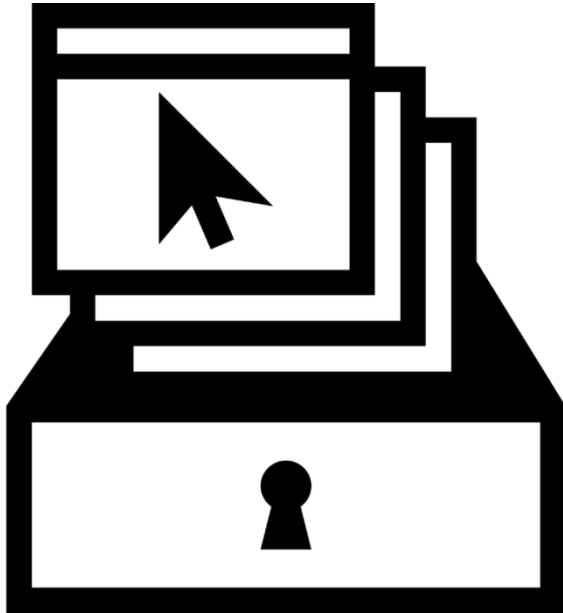
- reduce support + operations costs
- de-silo components + enable external integration
- prepare to evolve w/ the Web



"New York Reflection" by [Reto Fetz](#) under [CC BY-NC-SA 2.0](#)

aligning with web archiving

Web ARChive (WARC) format



compatible technologies

- Heritrix
- OpenWayback
- WarcBase
- Web Archiving Proxy

web archiving system APIs (WASAPI)

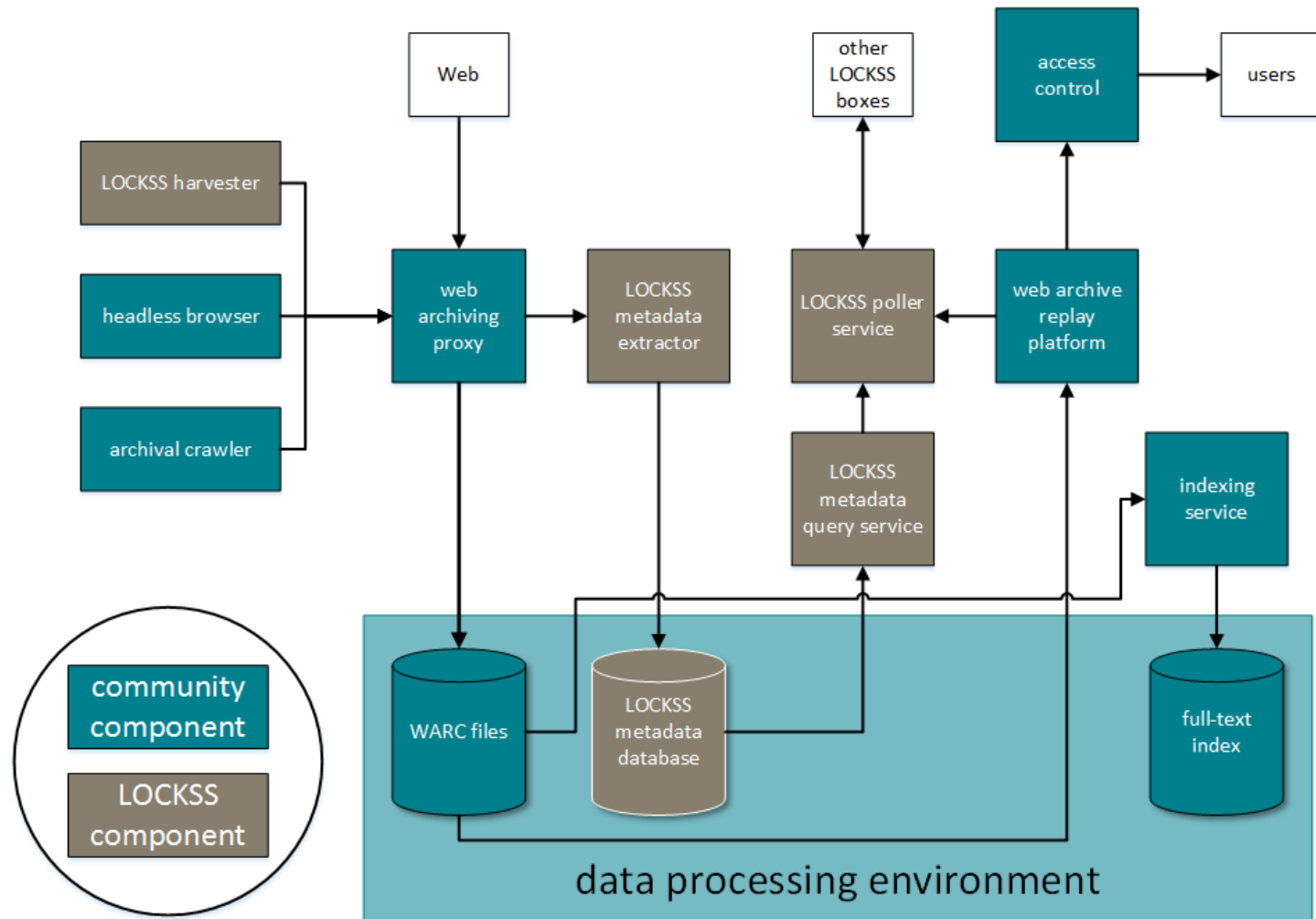
National Digital Platform Projects funded in August 2015

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





Software Components

bibliographic metadata extraction

functionality

- for web harvest + file transfer content
- map values in DOM tree to metadata fields
- retrieve downloadable metadata from expected URL patterns
- parse RIS + XML by schema

fields

- creator
- title
- published year
- volume
- issue
- article name

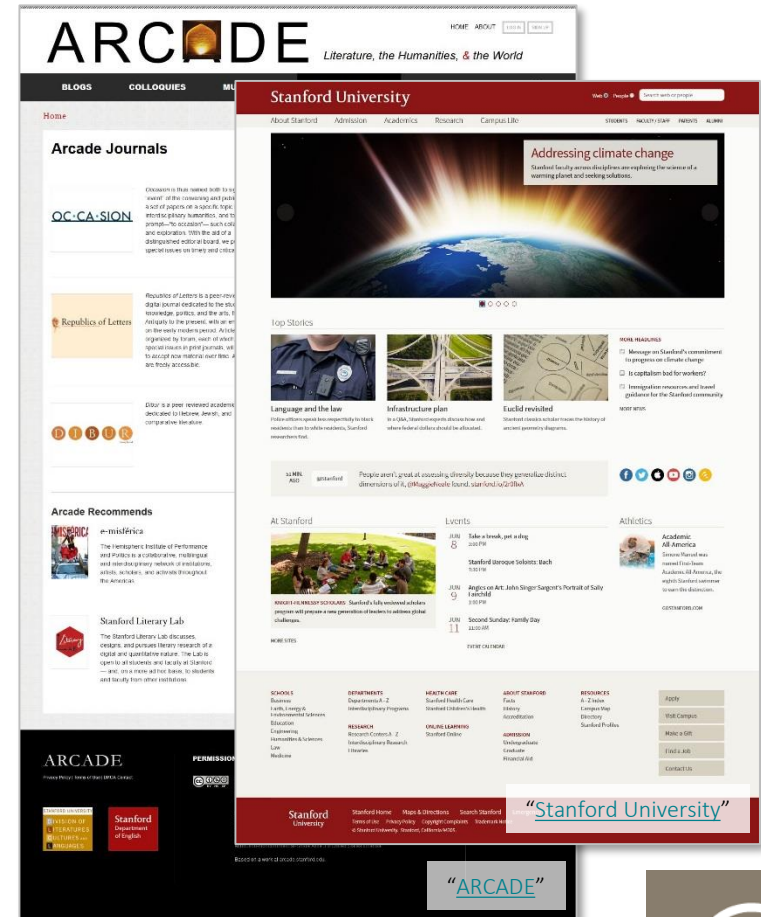
publisher/platform heuristics

- plugins making bibliographic objects + their metadata on many publishing platforms machine-intelligible
- a framework for creating such plugins
- useful in absence of standard conventions, e.g. [Signposting](#)

The logo for Atypen, featuring the word "Atypen" in a black serif font. A blue horizontal line with a crosshair-like symbol in the center passes through the letter 'p'.The logo for Digital Commons, consisting of a stylized white 'e' icon on a dark grey background, followed by the text "DIGITALCOMMONS" in white capital letters.The logo for HighWire, featuring a red circular icon made of dots on the left and the word "HighWire" in a black serif font on the right.The logo for Silverchair Information Systems, featuring a blue circular icon with horizontal lines above the text "SILVERCHAIR" in black, with "INFORMATION / SYSTEMS" in blue below.

use cases for metadata extraction

- apply to consistent subsets of content in larger corpora
- curate OA materials within broader crawls
- retrieve faculty publications posted online, license allowing
- describe sub-sites collected while self-archiving from a single institutional CMS



discovery via bibliographic metadata

- submit DOI or OpenURL query
- get OpenWayback access URLs
- integrate w/ OpenURL resolver

The screenshot shows the 'Find eJournal' page on the Stanford University Libraries website. The page has a header with the Stanford logo and 'STANFORD UNIVERSITY LIBRARIES'. Below the header, there are links for 'Home > eJournals', 'Find eJournal', and 'Find eBook'. The main search area is titled 'Citation Linker' and contains a form with various input fields and search options. The form includes a 'Journal Title' field with a search type selector (radio buttons for 'Starts with', 'Contains', and 'Exact', with 'Exact' selected). Below this is a date input field with a format example '1940-01-01' and a 'Go' button. Other fields include 'Volume', 'Issue', 'Start page', 'End page', 'ISSN', 'DOI', 'PMID', 'Author' (with sub-fields for 'Last name', 'First name', and 'Initials'), and 'Article title'. A 'Clear' button is located at the bottom right of the form. To the right of the form, there are links for 'Report a connection problem', 'Connect from off campus', and 'System status'. At the bottom right, there is a link for 'Other Stanford Libraries: Lane Medical Library, SLAC Research Library'. At the bottom of the page, there is a disclaimer about SUL licensed resources and a link to 'Stanford Libraries: "Find eJournal"'.

STANFORD UNIVERSITY LIBRARIES

Home > eJournals

Find eJournal Find eBook

By Title By Subject More Options Citation Linker

Use Citation Linker to find full text when you already know the citation for an article. Enter the journal's title (or ISSN, or DOI) and publication year. Fields with asterisks are recommended.

Journal Title * ☐ Starts with ☐ Contains ☒ Exact

Enter a date (format: 1940-01-01) or use the pulldown menus.

Date * OR year month day

Volume Issue

Start page * End page

ISSN DOI

PMID

Author Last name First name Initials

Article title

Report a connection problem

Connect from off campus

System status

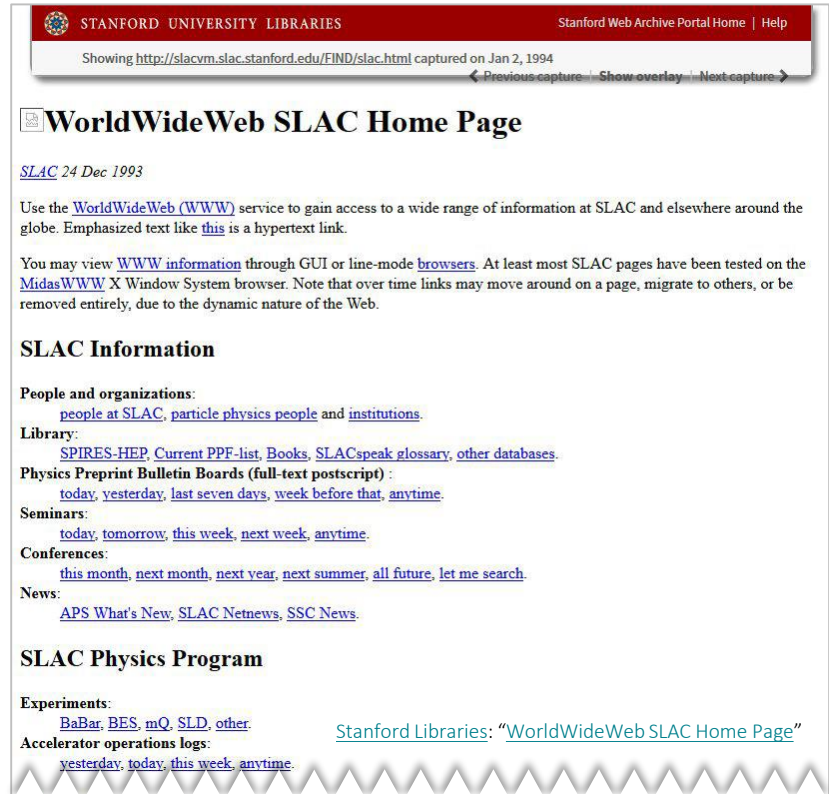
Other Stanford Libraries:
Lane Medical Library
SLAC Research Library

SUL licensed resources are subject to Terms of Use and contractual restrictions. Terms of Use may also be posted on the resource's website. Resources are for non-profit educational use of Stanford students, faculty, and staff. Systematic downloading, distribution, or retention of substantial portions of content is prohibited.

Stanford Libraries: "Find eJournal"

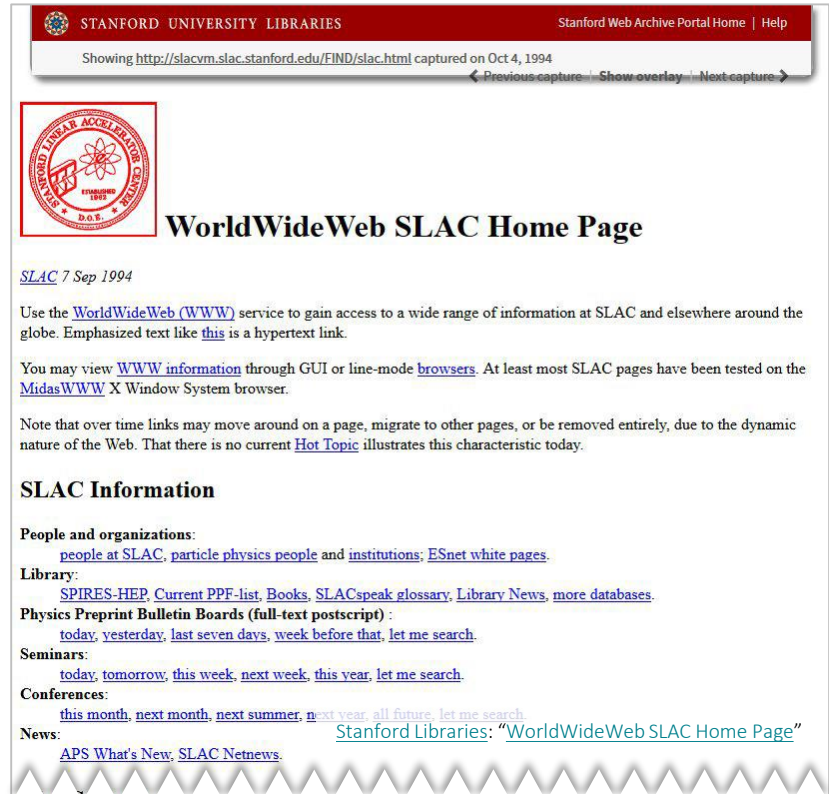
on-access format migration

- as described in [2005 D-Lib paper by DSHR et al](#)
- low obsolescence risk suggested by research from [Holden](#), [Jackson](#)
- implement upstream in OpenWayback
- example: X-BitMap → GIF migration



on-access format migration

- as described in [2005 D-Lib paper by DSHR et al](#)
- low obsolescence risk suggested by research from [Holden](#), [Jackson](#)
- implement upstream in OpenWayback
- example: X-BitMap → GIF migration

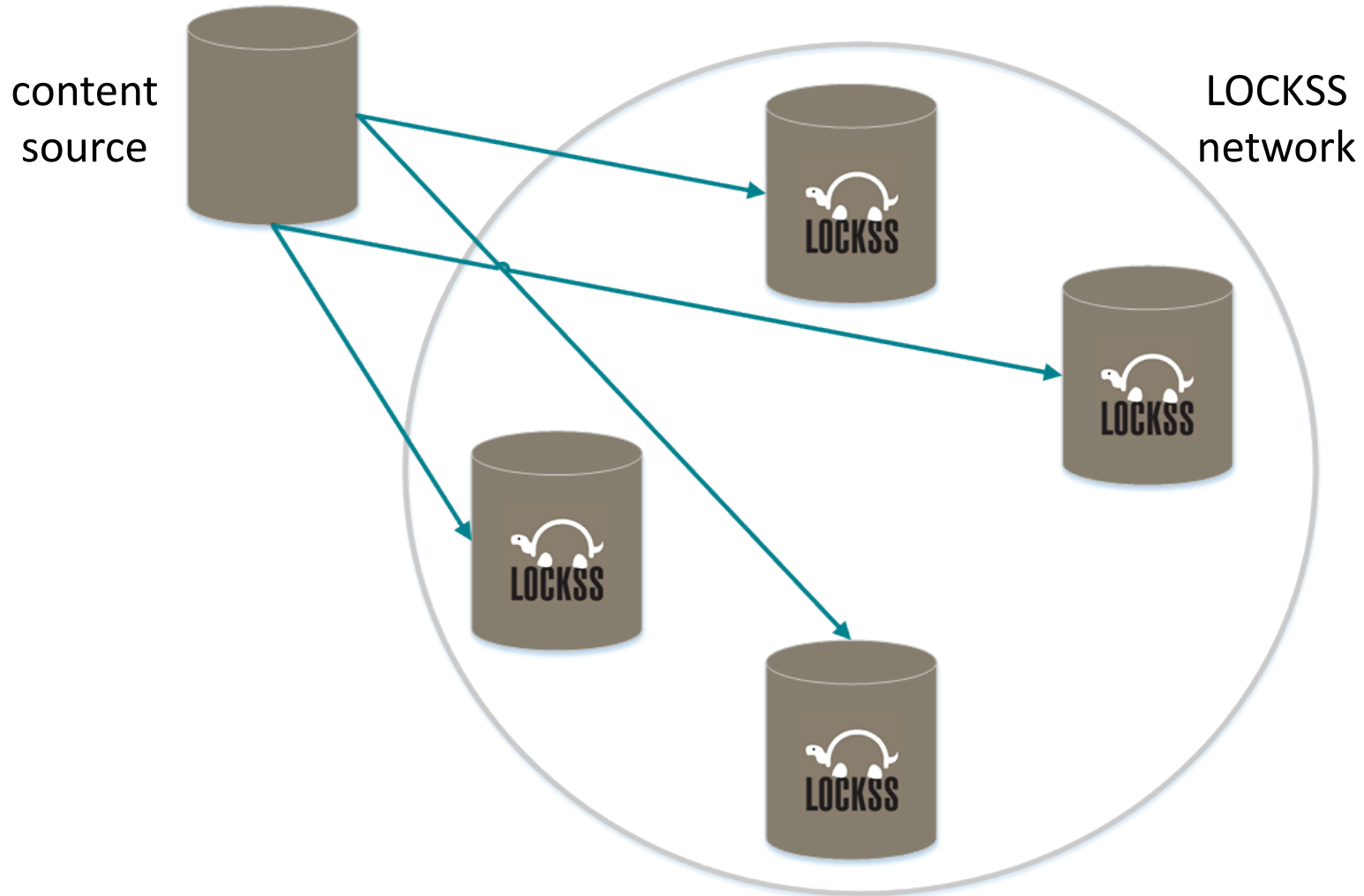


audit + repair protocol

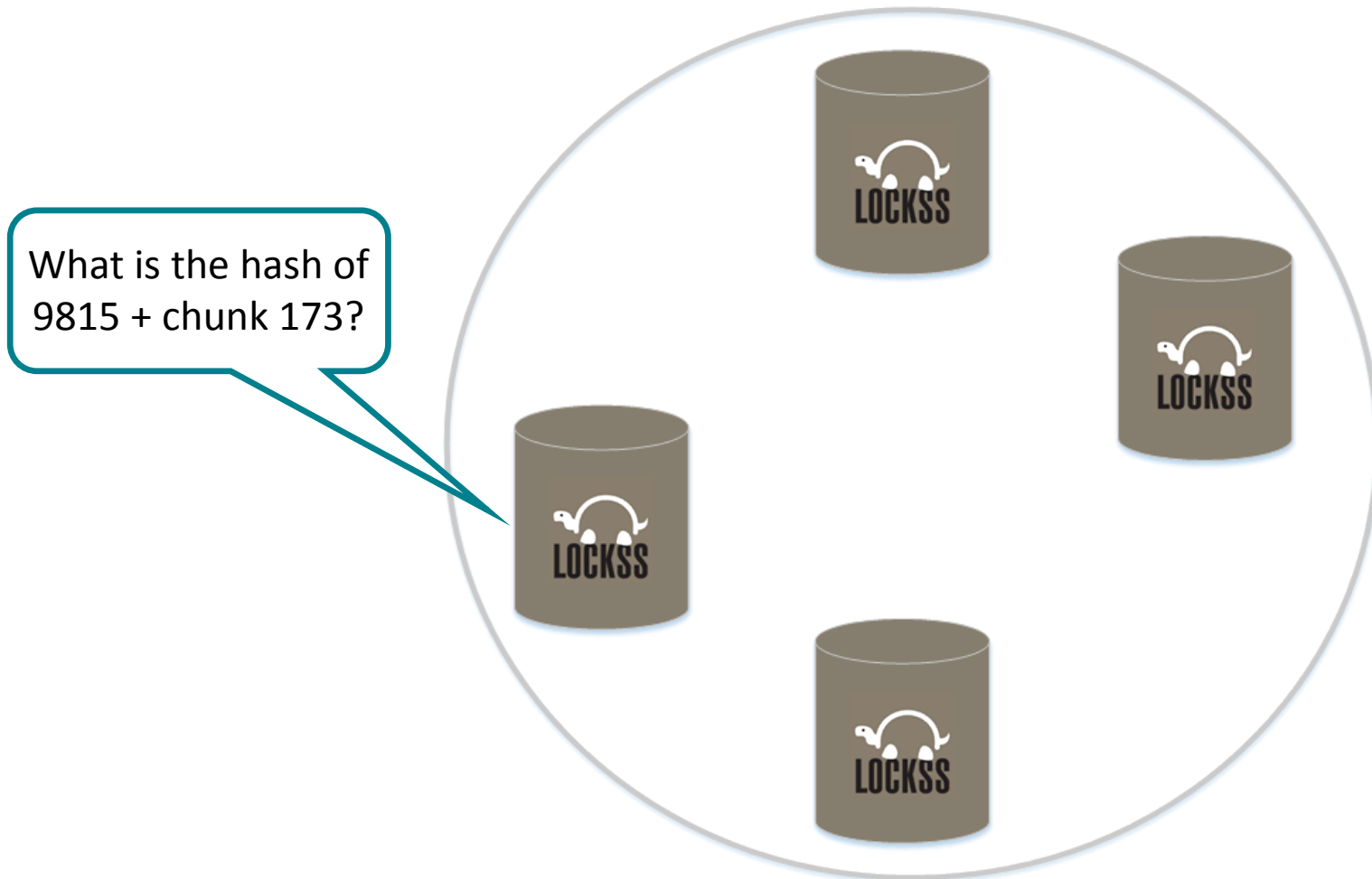
- core preservation capability
- network nodes conduct polls to validate integrity of distributed copies of data chunks
- more nodes = more security
 - more nodes can be down
 - more copies can be corrupted
 - ...and polls will still conclude
- nonces force re-hashing
- peers are untrusted
- polls are slow, to allow damage detection



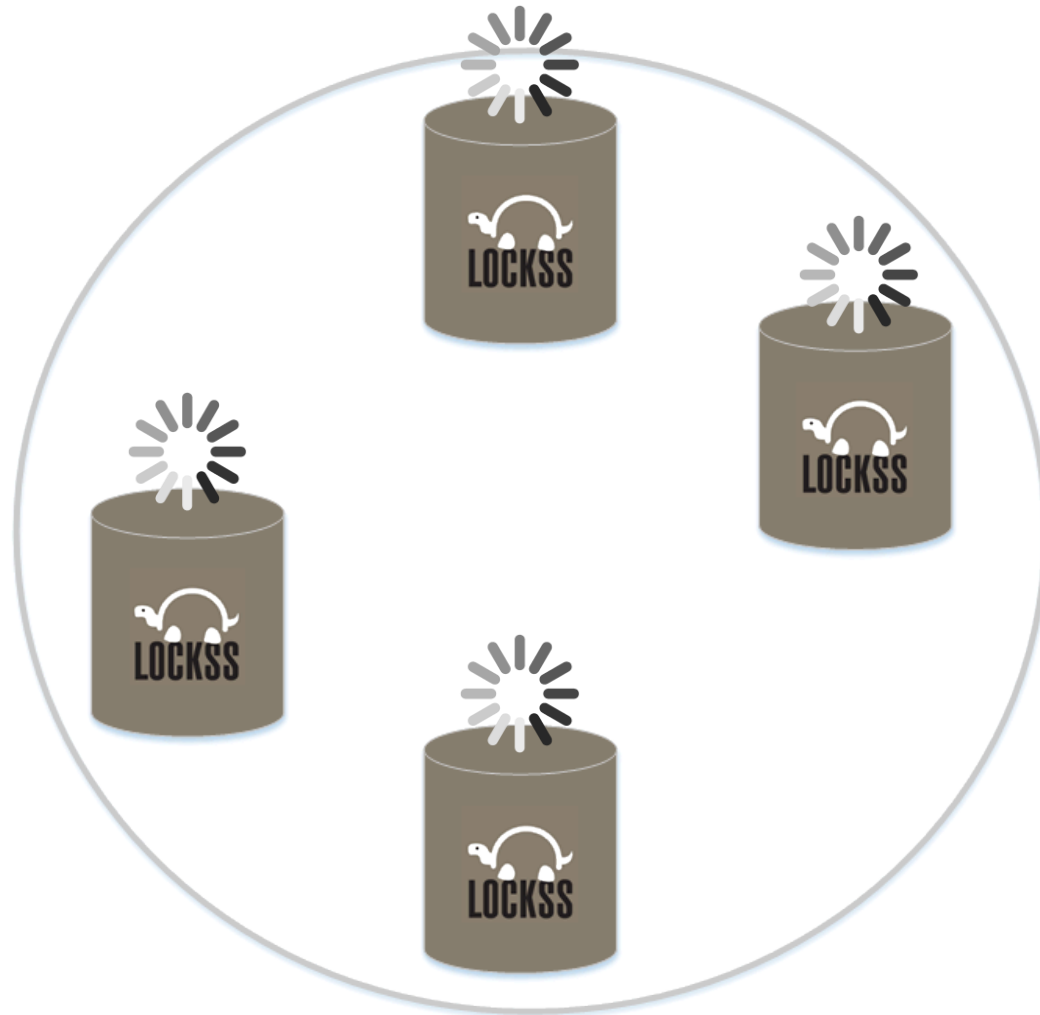
harvest from content source



initiate poll w/ nonce

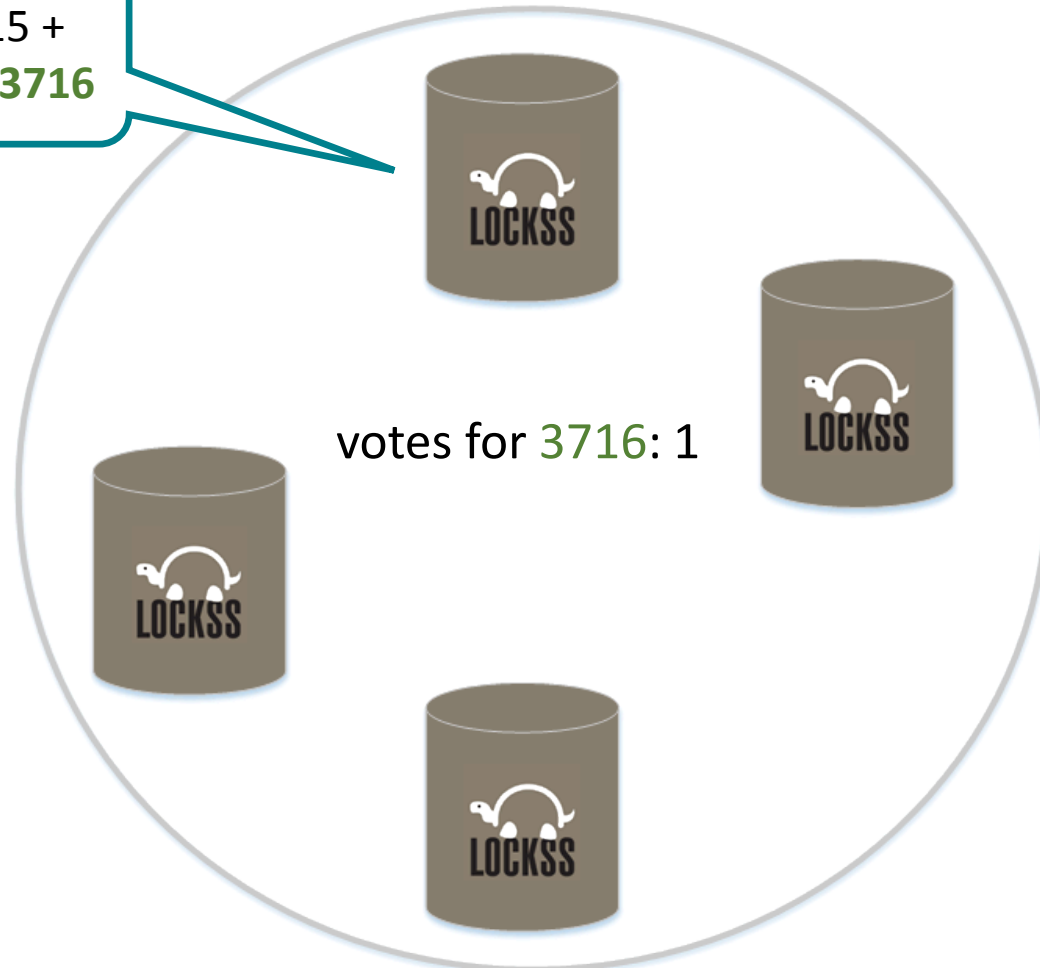


hashing

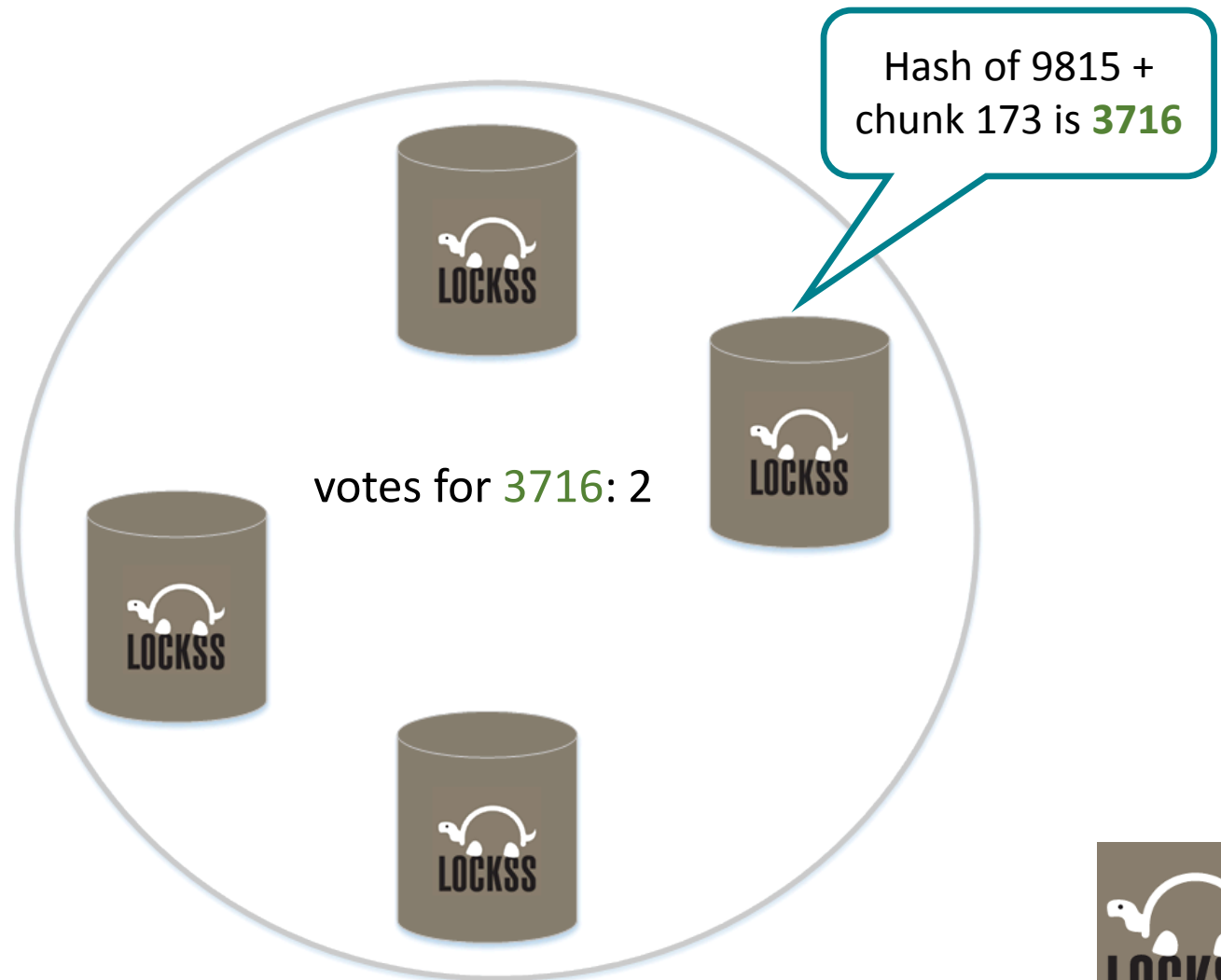


nodes respond to poll

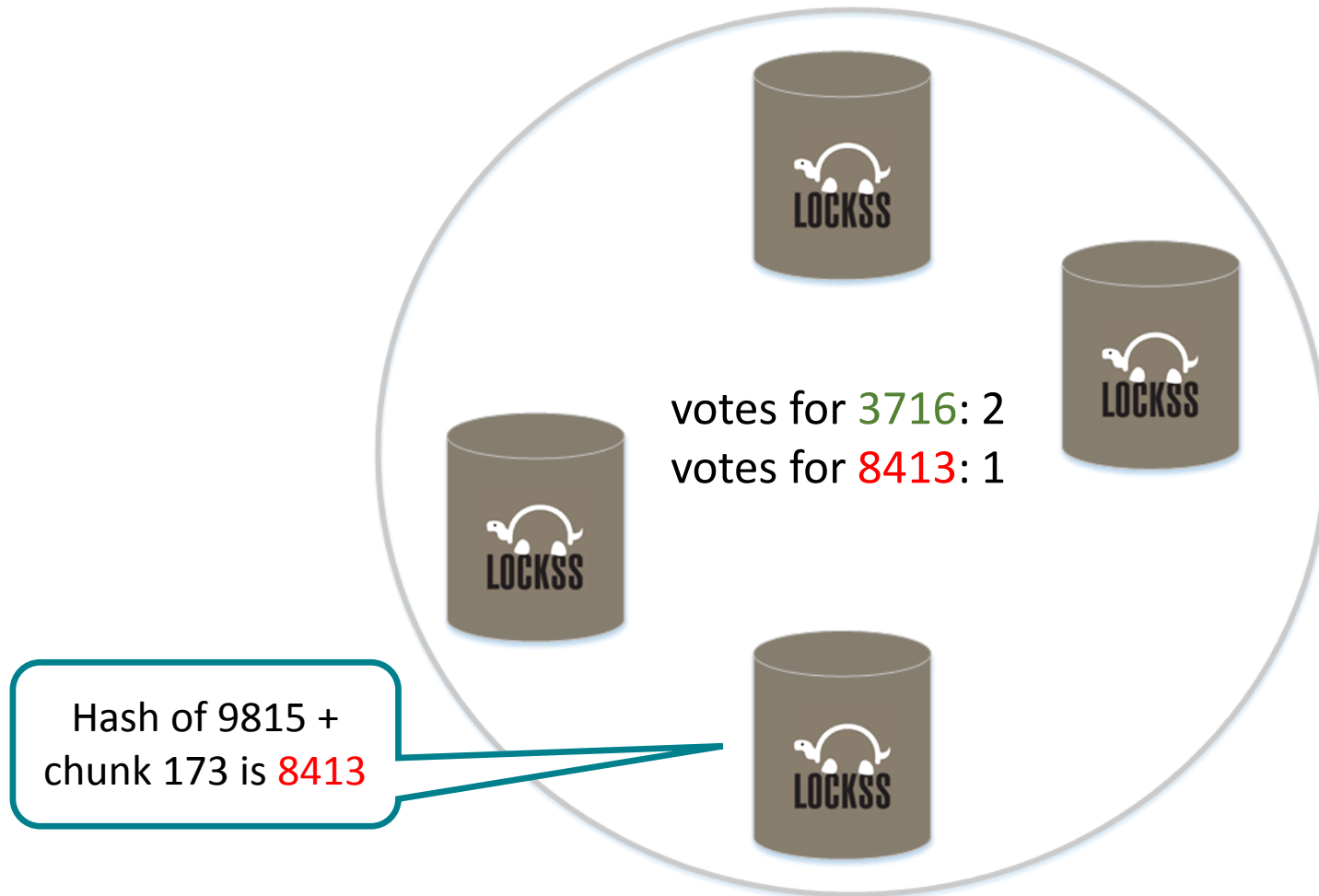
Hash of 9815 +
chunk 173 is **3716**



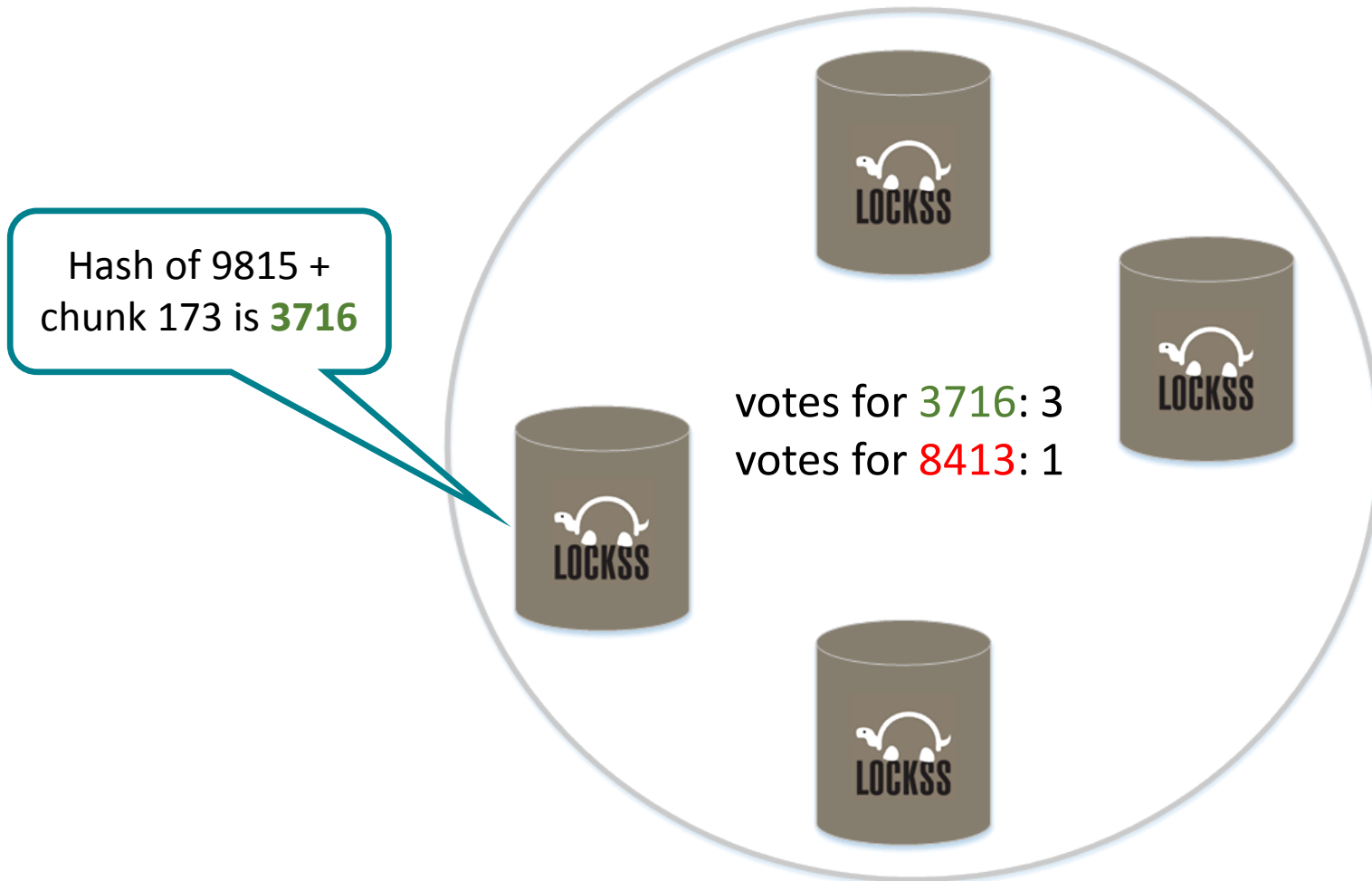
nodes respond to poll



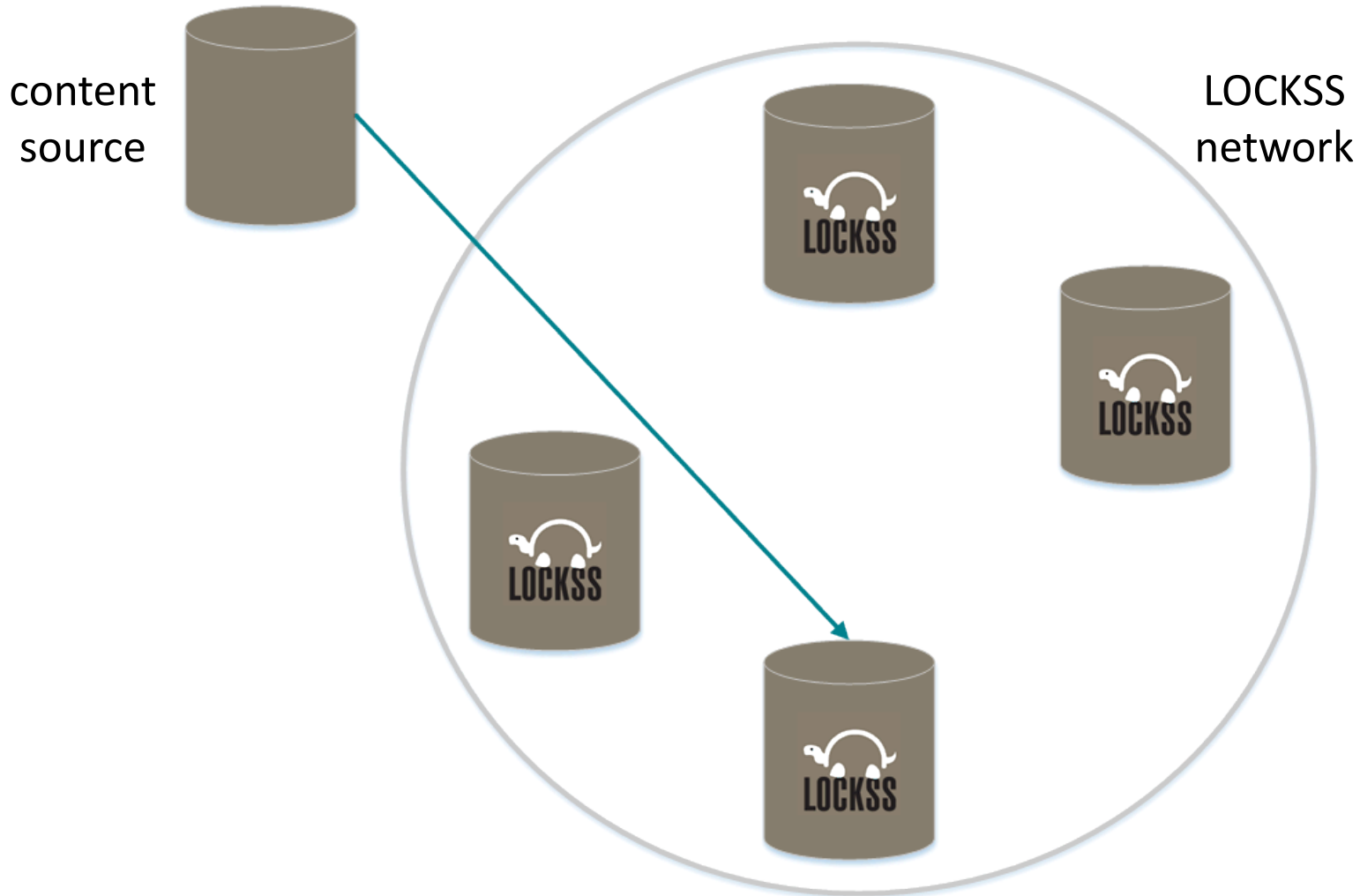
nodes respond to poll



initiate poll w/ nonce

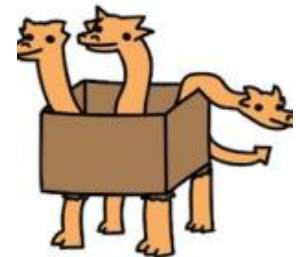


repair from content source



use cases for audit + repair

- other distributed digital preservation networks
- repository storage replication layers
- would like to support varied back-ends: tape, cloud, etc.



A photograph of a road curving to the right, bordered by a line of large, light-colored stones. The ground is covered in dry, brown autumn leaves and patches of green grass. The scene is captured in warm, golden-hour light. The word "Roadmap" is overlaid in a large, black, sans-serif font on a semi-transparent white rectangular background.

Roadmap

"Milestone?" by [Matteo De Toffoli](#) under [CC BY-NC-SA 2.0](#)

development progress

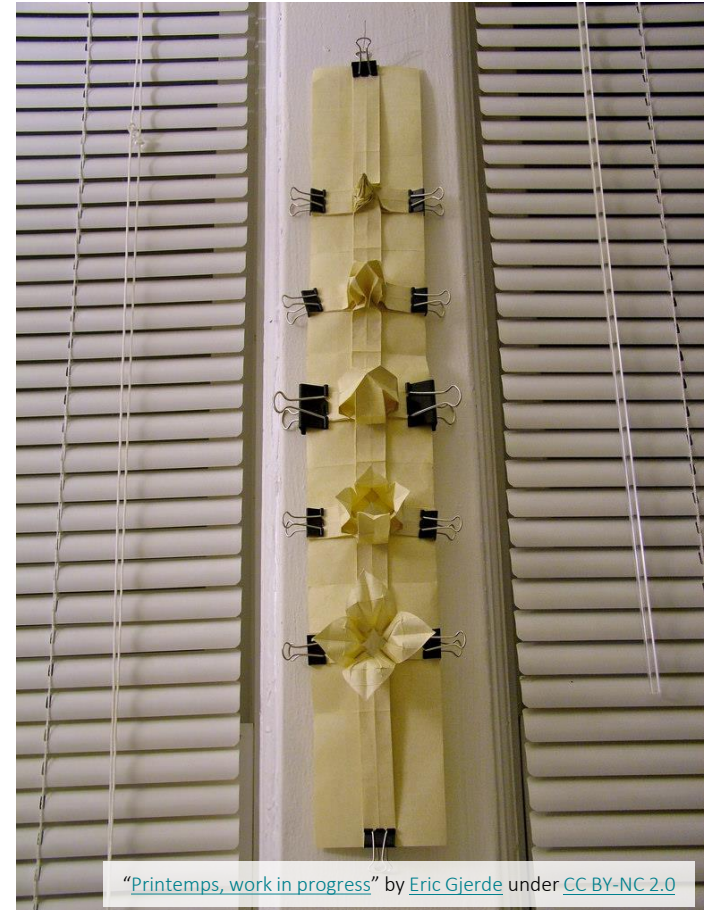
- access WARC-stored content via:
 - DOI
 - OpenURL
 - Memento (URL)
 - Solr full-text search
- web services:
 - metadata extraction
 - metadata query



"Milestones" by Dheeraj Nagwani under [CC BY-NC-ND 2.0](#)

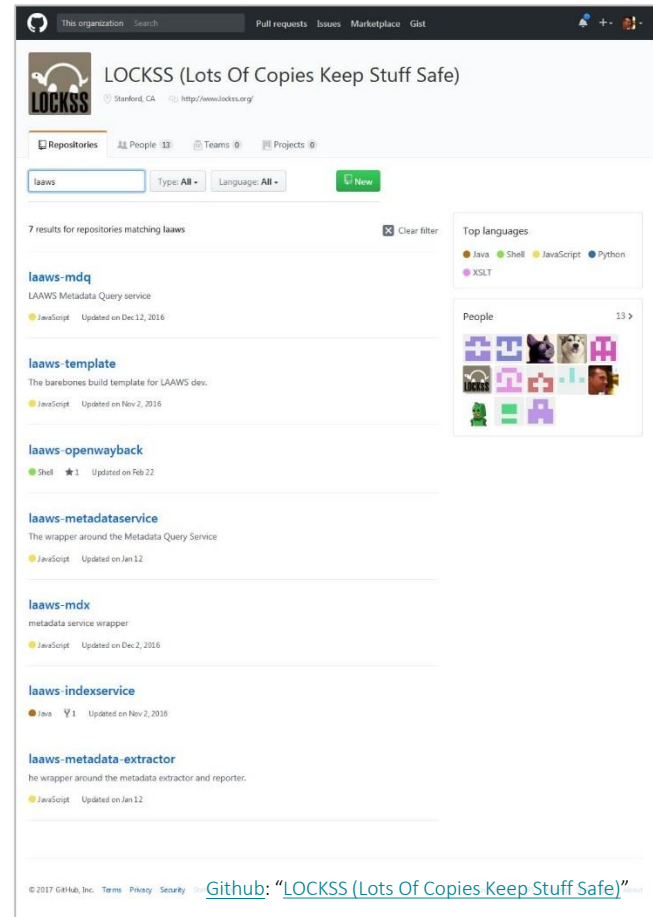
looking ahead

- by end of 2017
 - Docker-ize components
 - web harvest framework
 - polling + repair web service
- by end of 2018
 - IP address + Shibboleth access via OpenWayback
 - OpenWayback on-access format migration
 - full-text search web service



follow + plug in

- development (periodically) being pushed to [Github](#)
- moving toward more community-oriented software development
 - announcement of work cycles
 - sprint closeout reports + demos
 - community engagement



questions for you

- **what potential do you see** for LOCKSS technologies for web archiving, other use cases?
- what **standards or technologies could we use** that we maybe haven't considered?
- how could we help you to **use LOCKSS technologies?**
- how would you like to **see LOCKSS plug in more** to the web archiving community?

