LOTS OF COPIES KEEP STUFF SAFE

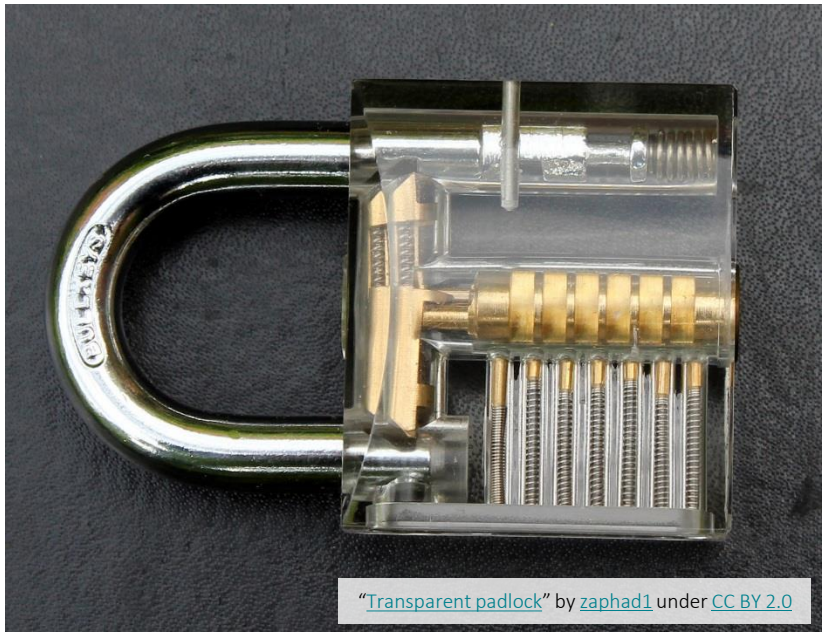
Unlocking LOCKSS with APIs

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

National Symposium on Web Archiving Interoperability
22 February 2017

a more interoperable LOCKSS

- beyond e-resources
  - a solution for **preserving** the **digital content your community cares about**
- APIs + interoperability
  - **maximize impact** by enabling integration + interconnection
  - **improve sustainability** by leveraging standardized community solutions



LOCKSS

lots of copies keep stuff safe

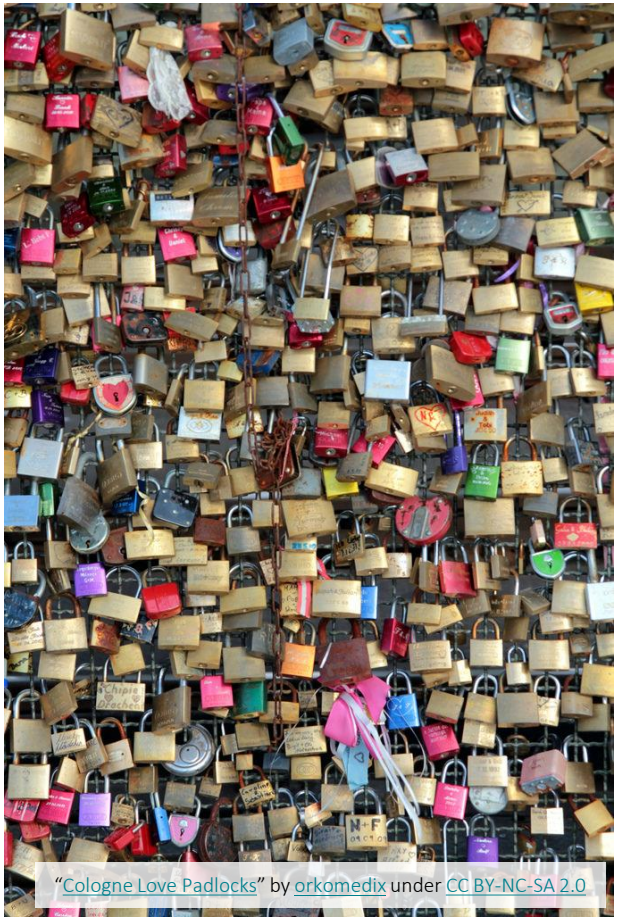
LOCKSS

Global LOCKSS Network

- subscription e-resources
- 150+ institutions
- each runs local LOCKSS node
- peer-to-peer preservation
- publishers opt-in
- post-cancellation access

lots of LOCKSS

- LOCKSS (principle)
- LOCKSS (program)
- LOCKSS (software)
- Global LOCKSS Network
- Private LOCKSS Networks
- Controlled LOCKSS (CLOCKSS)



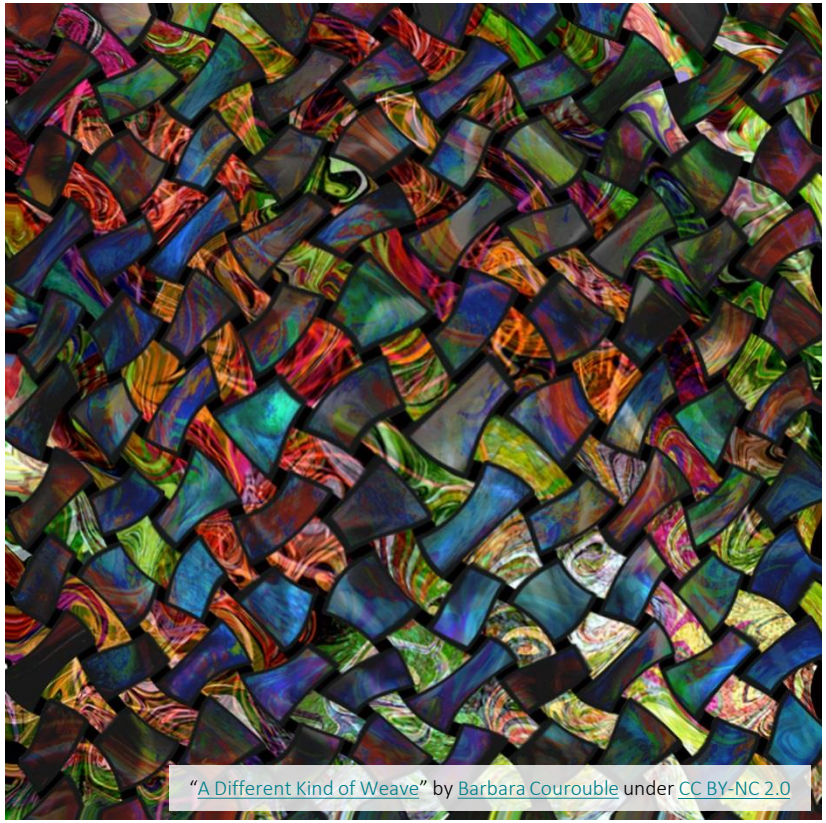
Private LOCKSS Networks (PLNs)

- community of interest
- jointly designate content
- run distributed nodes
- establish governance
- preservation via diverse:
  - technologies
  - institutions
  - networks

LOCKSS

integration opportunities

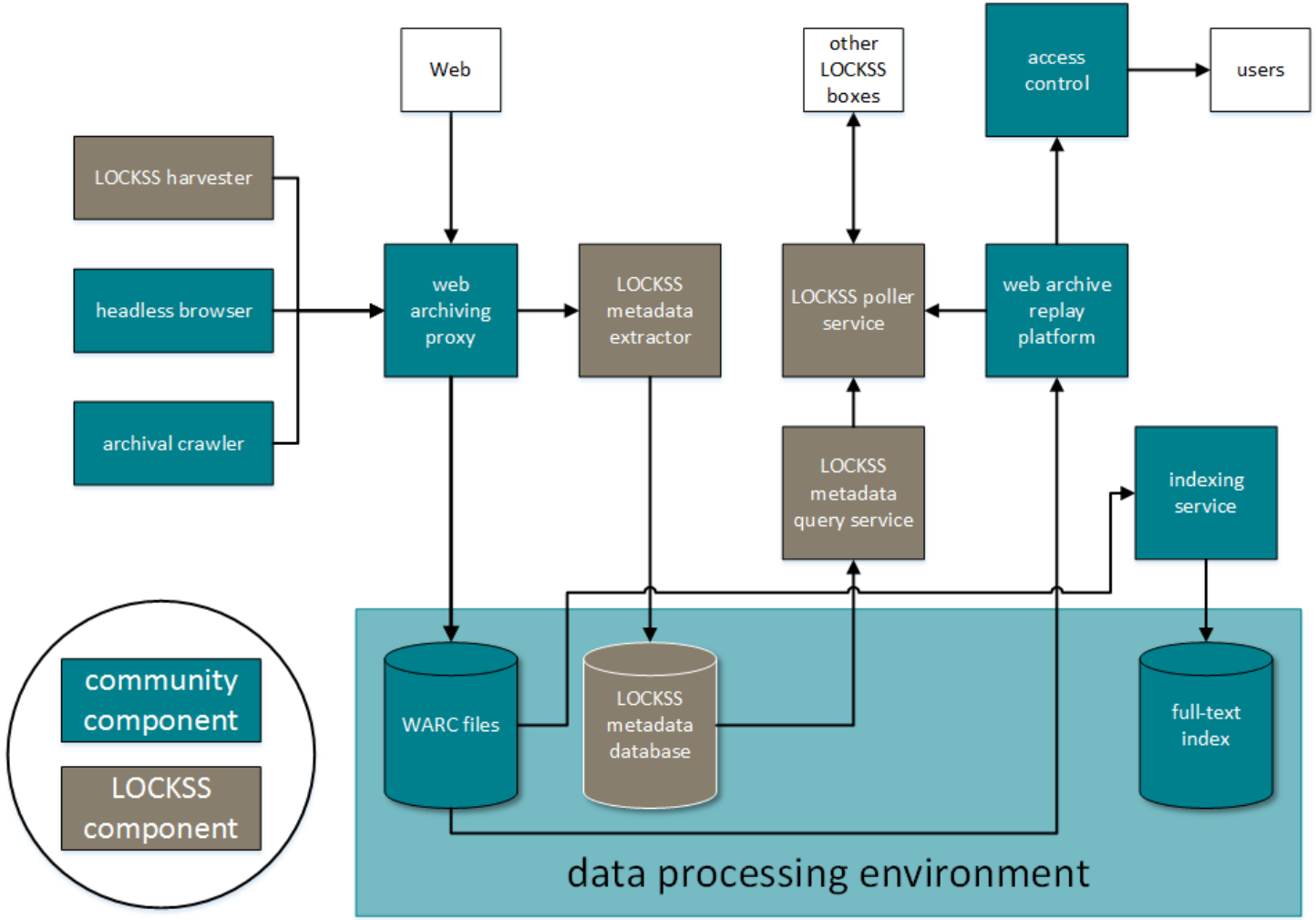
- polling + repair
  - repository replication
  - other distributed digital preservation systems
- access
  - Dockerized Solr indexing for WARC'ed content
  - DOI + OpenURL access to web archives
- metadata extraction



why re-architect LOCKSS?

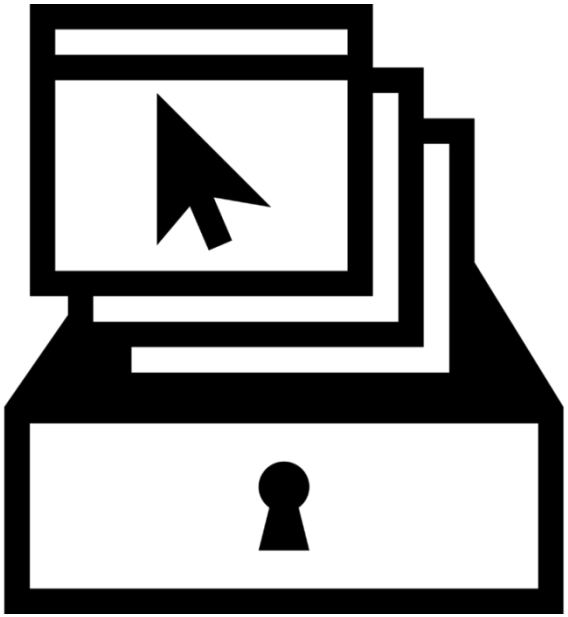
- reduce support + operations costs
  - leverage web-scale open-source software
  - align w/ web archiving mainstream
- de-silo components + enable external integration
  - metadata extraction
  - archive access via DOI + OpenURL
  - polling + repair protocol
- prepare to evolve w/ the Web
  - web services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

leveraging community components



aligning with web archiving

Web ARChive (WARC) format



compatible technologies

- Heritrix
- OpenWayback
- WarcBase
- Web Archiving Proxy

API candidates

- capture tool/proxy interconnect
- capture tool management
- **data import/export**
- query + extraction
- integrity audit + repair
- descriptive metadata
- logs + analytics
- renderings/derivative formats
- federated data delivery
- federated replay
- federated full-text search

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.

development roadmap

- 2017
  - Docker-ize components
  - web harvest framework
  - polling + repair web service
- 2018
  - IP address + Shibboleth access via OpenWayback
  - OpenWayback format negotiation framework
  - full-text search web service

LOCKSS