LOTS OF COPIES KEEP STUFF SAFE

LOCKSS Update: Offering Community Web Service Components

Art Pasquinelli
LOCKSS Partnerships Manager
Stanford University Libraries

PASIG Oxford
September 13, 2017

overview

- LOCKSS background and status
- software re-architecture
- roadmap

LOCKSS background and status

community-centric

- preservation is an **active** community effort
- lots of **communities** keep stuff safe
- enable preservation of the **content your community cares about**
- **enable** libraries to be **digital libraries**

LOCKSS

open-source software

- complements digital preservation
- mitigates lock-in
- facilitates data portability
- builds on open standards
- enables collaboration
- enhances security
- empowers adopters

LOCKSS

today

- new collaborative momentum
- self-sustaining organization
- large community of LOCKSS networks, hundreds of institutions
- TRAC-certified technologies
- handling all types of content
- increased focus on archives, research, repositories, museums, web archiving
- **Mellon Foundation support for SW re-architecture**

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**

decentralized community copies

- no monopoly on copy-making
- de-correlated, independent copies
- no central point of failure or vulnerability
- local community custody, focus, control

connecting w/ new communities

- streamline system←→network **data exchange**
- promote **API-oriented** architectures
- **contribute upstream** to shared tools
- broaden, diversify **community outreach**

software re-architecture

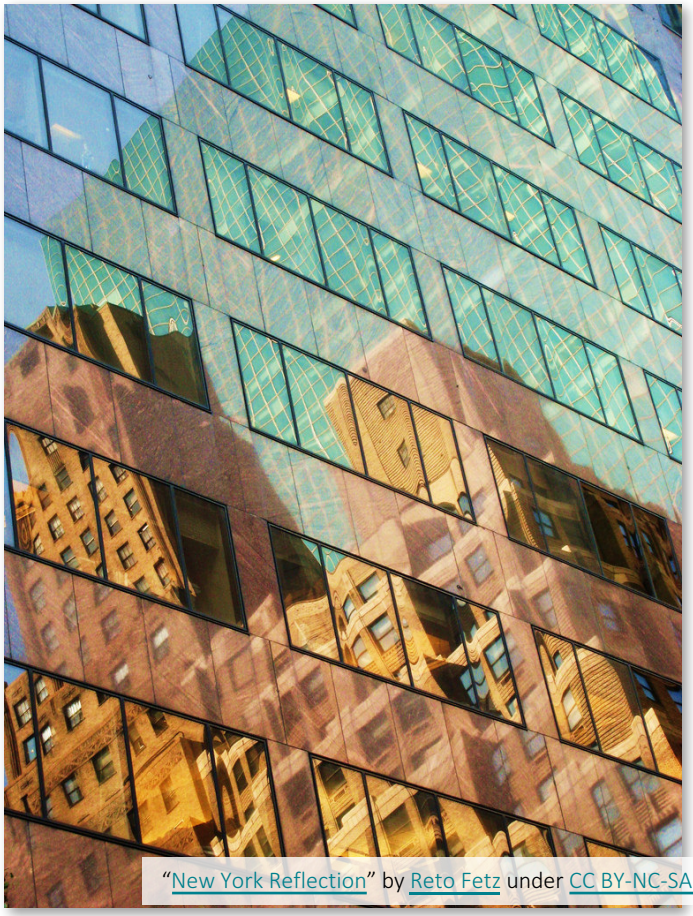
what we are doing

Press Release, September 7, 2017

“The core of the LOCKSS software is a peer-to-peer data **integrity validation and repair mechanism**, a feature built upon peer-reviewed research to mitigate the real threats that centralization poses to the long-term persistence of digital information. This and other LOCKSS software elements, including tooling for **automated metadata extraction and enhancements for discovery of scholarly communications** within web archives, will be made available to the community as documented web services.“

why re-architect LOCKSS?

- reduce support and operational costs
- de-silo components + enable external integration
- evolve with new web technology and services
- increase both community and partner enhancements



polling and 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

LOCKSS

use cases for polling and repair

- distributed digital preservation networks consortia, and communities
- repository storage replication layers
- existing communities with specialized, restricted, or at-risk content
- would like to support varied back-ends: tiered storage, cloud, etc.

automated metadata extraction

functionality

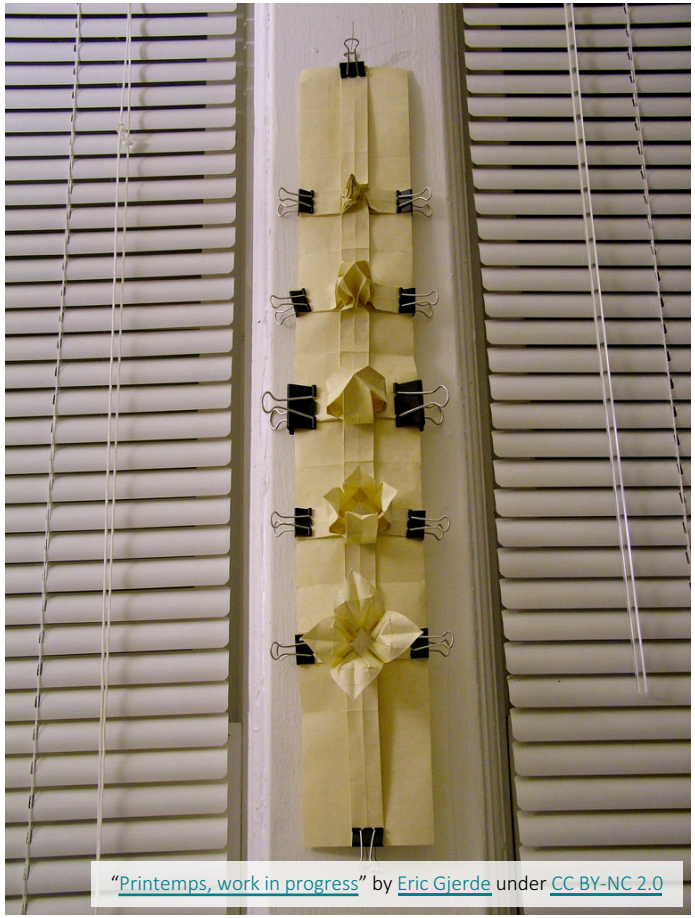
- for both web harvest and file transfer content
- map values in Document Object Model (DOM) tree to metadata fields
- retrieve downloadable metadata from expected URL patterns

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

looking ahead

- early 2018
  - Docker containerization
  - web harvest framework
  - polling + repair web service
- late 2018
  - access control in OpenWayback
  - full-text search web service



questions for you

- **what potential do you see** for LOCKSS technologies?
- what **standards or technologies could we use?**
- how could we help you to **use LOCKSS technologies**?
- how would you like to **see LOCKSS plug in more** to other communities?

I work here

X