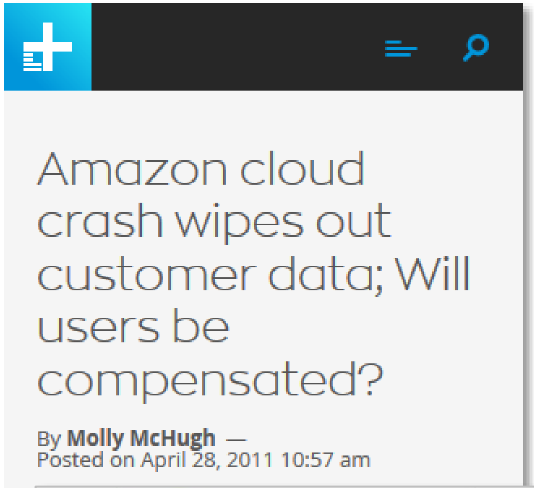
LOCKSS

Continuing to Keep Stuff Safe with Lots of Copies, Communities, and Innovation

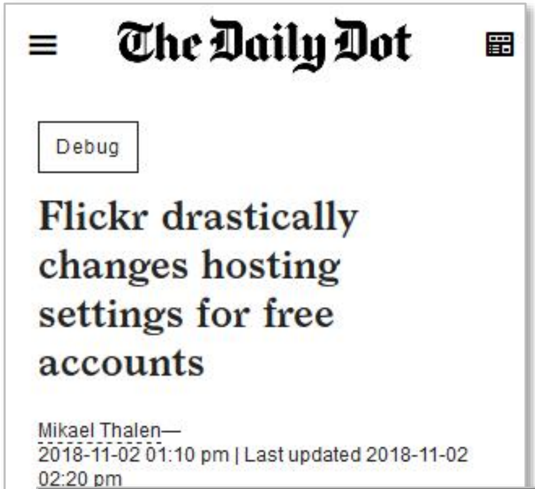
Nicholas Taylor (@nullhandle)
Program Manager, LOCKSS and Web Archiving
Stanford Libraries

LOCKSS open webinar
29 January 2019



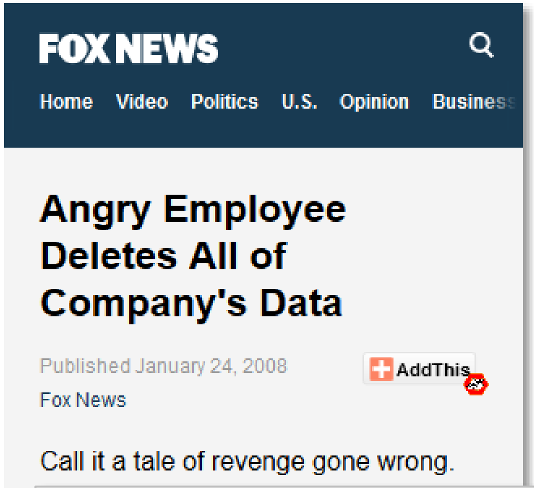
Amazon cloud crash wipes out customer data; Will users be compensated?

By Molly McHugh —
Posted on April 28, 2011 10:57 am



Flickr drastically changes hosting settings for free accounts

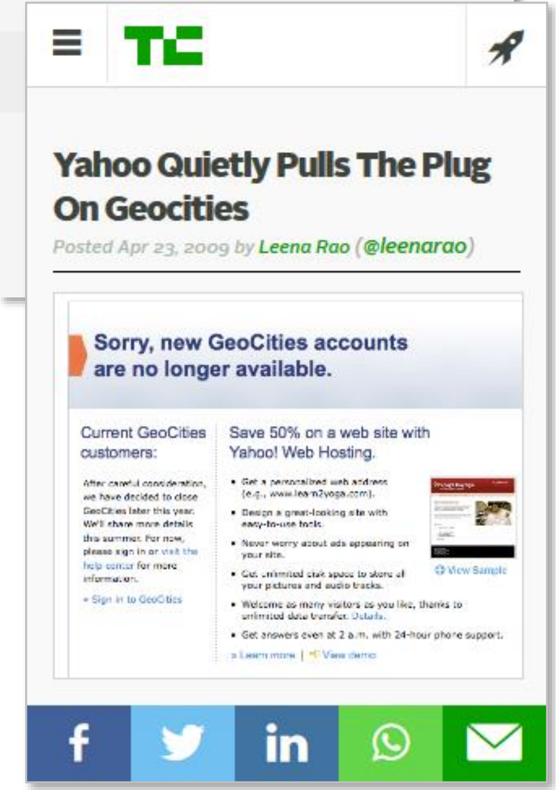
Mikael Thalen—
2018-11-02 01:10 pm | Last updated 2018-11-02 02:20 pm



Angry Employee Deletes All of Company's Data

Published January 24, 2008
Fox News

Call it a tale of revenge gone wrong.



Yahoo Quietly Pulls The Plug On Geocities

Posted Apr 23, 2009 by Leena Rao (@leenarao)

Sorry, new GeoCities accounts are no longer available.

Current GeoCities customers:

After careful consideration, we have decided to close GeoCities later this year. We'll share more details this summer. For now, please sign in or visit the help center for more information.

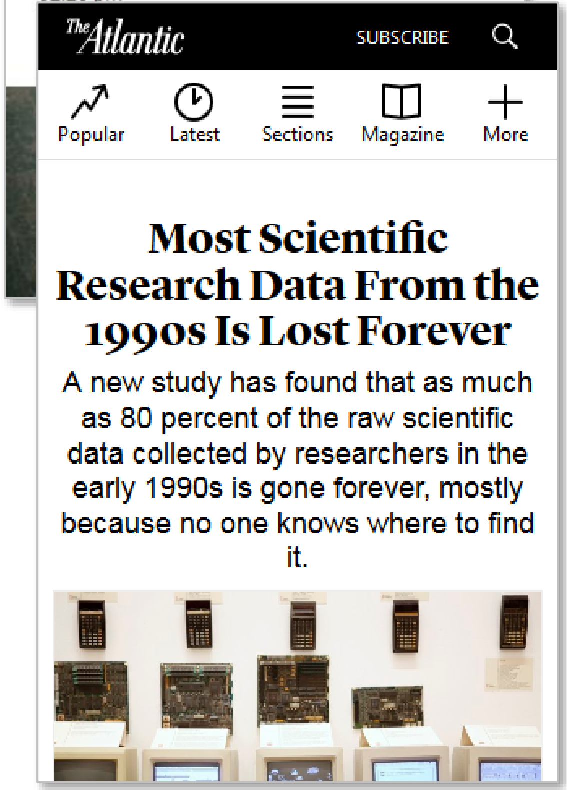
» Sign in to GeoCities

Save 50% on a web site with Yahoo! Web Hosting.

- Get a personalized web address (e.g., www.learn2yoga.com).
- Design a great-looking site with easy-to-use tools.
- Never worry about ads appearing on your site.
- Get unlimited disk space to store all your pictures and audio tracks.
- Welcome as many visitors as you like, thanks to unlimited data transfer. Details.
- Get answers even at 2 a.m. with 24-hour phone support.

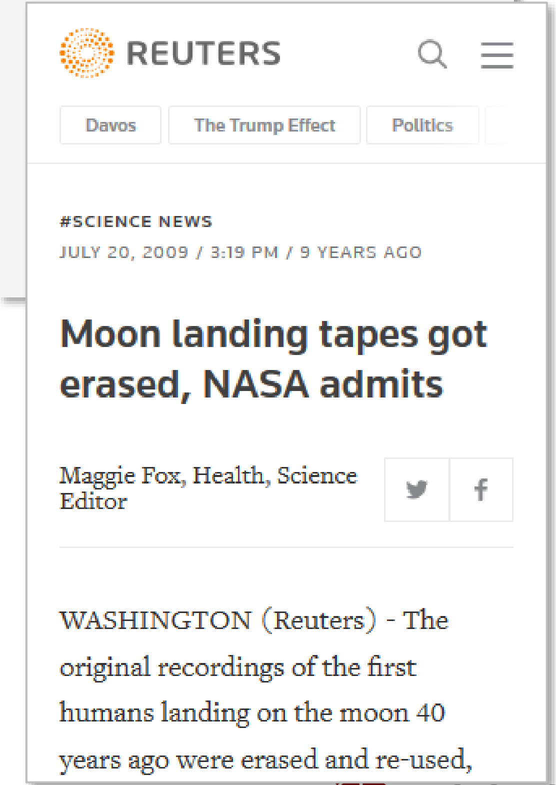
» Learn more | View demo

View Sample



Most Scientific Research Data From the 1990s Is Lost Forever

A new study has found that as much as 80 percent of the raw scientific data collected by researchers in the early 1990s is gone forever, mostly because no one knows where to find it.



#SCIENCE NEWS
JULY 20, 2009 / 3:19 PM / 9 YEARS AGO

Moon landing tapes got erased, NASA admits

Maggie Fox, Health, Science Editor

WASHINGTON (Reuters) - The original recordings of the first humans landing on the moon 40 years ago were erased and re-used,

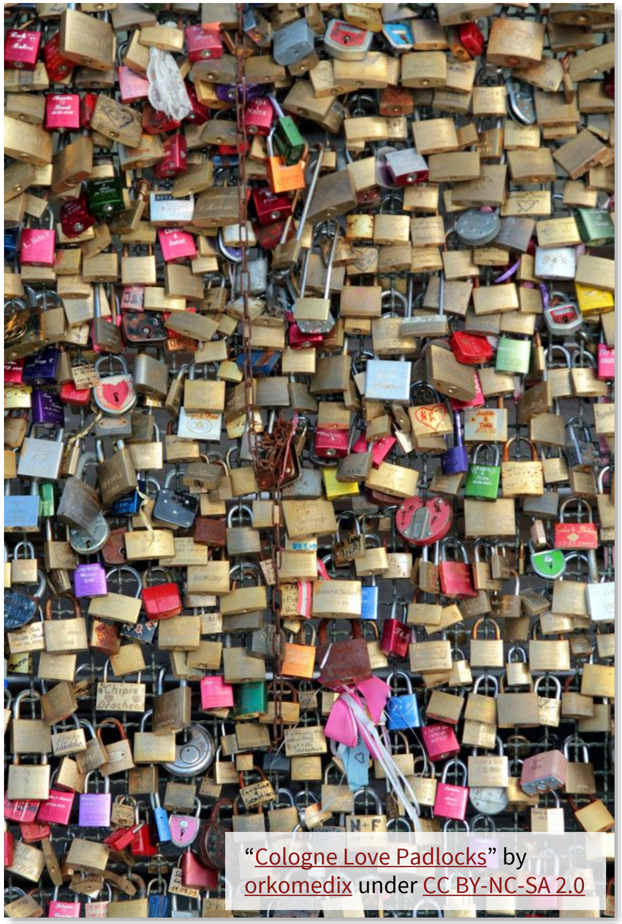
understand + mitigate threats

- long-term **data integrity** is hard
- needs architecture informed by **actual leading threats** to data
- don't underestimate:
  - people making **mistakes**
  - **attacks** on information
  - organizational **failure**

"Fragile" by Garrett Coakley under CC BY-NC 2.0

what is LOCKSS?

- a widely-accepted **principle** for the persistence of digital info
- a digital library-focused **program** of Stanford Libraries
- a research-informed **software** app for p2p digital preservation
- an international **community** of institutions + networks



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

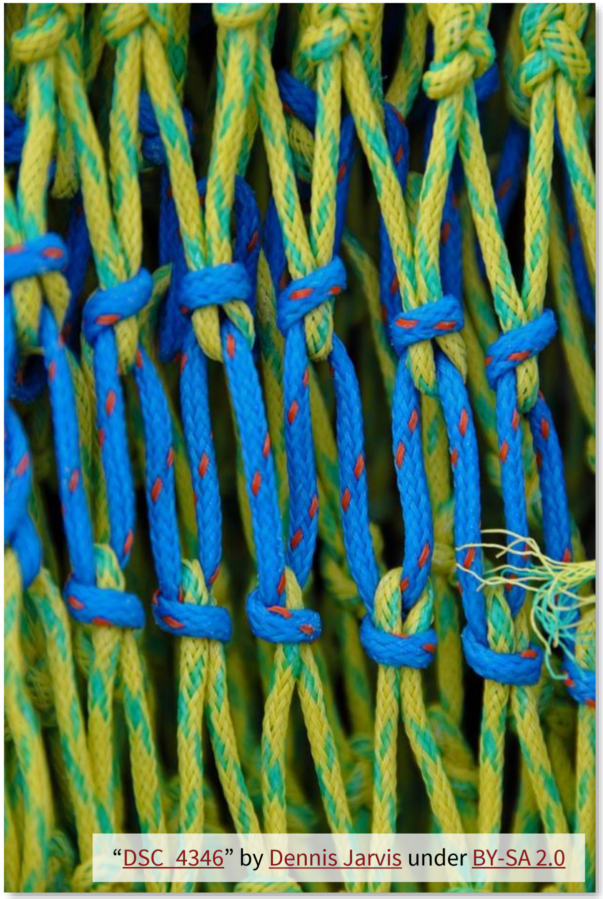
more than lots of copies

- lots of copies is **necessary but not sufficient**
- central points of failure **can undermine all copies at once**
- LOCKSS provides:
    - continual **integrity checking + repair**
    - b/t mutually-distrusting, **independent peers**
    - on a network that **your community controls**

"Domino's" by david pacey under CC BY 2.0

routine audit + repair

- ensuring **long-term data integrity**
  - must **read data** to know it's good
  - easier to **repair data sooner**
- network nodes **conduct polls** to validate integrity of distributed copies
- more nodes = **more security**
  - more nodes can be **down**
  - more copies can be **corrupted**
  - …and polls will **still conclude**



"DSC_4346" by Dennis Jarvis under BY-SA 2.0

community + digital preservation

- communities **complement** LOCKSS:
  - **resilience** against organizational failure
  - native **heterogeneity**
- preservation is an **active** community effort
- lots of **communities** keep stuff safe

"Redwood Canopy" by Floyd Stewart under CC BY-NC-SA 2.0

How LOCKSS Works

“gears” by starfive under CC BY-NC-SA 2.0

distributed digital preservation

- align w/ **best practice**
- achieve **resilience** not possible w/ centralized solution
- for use either:
  - as **dedicated** preservation solution
  - to **supplement** local preservation (e.g., for most important materials)
- particular to LOCKSS among service providers:
  - strongly **research-based**
  - articulated **threat model**
  - supports **local custody**

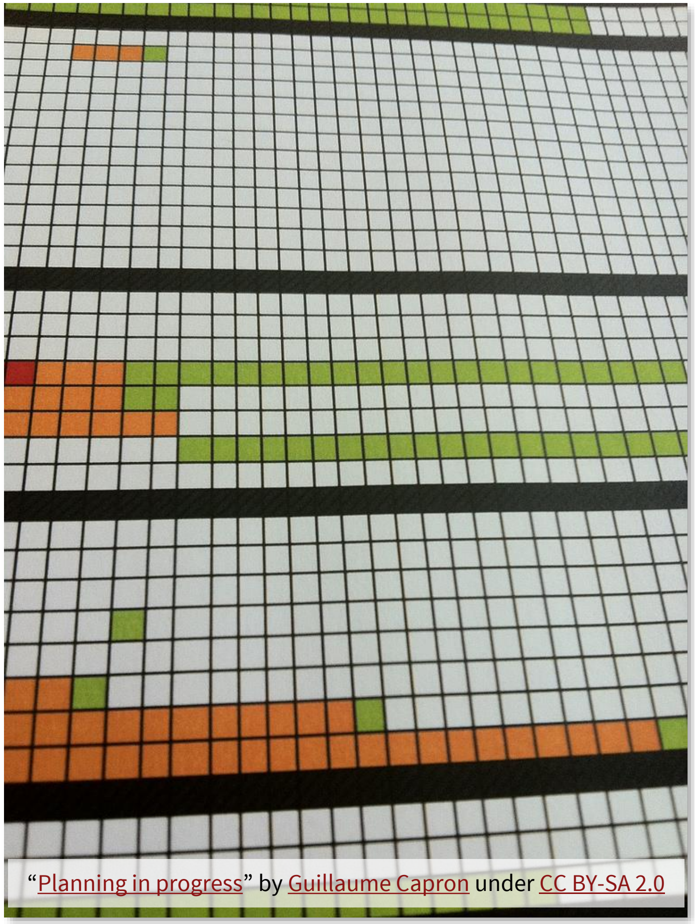
"Stone stacks" by Jack Malvern under CC BY-NC 2.0

content lifecycle

- ingest content
  - web harvest, OAI-PMH, direct interconnect, or drag-and-drop via LOCKSS-O-Matic
- manage content
  - web-accessible GUI to monitor preservation activity (+ select new content for archiving, in some networks)
- preserve content
  - each node retrieves content independently
  - once stored, audit + repair takes place automatically, on ongoing basis
- deliver content
  - proxy server, web server, OpenURL

setup, support, costs

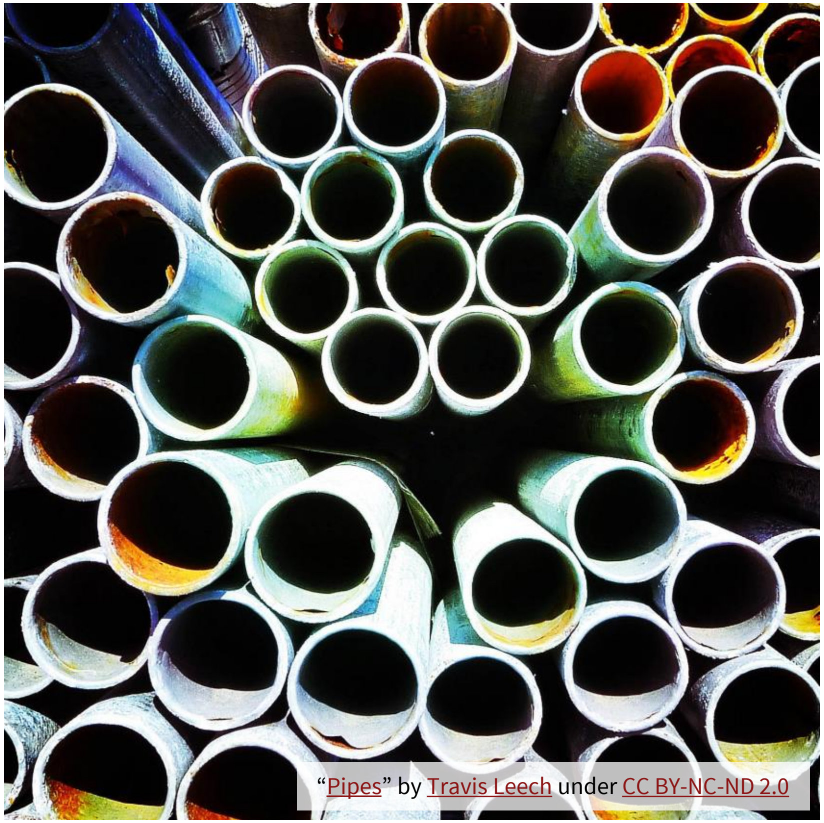
- **organize your community** around content of shared concern
- we will **consult on fit**, technical requirements, workflow integration
- **pilot implementation** w/ subset of nodes to validate workflows
- **production implementation**
- **ongoing support**
- participants are asked to join the LOCKSS Alliance (annual membership fee)



"Planning in progress" by Guillaume Capron under CC BY-SA 2.0

start new or join existing network

- start a new network
  - recommend 4+ copies
  - we can host node(s)
- also an option to join an existing network
  - reach out to network communities directly
- we are exploring how to **better support needs of individual orgs** that aren’t aligned w/ a logical community



“Pipes” by Travis Leech under CC BY-NC-ND 2.0

Use Cases

“ROC/Bredero” by Jesper2cv under CC BY-NC-ND 2.0

post-cancellation access for e-resources

- networks:
  - Global LOCKSS Network
- restore **best features of print journal holdings** lost w/ online publishing transition:
  - **local custody** (vice contingent access)
  - lots of **decentralized copies** (vice fewer, centralized copies)
- to better assure:
  - preservation of **scholarly record**
  - continuing **library role as steward**

LOCKSS

dark archive for scholarly publications

- networks:
  - CLOCKSS Archive
  - Public Knowledge Project Preservation Network
- CLOCKSS
  - **co-governed** by libraries + publishers
  - content **triggered OA** when no longer available
  - top CRL TRAC audit score
- PKP PN
  - OA content **hosted on OJS**
  - free + seamless to use for folks publishing on OJS

government information

- networks:
  - Canadian Government Information
  - Digital Federal Depository Library Program
- **can’t necessarily depend** on government for permanent access
- save + **re-decentralize** government information

institutional repository content

- networks:
  - Alabama Digital Preservation Network
  - MetaArchive Cooperative
  - WestVault
- **all types** of content
- service models:
  - **all depositors** also host infrastructure
  - **subset of orgs** hosts infrastructure but serves whole community
- governance + infrastructure both **community-based**

COPPUL

MetaArchive
COOPERATIVE

The Alabama Digital Preservation Network
Preserving Alabama's Digital Resources

web archives

- networks:
  - Ivy Plus Libraries Confederation Preservation Network
- growing relative importance of web archives for **collection development**
- decentralized local custody + preservation to **complement Archive-It**

Ivy Plus Libraries

national / nationally-licensed scholarly publications

- networks:
  - Cariniana
  - German national network
- natural national interest in **preserving own OA output**
- national consortia want **jurisdictional control** over licensed scholarly content

Cariniana
Rede Brasileira de Serviços de Preservação Digital

Software Re-Architecture

“The two bridges” by Frank Schulenburg under BY-SA 2.0

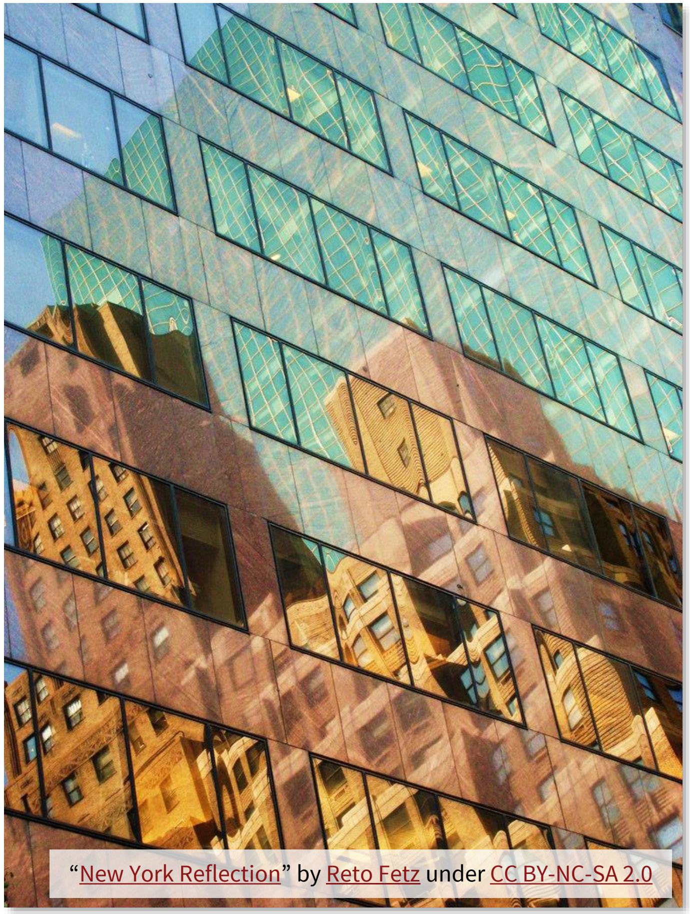
software re-architecture motivation

- **monolithic** Java application
- only deployable as **end-to-end solution**
- lacking **modern APIs**
- maintaining functionality on our own that others **increasingly address as a community**
- undertook re-architecture 2017-2019 w/ Mellon Foundation funding

“The SF Bay Bridge reopens” by Benjy Feen under CC BY-NC-SA 2.0

software re-architecture goals

- capitalize on **work of broader communities**
- de-silo + enable **external integrations**
- empower community of practice w/ **better documentation** + well-defined APIs
- evolve w/ web + **digital preservation ecosystem**



"New York Reflection" by Reto Fetz under CC BY-NC-SA 2.0

anticipated outcomes

- collaborate to **build new hybrid solutions**
- align better w/ **community workflows + interfaces**
- **simplify adaptation** of LOCKSS to local needs
- support LOCKSS technical **community of practice**
- **expand contexts** where LOCKSS can contribute to digital preservation

" Bloom" by Jak W! under CC BY-NC 2.0

new integration possibilities

- **Data Life-Cycle Management (DLCM)**
  - Swiss universities collaborative research data management
- **Software Preservation Network (SPN)**
  - promoting best practices + piloting distributed emulation infrastructure
- **Webrecorder**
  - high-fidelity web capture + replay software

fixity service

- some content too big for lots of copies
- instead, make lots of copies of checksums
- subject to LOCKSS polling + repair
- provide API endpoint
- compare w/ hash result generated by external system

"Measure twice, cut once..." by GretaMichelle Joachim under CC BY 2.0

cloud friendl(ier)

- may enable some use cases; improve handling of others
- technically feasible, but not economically optimized
- explore using cloud in concert w/ fixity service
- benchmark cloud costs (revisiting prior research on LOCKSS in the cloud)
- leverage w/o ceding value of distributed, local content custody

"State-of-the-art cloud storage. 2015." by Samarth Shyamanur under CC BY-NC 2.0

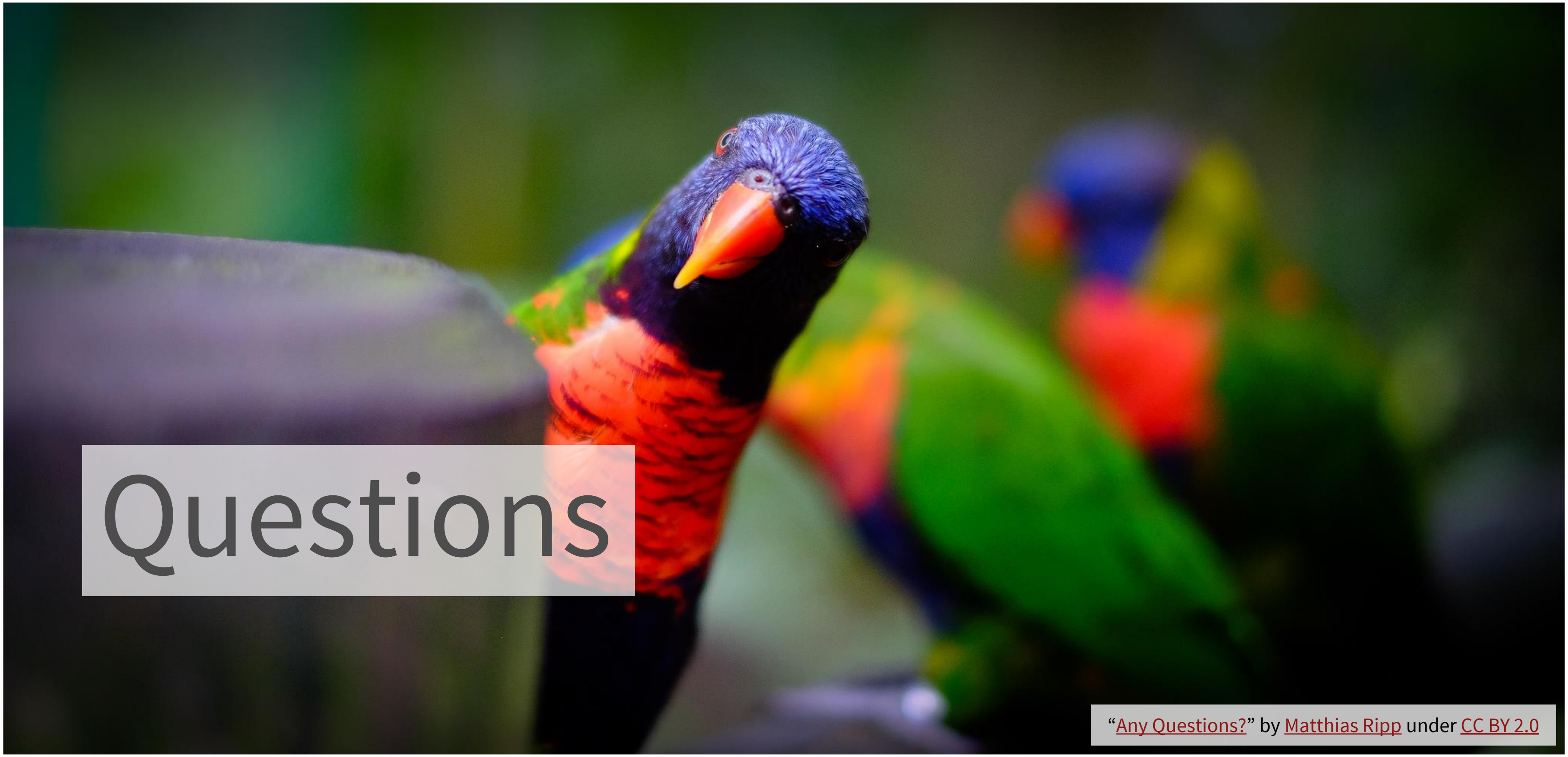
Wrap Up

"Birmingham City Centre - Nov 2013 - The New Library Wrapped and Ready for Christmas" by Gareth Williams under CC BY 2.0

takeaways

- LOCKSS is a **general-purpose** digital preservation platform
- re-architecture will provide **improved integration + interoperability**
- learn more at our **new website**: lockss.org
- **please contact us** with any questions, or ideas on how we can work together

"Partir" by Chiara Conticelli under Fair Use



Questions

"Any Questions?" by Matthias Ripp under CC BY 2.0