



STANFORD UNIVERSITY LIBRARIES

Building Web Archiving Technology, Together

Nicholas Taylor
[Web Archiving](#) Service Manager
[Stanford University Libraries](#)

[Web Archives 2015: Capture, Curate, Analyze](#)
November 13, 2015

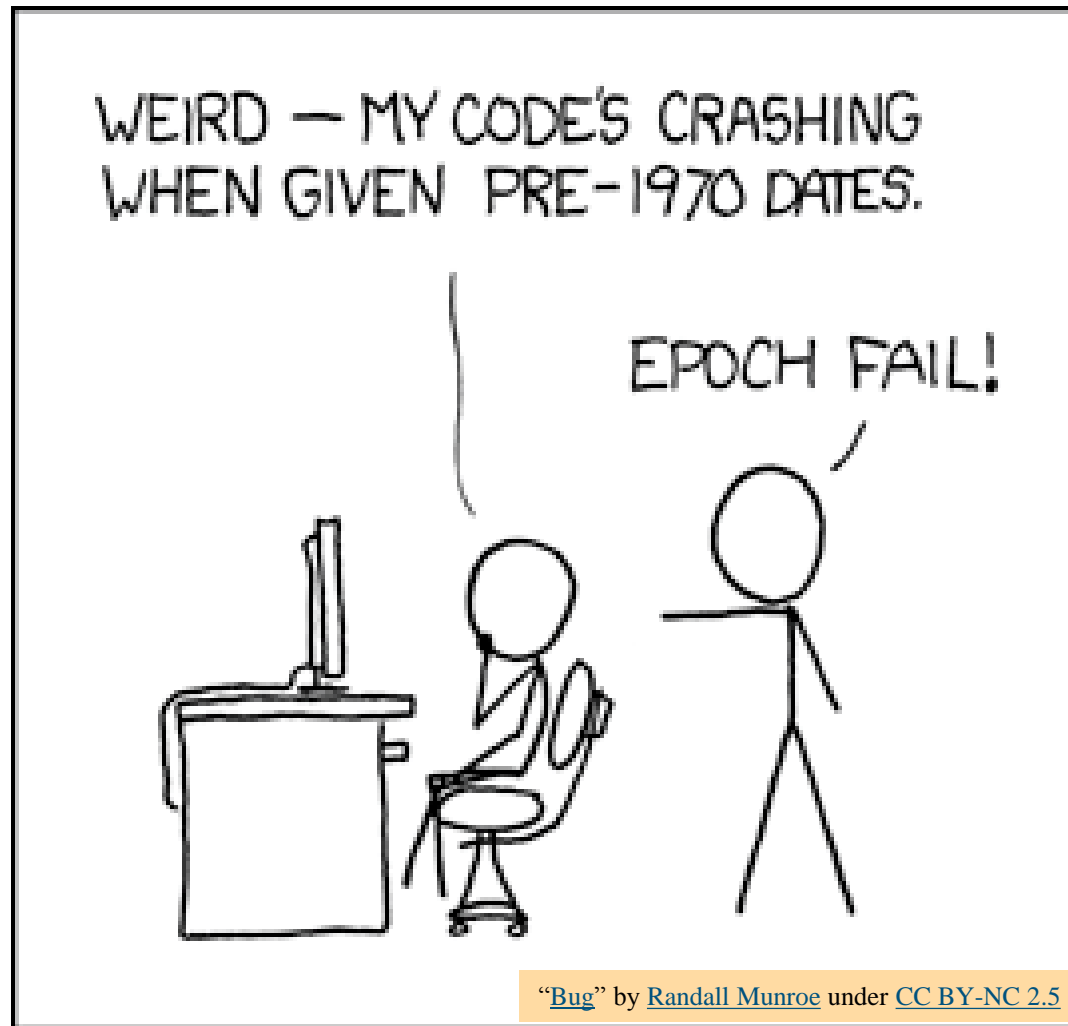
overview

- **why** build together?
- **community** for collaborative work
- **APIs** for collaborative work

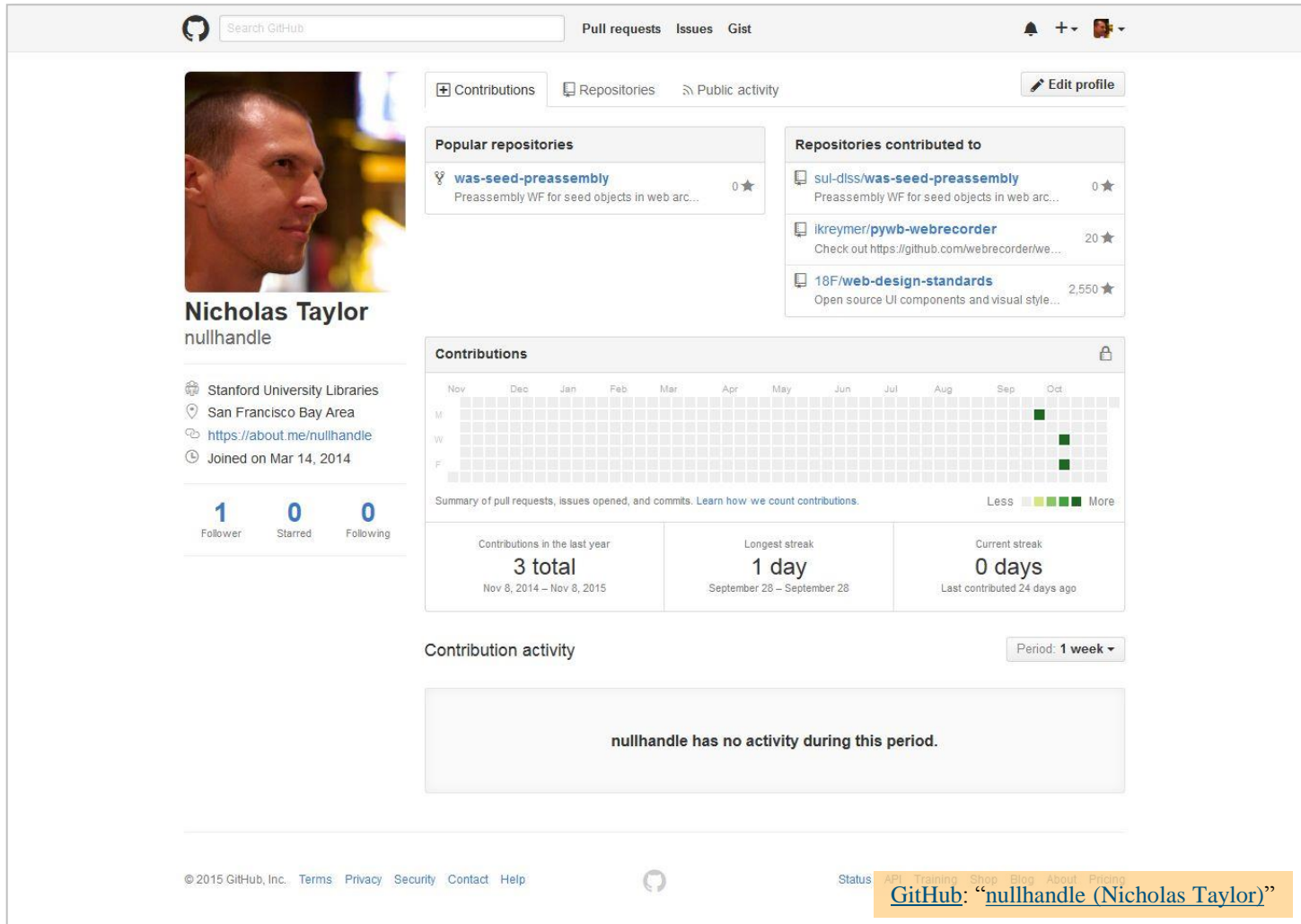


“LAX on take off” by [Doug](#) under [CC BY-NC-ND 2.0](#)

not a programmer



aspiring OSS contributor



Search GitHub

Pull requests Issues Gist

+ -

Contributions Repositories Public activity Edit profile

Popular repositories

- [was-seed-preassembly](#) 0 ★
Preassembly WF for seed objects in web arc...

Repositories contributed to

- [sui-dlss/was-seed-preassembly](#) 0 ★
Preassembly WF for seed objects in web arc...
- [lkreymer/pywb-webrecorder](#) 20 ★
Check out <https://github.com/webrecorderwe...>
- [18F/web-design-standards](#) 2,550 ★
Open source UI components and visual style...

Contributions

Summary of pull requests, issues opened, and commits. [Learn how we count contributions.](#) Less More

Contributions in the last year
3 total
Nov 8, 2014 – Nov 8, 2015

Longest streak
1 day
September 28 – September 28

Current streak
0 days
Last contributed 24 days ago

Contribution activity Period: 1 week

nullhandle has no activity during this period.

© 2015 GitHub, Inc. [Terms](#) [Privacy](#) [Security](#) [Contact](#) [Help](#)

Status [API](#) [Training](#) [Shop](#) [Blog](#) [About](#) [Pricing](#)

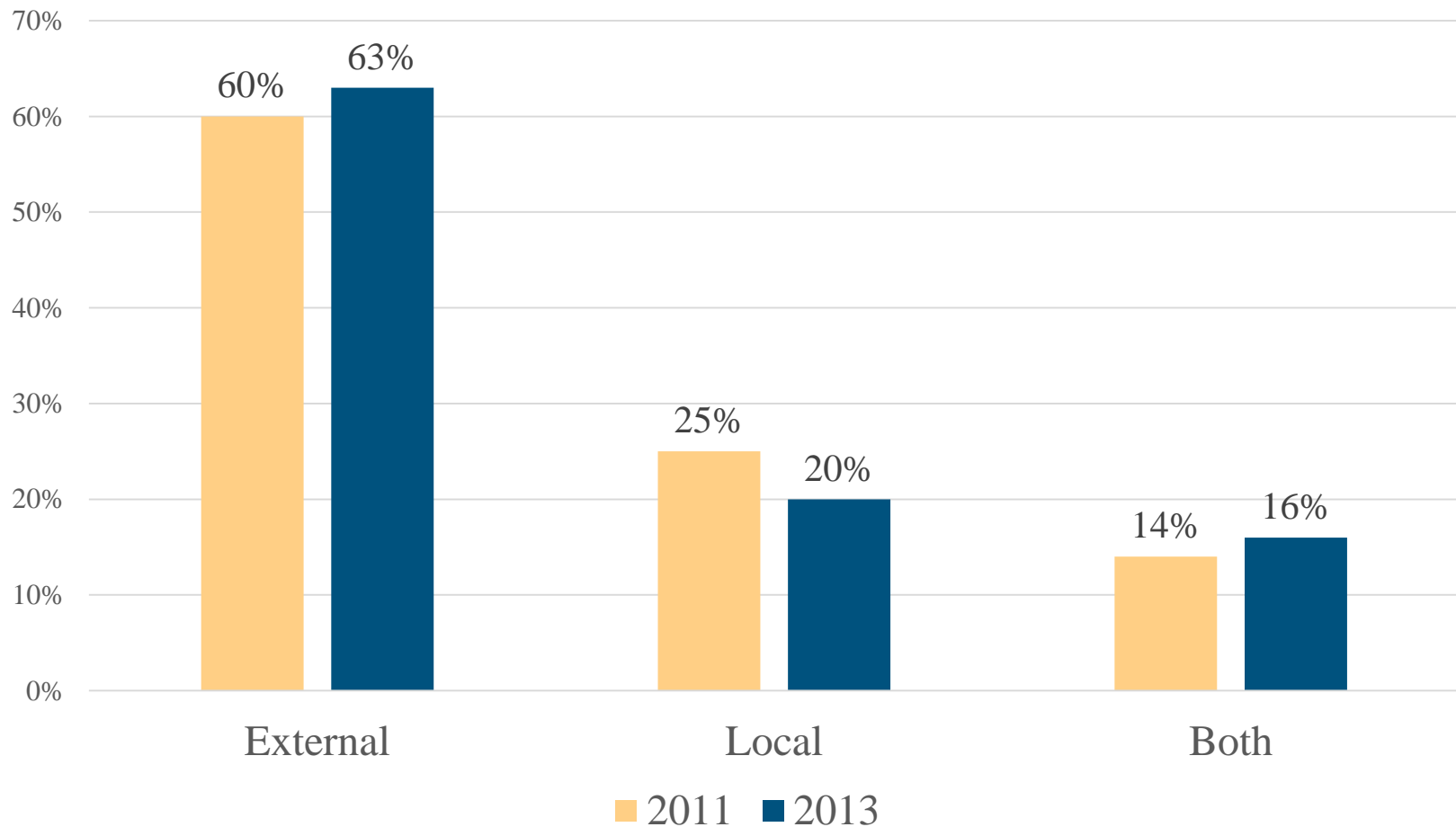
GitHub: “nullhandle (Nicholas Taylor)”

studying the landscape

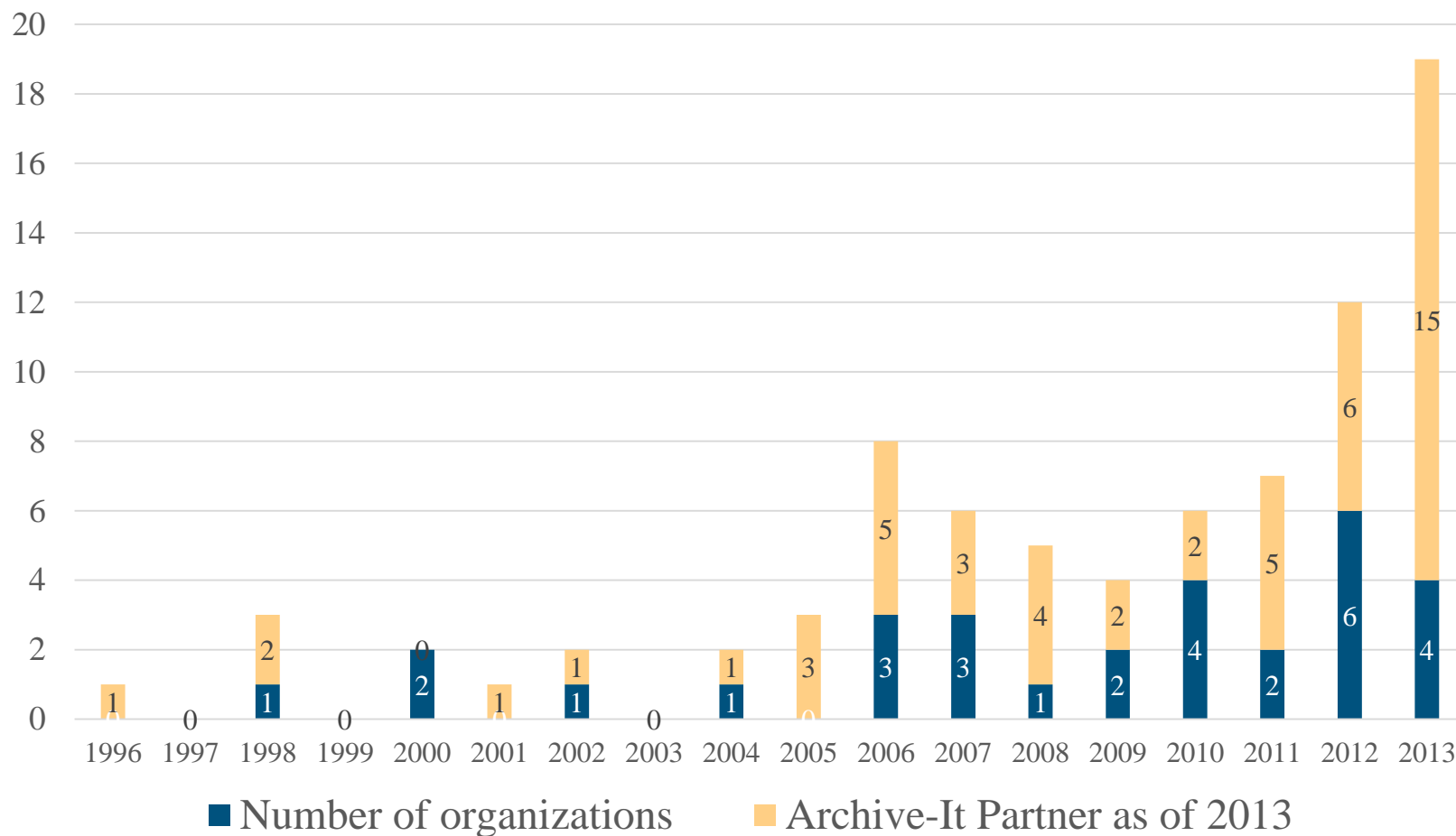


“2010 Grand Canyon Celebration of Art 172” by [Grand Canyon National Park](#) under [CC BY 2.0](#)

a centralized enterprise

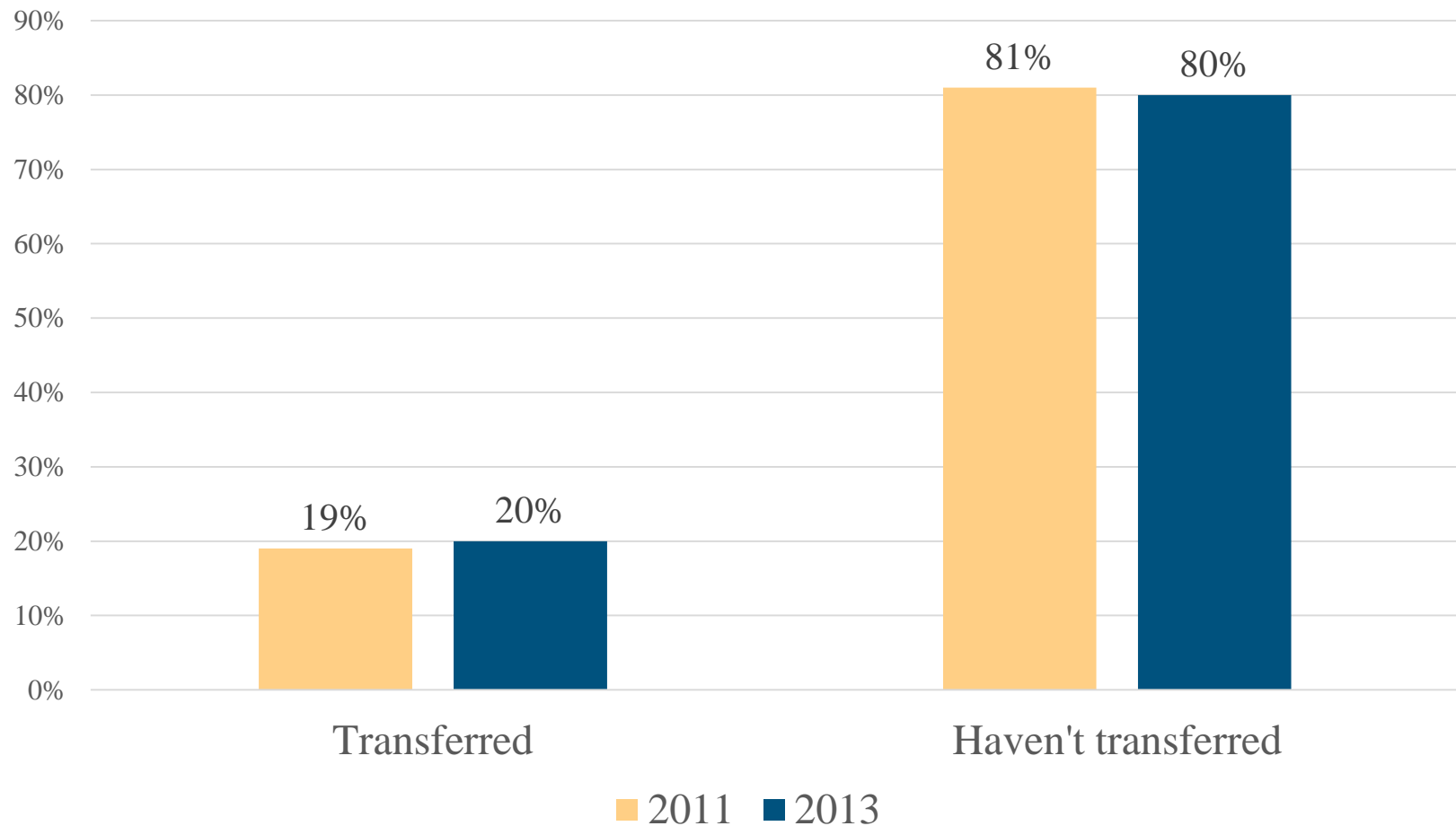


a centralized enterprise



NDSA: "Web Archiving in the U.S.: A 2013 Survey"

minimal local preservation



evolving web



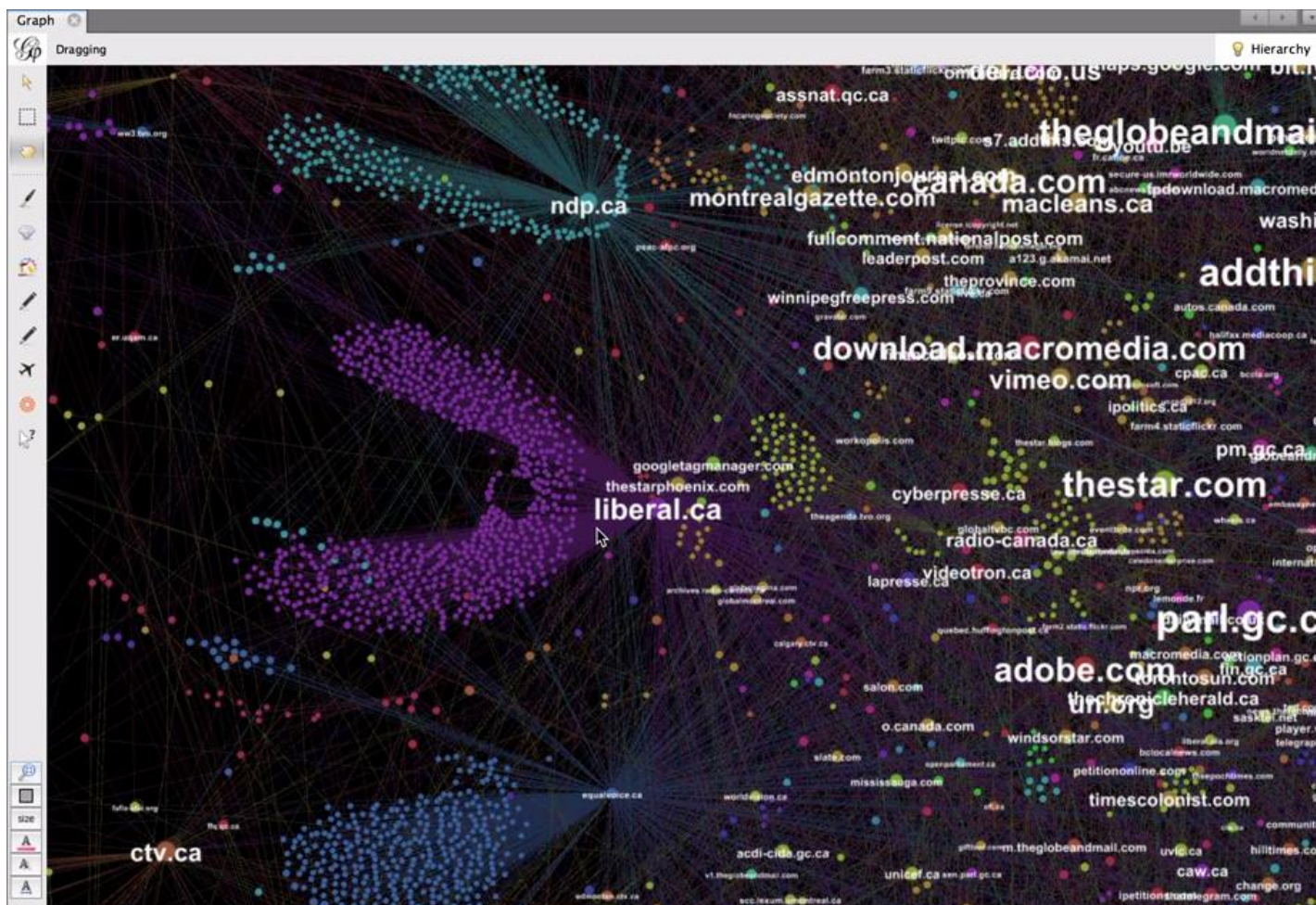
“[Light Writing - Spider Web](#)” by [oz dean](#) under [CC BY-ND 2.0](#)

opportunities for preservation




“standing out” by [kenda bustami](#) under [CC BY 2.0](#)

opportunities for research



“Exploring the Canadian Political Interest Group and Political Parties Web Sphere” by [Ian Milligan](#) under [Standard YouTube License](#)

not the only one



University of California
CDL
California Digital Library

November 10, 2015
About CDL
Services and Projects
Information Gateways

Home > News and Media > CDLINFO News > Announcing a New Partnership: California Digital Library, UC Libraries, and Internet Archive's Archive-It Service

Announcing a New Partnership: California Digital Library, UC Libraries, and Internet Archive's Archive-It Service

January 14, 2015 | Category: Digital Preservation (UC3),Newsletter,Web Archiving | Author: Rosalie Lack

The CDL and the UC Libraries are partnering with Internet Archive's Archive-It Service. In the coming year, CDL's Web Archiving Service (WAS) collections and all core infrastructure activities, i.e., crawling, indexing, search, display, and storage, will be transferred to Archive-It. The CDL remains committed to web archiving as a fundamental component of its mission to support the acquisition, preservation and dissemination of content. This new partnership will allow the CDL to meet its mission and goals more efficiently and effectively and provide a robust solution for our stakeholders.

Why now?

Eight years after the release of WAS, we found ourselves at a critical juncture. The constantly changing and ever-increasing complexity of the web poses significant challenges to the current web archiving toolset and requires frequent upgrades to stay ahead. It became clear that there was a significant opportunity cost to maintaining WAS, which would not leave us with the capacity to develop new added-value web archiving services, such as tools for researchers, computational analysis of aggregated archival corpora, or work toward integrating web archives with other format types.


Collaboration is the Solution

In 2014, the CDL held a series of meetings with peer institutions to investigate the possibility of collaborating on web archiving solutions. We ultimately came to the conclusion that running the core web archiving infrastructure is not the best use of our limited resources. Instead, enlisting the services of Archive-It was the most efficient solution because it will permit the CDL and its partners to reallocate their local resources to activities through which they can uniquely add stakeholder value to the baseline function provided by Archive-It.

Thus, the CDL is currently exploring opportunities with Harvard, MIT, Stanford, UCLA, and others to work closely with Archive-It to create an expanded roster of added-value tools and services. Our goals are to define technical needs as well as the organizational structure that can ensure creation of new tools and services and make them broadly available across the community.

[CDL: "Announcing a New Partnership"](#)

Harvard Wiki
Spaces
Browse


Web Archiving Environmental Scan
Web Archiving Environmental Scan Home

Created by Scott Helms, last modified by Abigail Bordeaux on Aug 04, 2015

Welcome to the Web Archiving Environmental Scan space.

The purpose of the environmental scan is to develop a landscape view of current issues, trends and needs in:

- the provision and maintenance of web archiving infrastructure and services
- the collection and provision of web archives to users
- the use of web archives by researchers

We welcome contributions to the wiki. Contributors external to Harvard will need to [request an XID](#) in order to edit. An XID is simply an ID, mainly for people outside Harvard, that gives you a credential to access designated systems. Once you have it, contact abigail_bordeaux@harvard.edu and andrea_goethals@harvard.edu and we'll grant you edit privileges.

Our [help space](#) and the [Confluence User Guide](#) are good spots for initial questions.

5 Child Pages

- [Report Outline](#)
- [About this Project](#)
- [How to Contribute](#)
- [Profiles](#)
- [Areas of Exploration](#)

Copyright © 2014-2015 The [HUL: "Web Archiving Environmental Scan Home"](#)
| [ICommuns Privacy Policy](#) | [Support](#) | [Atlassian Confluence 5.6.3](#)

a response

National Digital Platform Projects funded in August 2015

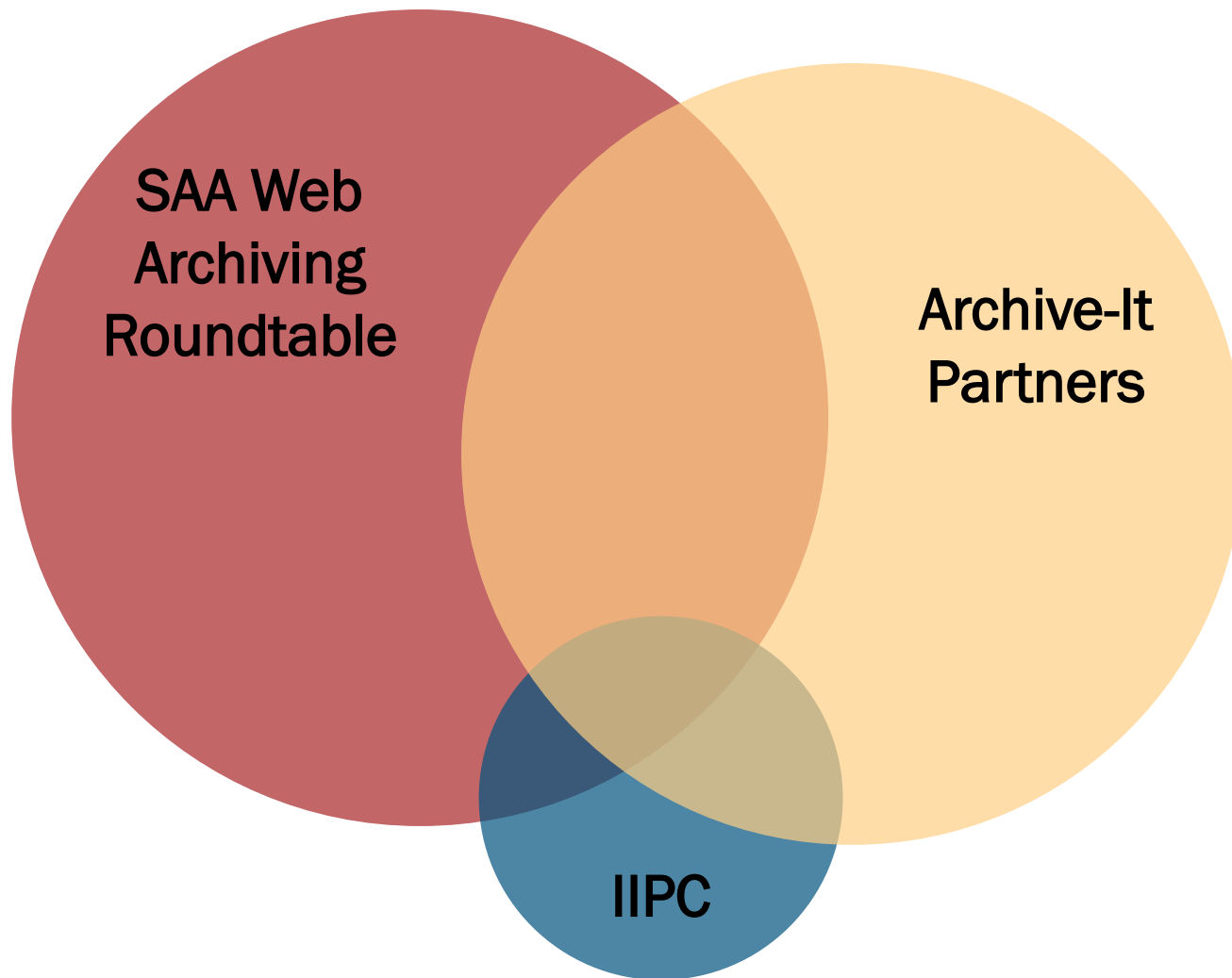
Systems Interoperability and Collaborative Development for Web Archiving

(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.






community analysis



Archive-It

 Archive-It Support

Nicholas | [logout](#)

HOME | FORUMS | [SUBMIT A REQUEST](#) | [CHECK YOUR EXISTING REQUESTS](#)

Forums / Archive-It 5.0 Feature Requests

Search

Archive-It 5.0 Feature Requests

Overview | Recent

Harvesting (8) »

- 💡 View content from test crawls and option to save permanently
- 💡 Ability to resume crawls that have hit the time limit
- 💡 Archive and export specific file types

Access (6) »

- 💡 Embargo periods for collections and/or seeds
- 💡 Updated look and feel of Wayback banner
- 💡 Creation of one Wayback calendar page for linked/identical URLs

Collection Management (including Scop... (24) »

- 💡 Splitting and merging collections; moving & sharing seeds between collections
- 💡 Automatic metadata extraction to improve description (such as PDF files)
- 💡 Ability to browse a seed's entire crawl history

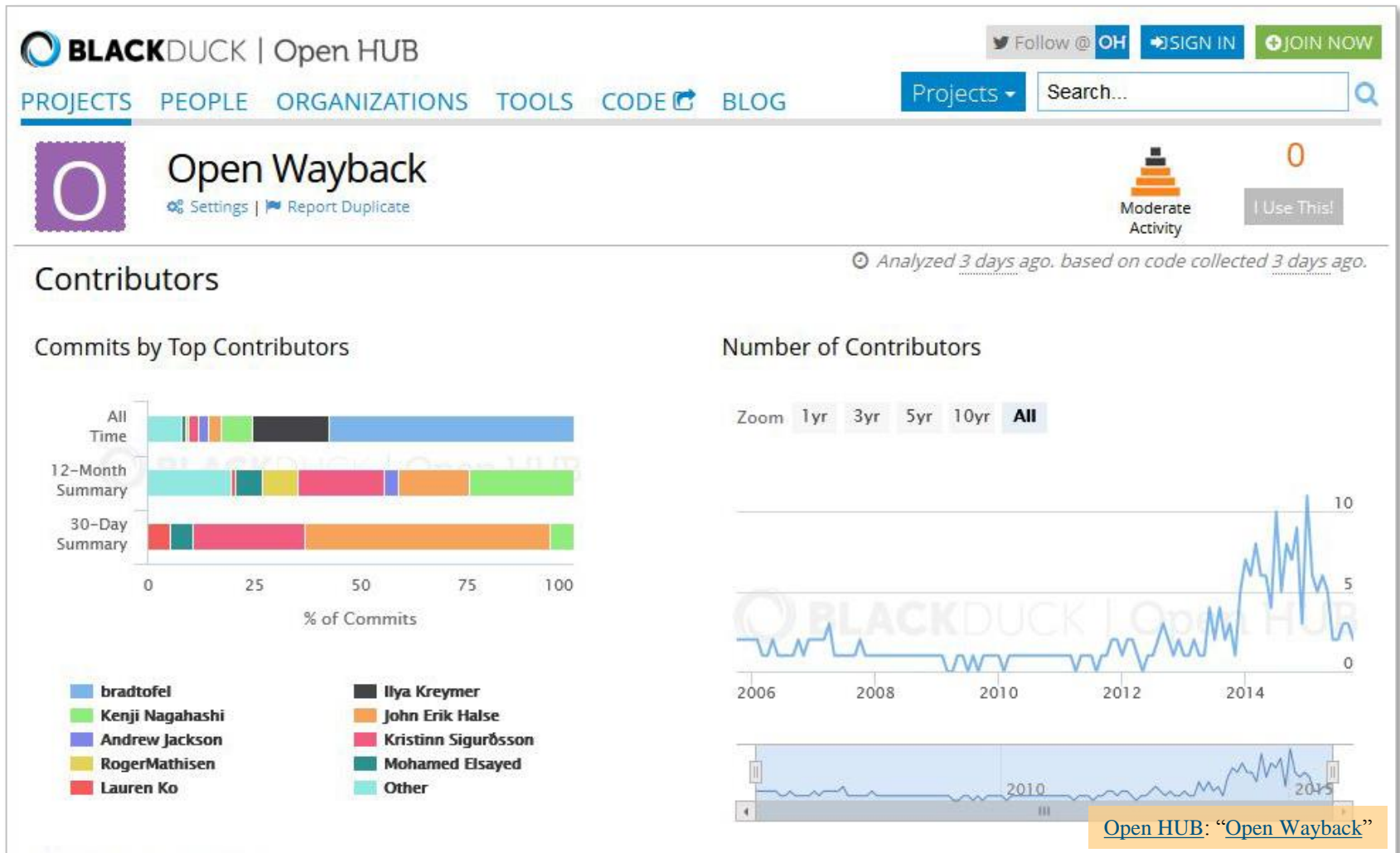
Reports/QA (14) »

- 💡 Remove 'out of scope' or unwanted content from your collection
- 💡 Access statistics for archived content and Archive-It.org homepage
- 💡 Alerts when the capture status of a site has changed since a previous crawl

Support Software by Zendesk

Archive-It: “Archive-It 5.0 Feature Requests”

IIPC



models of software production

(irrespective of license)

- **sole** source
 - *single developer*
- **closed** source
 - *team/corporate dev; no outside contributions*
- **club** source
 - *pool resources for solo/team/corporate dev*
- **community** source
 - *direct and distributed community participation*
- **open** source
 - *grassroots, democratic, meritocratic participation*

club source examples

- Archivematica, AtoM (Artefactual)
- ArchivesSpace (Lyrasis)
- Bitcurator (Educopia)
- Fedora (DuraSpace)
- JHOVE (OPF)
- LOCKSS (Stanford University)
- Omeka (George Mason University)

community source examples



In a Nutshell, Project Hydra...

- ... has had **11,717 commits** made by **105 contributors** representing **73,546 lines of code**
- ... is **mostly written in Ruby** with a **low number of source code comments**
- ... has a **well established, mature codebase** maintained by a **very large development team** with **stable Y-O-Y commits**
- ... took an estimated **19 years of effort** (COCOMO model) starting with its **first commit in October, 2009** ending with its **most recent commit 3 days ago**



In a Nutshell, Blacklight...

- ... has had **2,887 commits** made by **67 contributors** representing **14,774 lines of code**
- ... is **mostly written in Ruby** with an **average number of source code comments**
- ... has a **well established, mature codebase** maintained by a **large development team** with **increasing Y-O-Y commits**
- ... took an estimated **4 years of effort** (COCOMO model) starting with its **first commit in October, 2009** ending with its **most recent commit 2 days ago**

community architecture

- privileges **community** over code
- recognizes **distribution** of investment
- embraces community **diversity**
- models **open** processes and governance
- encourages **varied contributions**
- serves **community needs**

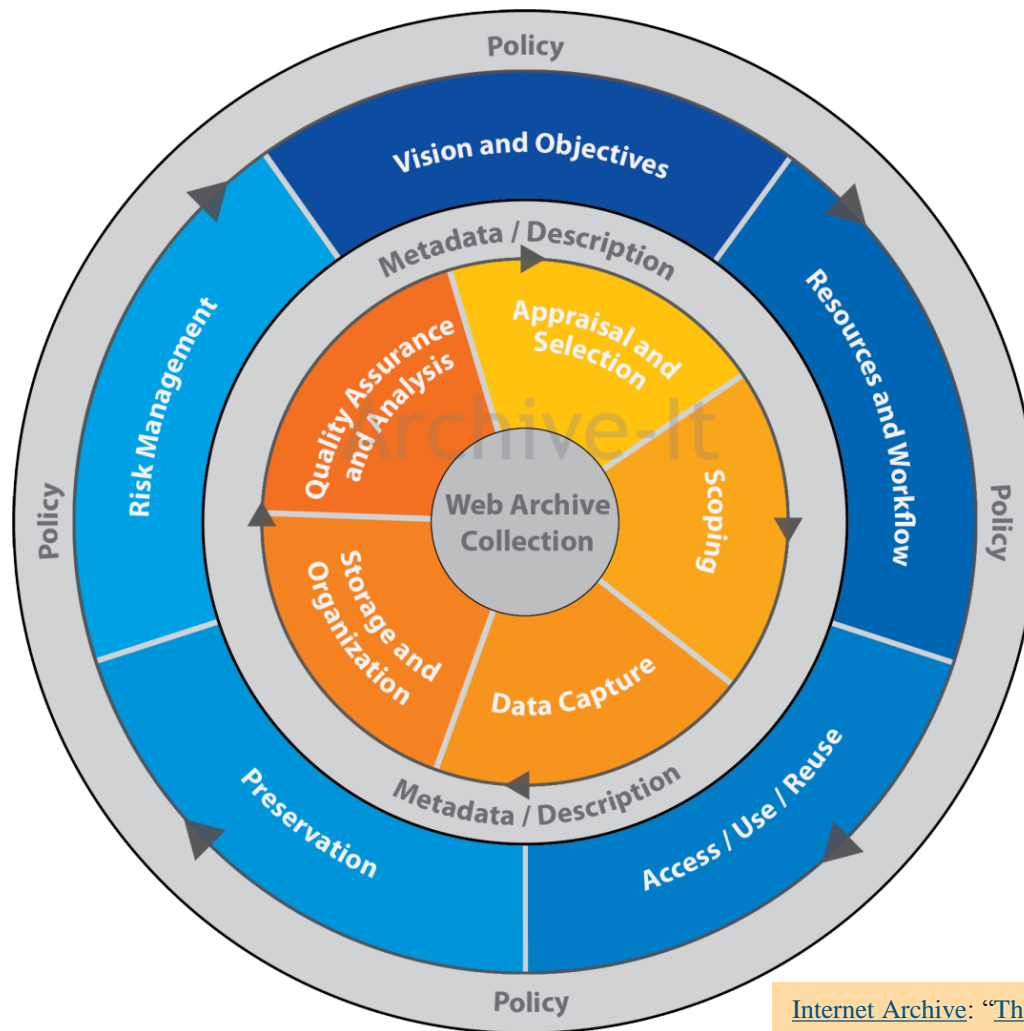


STANDARDS

success of a standard

- **capture:** [DeDuplicator](#), [Heritrix](#), [python-heritrix](#), [SiteStory](#), [WAIL](#), [WARCreate](#), [WarcMITMProxy](#), [WarcProxy](#), [Webrecorder](#), [wget](#), [Wpull](#)
- **access:** [OpenWayback](#), [pywb](#), [warc-proxy](#), [WarcManager](#), [Wayback Machine](#), [Web Archive Discovery](#), [WebArchivePlayer](#)
- **utilities:** [JHOVE2](#), [JWAT](#), [Megawarc](#), [pylibwarc](#), [WARCAT](#), [Warcbase](#), [warctools](#), [Web Archive Commons](#)

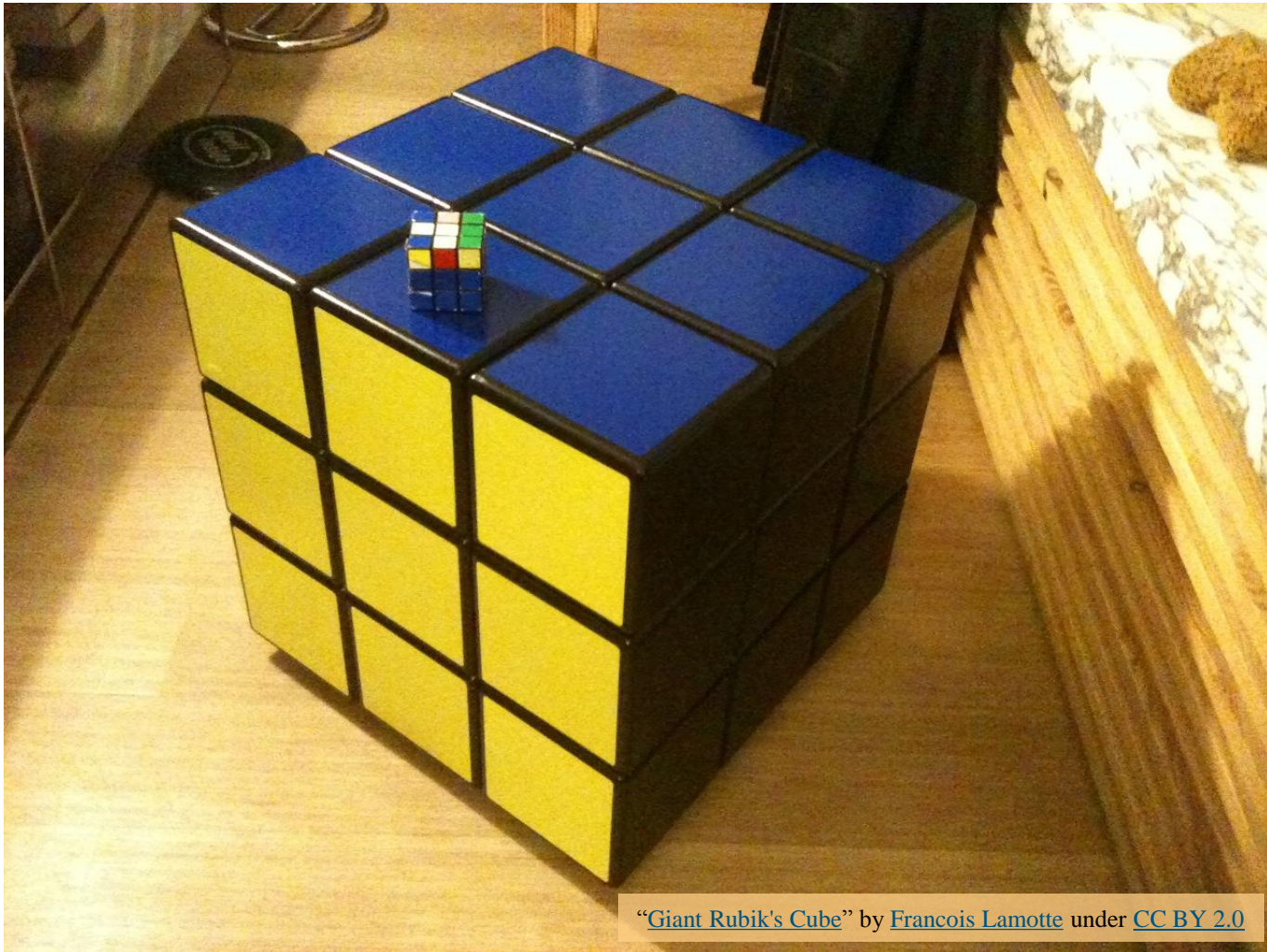
web archiving lifecycle



missed opportunities?

[illegible]

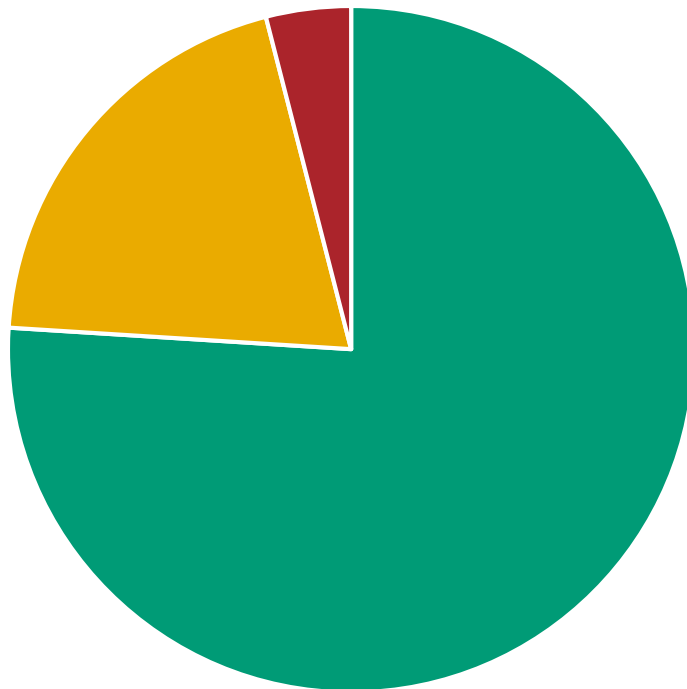
smaller, modular components



“Giant Rubik's Cube” by [Francois Lamotte](#) under [CC BY 2.0](#)

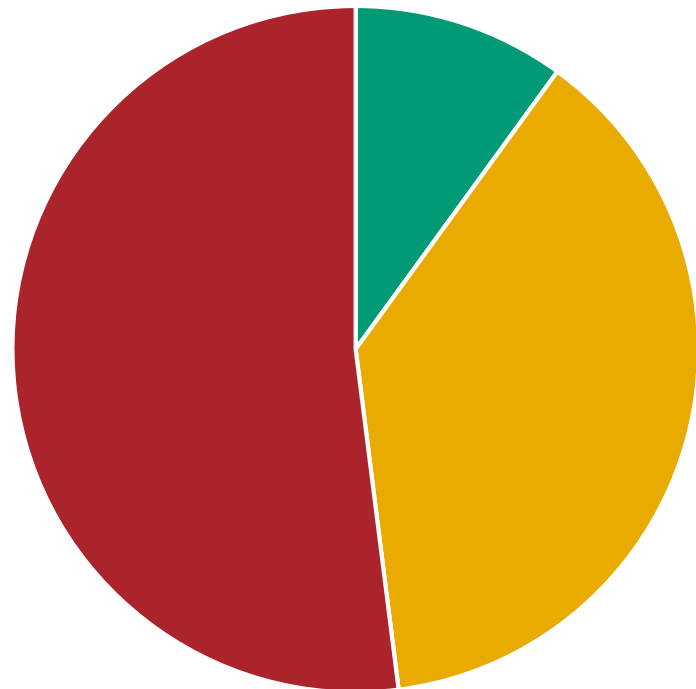
smaller projects do better

small projects (<\$1 million)



■ on time/budget ■ challenged ■ failed

large projects (>\$10 million)



■ on time/budget ■ challenged ■ failed

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

IIPC community interest in APIs

contribution type	% of respondents	# of respondents
help define functional requirements	94%	15
contribute use cases	81%	13
help define technical details	69%	11
help schedule and run meetings	19%	3
implement and test	6%	1

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

