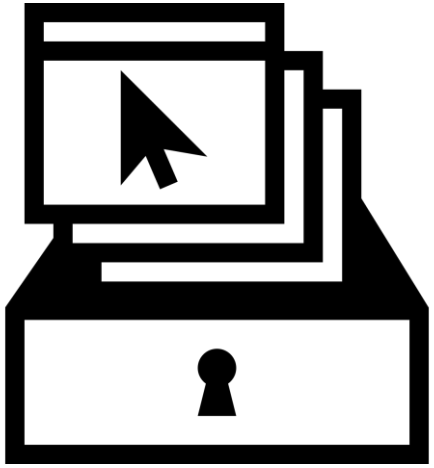
STANFORD UNIVERSITY LIBRARIES

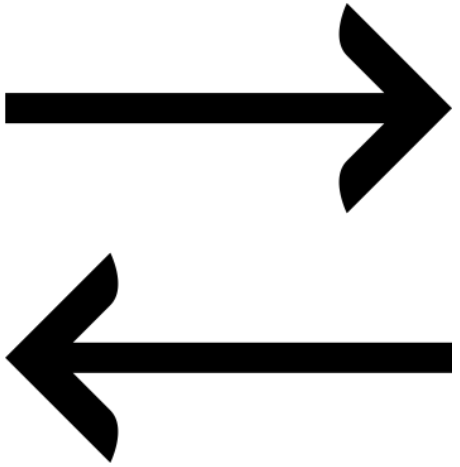
WASAPI Technical Working Group Update

Nicholas Taylor
Web Archiving Service Manager
Stanford University Libraries

IIPC General Assembly: Building API-Based Web Archiving Systems and Services
April 12, 2016

WASAPI





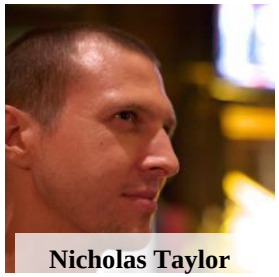
Technical Working Group

Kristine Hanna
Internet Archive / Archive-It

Jefferson Bailey
Internet Archive / Archive-It

Cathy Hartman
University of North Texas

Christie Moffatt
National Library of Medicine



Nicholas Taylor
Stanford University

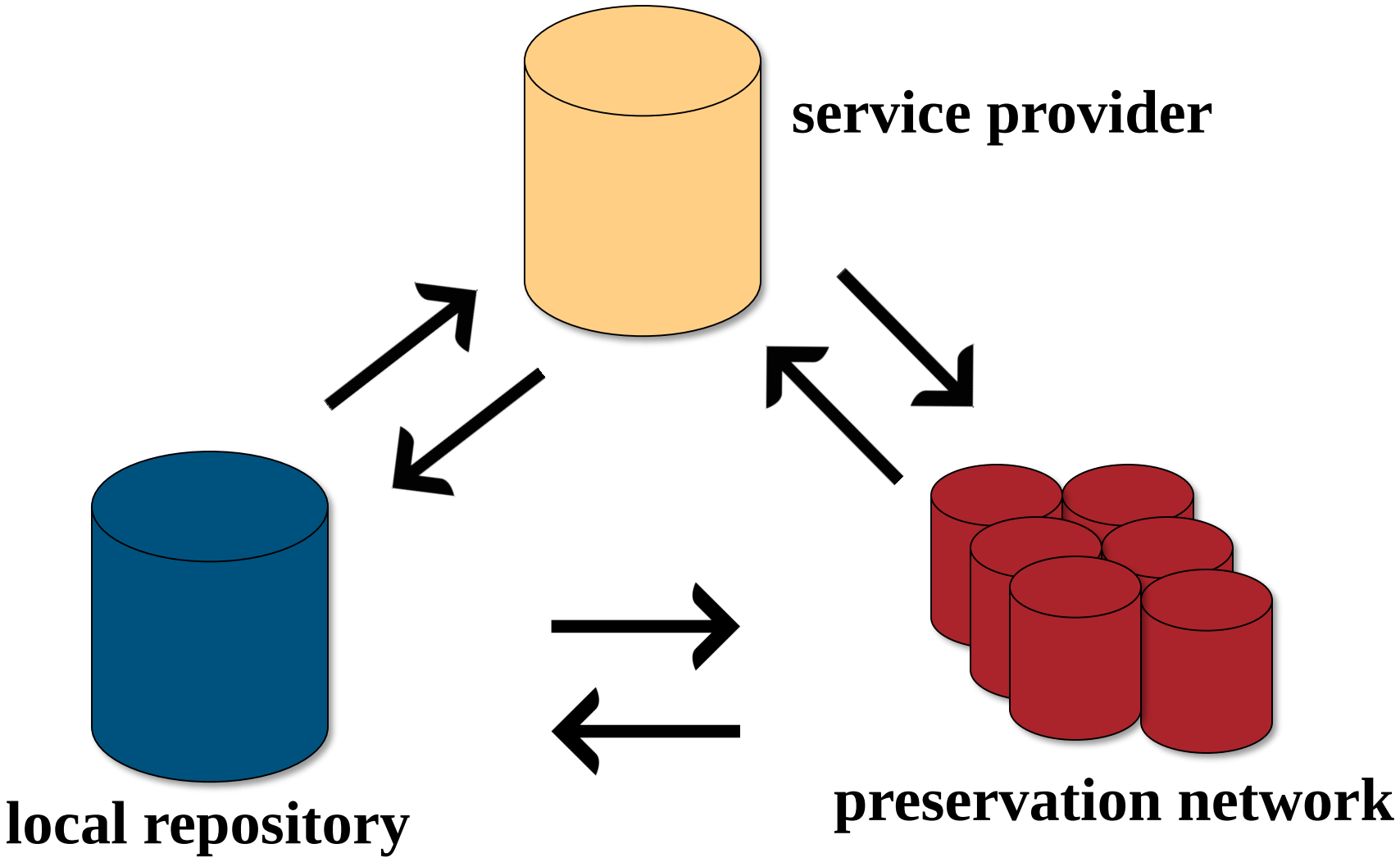
Abbie Grotke
Library of Congress

Edward McCain
University of Missouri

related API work

- CDX Server API (IA, IIPC)
- derivative formats (Archive-It, BL)
- crawl logs/partner data (Archive-It)
- Wayback Machine APIs (IA)
- proliferating capture tools (GWU, IA, Rhizome)
- Cobweb (CDL, Harvard, UCLA)

data flow



test cases

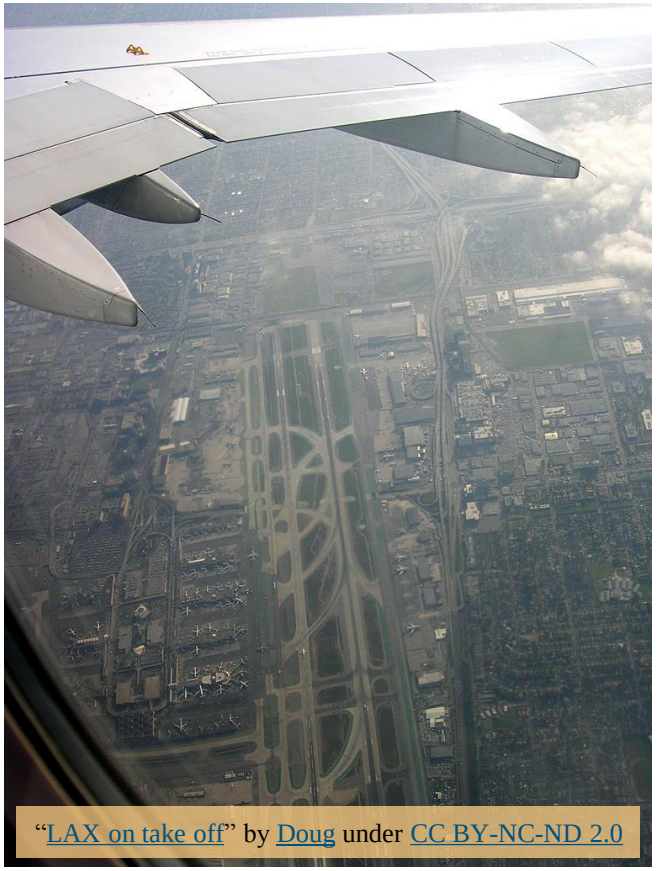
- Archive-It →
  - partner IR/local use
  - DPN
  - LOCKSS (PLN)
- CDL → Archive-It (migration)
- DLSS → IA (WebBase)
- [EoT partners] ← → [EoT partners]
- IA global Wayback→
  - LOCKSS (OA content)
  - national libraries
- LOCKSS (.gov) → IA
- [any web archive] →
  - researcher
  - original publisher

questions

- what's in extension vs. core?
- what abstracted elements sufficient for crafting request across archives?
- what co-bundled metadata?

overview

- Stanford Web Archiving
- CDL WAS Transitioning
- A more collaborative future



"LAX on take off" by Doug under CC BY-NC-ND 2.0

STANFORD WEB ARCHIVING

"Stanford Dish" by Ed Bierman under CC BY 2.0

web archiving activities

- LOCKSS

1999 – present

- WebBase

2001 – 2012

- Archive-It

2007 – present

- CDL WAS

2008 – 2015

Middle East Politics collection

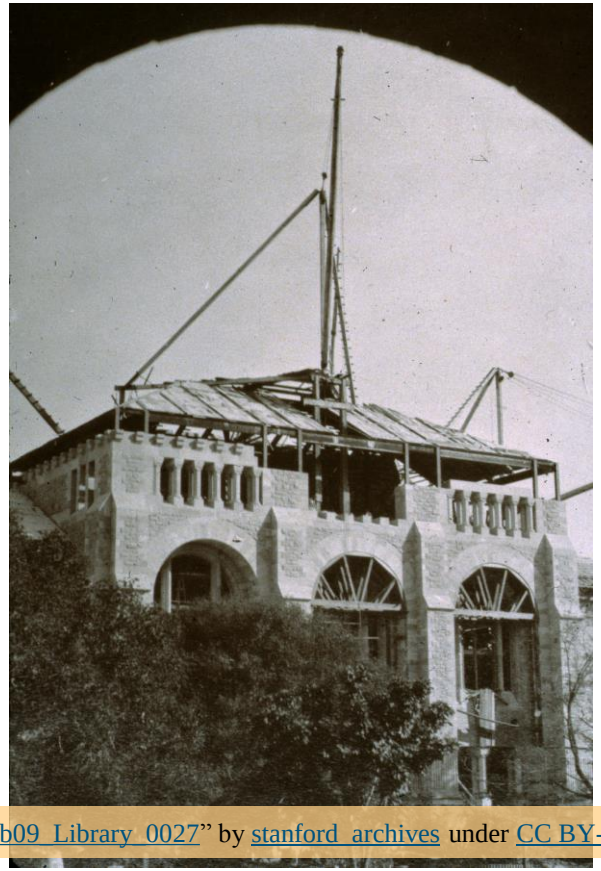
- duration: 2008 – 2015
- size: ~10 TB
- count: 185 websites
- contents: blogs, political orgs, NGOs

African Politics collection

- duration: 2008 – 2015
- size: ~15 TB
- count: 199 websites
- contents: campaigns, news, political parties

Digital Library Buildout 2

- identify needs
- secure funding
- programmatize
  - staffing
  - use cases
  - policy
  - collection development
  - service model
  - technical architecture



"PC0141_b09_Library_0027" by stanford_archives under CC BY-NC-SA 2.0

"Transition" by RicardoRQ under CC BY-NC-SA 2.0

challenges

quality assurance

- backlog
- purge soft 404s

data accessioning

- ingest congestion
- non-working workflows

data transfer

- data volume
- retrieved everything?
- checksums match?

description + discovery

- crosswalk metadata
- improve metadata

challenges

quality assurance

- backlog
- purge soft 404s

data accessioning

- ingest congestion
- non-working workflows

data transfer

- data volume
- retrieved everything?
- checksums match?

description + discovery

- crosswalk metadata
- improve metadata

challenges

quality assurance

- backlog
- purge soft 404s

data accessioning

- ingest congestion
- non-working workflows

data transfer

- data volume
- retrieved everything?
- checksums match?

description + discovery

- crosswalk metadata
- improve metadata

challenges

quality assurance

- backlog
- purge soft 404s

data accessioning

- ingest congestion
- non-working workflows

data transfer

- data volume
- retrieved everything?
- checksums match?

description + discovery

- crosswalk metadata
- improve metadata

A MORE COLLABORATIVE FUTURE

"There's No Place Like The Death Star" by JD Hancock under CC BY 2.0

share collection content

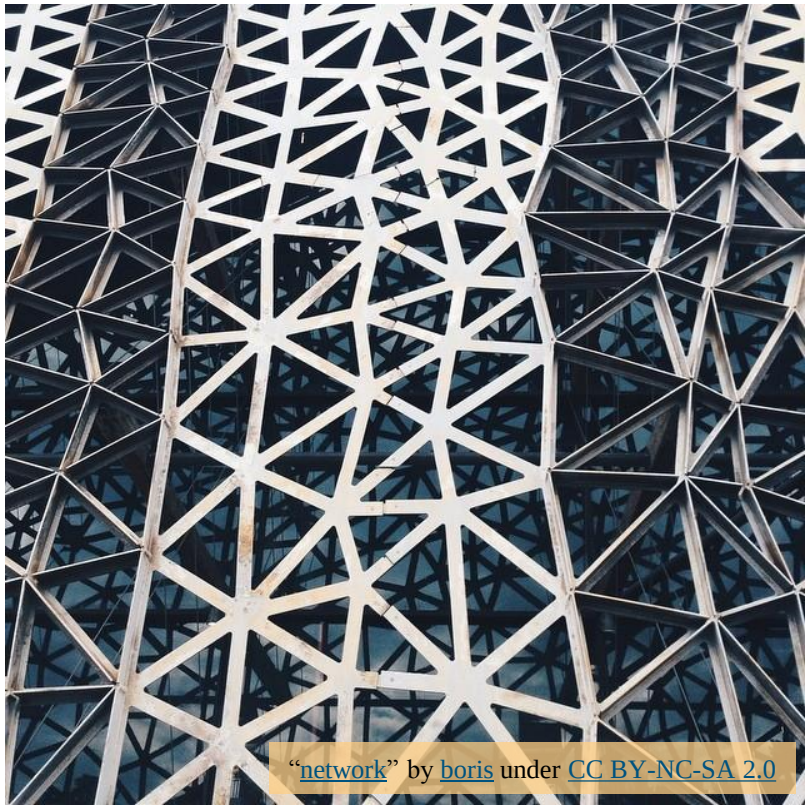
- advantages
  - larger, unified collection(s)
  - distributed preservation
- challenges
  - missing/mixed provenance
  - institutional ownership
  - ad hoc data transfer
  - redundant effort
- opportunity: data transfer APIs (WASAPI)

collaborative collecting

- advantages
  - distribute curation costs
  - more comprehensive collection
- challenges
  - curatorial roles
  - cost sharing
  - institutional ownership
- opportunity: collaborative collecting interface (Cobweb)

distributed services

- changing landscape
  - CDL transition
  - Archive-It predominance
  - Harvard environmental scan
- community interest in APIs
- SUL (web archiving + LOCKSS) needs



let's combine forces

"Stages of flow" by Peter Thoeny under CC BY-NC-SA 2.0