SUL Web Archiving Before and Beyond the CDL WAS Transition

Nicholas Taylor
Web Archiving Service Manager
Stanford University Libraries

IIPC General Assembly: Collaborative Initiatives and Opportunities in the Wake of the CDL WAS Transition
April 12, 2016
overview

• Stanford Web Archiving
• CDL WAS transitioning
• A more collaborative future
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
original architecture plan (local)
actual architecture (hybrid)
planned architecture (hybrid)
challenges

quality assurance
- backlog
- purge soft 404s

data transfer
- data volume
- retrieved everything?
- checksums match?

data accessioning
- ingest congestion
- non-working workflows

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 transfer
• data volume
• retrieved everything?
• checksums match?

data accessioning
• ingest congestion
• non-working workflows

description + discovery
• crosswalk metadata
• improve metadata
challenges

quality assurance
• backlog
• purge soft 404s

data transfer
• data volume
• retrieved everything?
• checksums match?

data accessioning
• ingest congestion
• non-working workflows

description + discovery
• crosswalk metadata
• improve metadata
A MORE COLLABORATIVE FUTURE
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, UNT Nomination Tool)
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