Exploring a National Collaborative Model for Web Archiving

Andrea Goethals
Harvard University

Stephen Abrams
California Digital Library
Why cooperate?

• The problem is too large for any one organization to respond effectively
  – Web scale
  – Technological “arms race”
    • Static HTML4/PDF ⇔ Dynamic JavaScript, AJAX, Flash, HTML5, paywalls, ...
“Web archiving is hard”

• CDL case study (WAS)
  – Production service since 2008; currently supporting 42 curatorial units, 280 collections, 135 TB
  – Relying on the “standard” FOSS stack: Heritrix, Nutch, OpenWayback, with lots of Ruby/Rails/Rake “glue” and a locally-developed curatorial interface
  – Large infrastructural footprint: 11 servers, 100 TB DAS (staging), and 150 TB SAN (archival/access)
  – Approx. 2.5 FTE just to meet operation demands
    • Little time available for necessary improvements: Heritrix/OWB upgrades, Nutch to Solr replacement, deduplication
“Web archiving is hard”

• Harvard case study (WAX)
  – Production operation since 2009; currently supporting 3 curatorial units
  – Reliance on IIPC software (Heritrix, Wayback, NutchWAX, hcc), general open-source tools (Quartz scheduler, Tomcat, JBoss, Hadoop), custom Java modules to control the process (Harvester, Importer, Indexer, Archiver), and custom curatorial interface
  – Bad timing (2009 start of Library reorg)
“Web archiving is hard”

- Harvard case study (WAX)
  - 2009-2014 WAX stagnates - years of technical debt - the underlying software hasn’t been upgraded - many versions behind, still using ARC, still only used by 3 curatorial units
  - Estimated 2.5 FTE for one year to upgrade WAX and expand curatorial units; then 3 FTE on an on-going basis; 2 FTE is closer to what Library wants to commit to providing a service to curators
Benefits of collaboration

• Enable the collection of a rich body of Internet content from around the world
• Foster the development and use of common tools, techniques and standards that enable the creation of international archives
• Encourage and support national libraries, archives and research organizations everywhere to address Internet archiving and preservation

— International Internet Preservation Consortium (IIPC)
Steps toward collaboration

• CUL-hosted *Web Archiving Policies and Practice in the US* summit, May 2012
  – “… an articulation of a small number of model programs for web archiving, and development of ‘best practices’ for documenting program elements”
  – [https://webarch.cul.columbia.edu/](https://webarch.cul.columbia.edu/)
Steps toward collaboration

• CDL-hosted summit, June 2014
  – CDL, Columbia, George Washington, Harvard, IA, LC, North Texas, Stanford
  – “... more robust collaboration was desirable in order to collectively address these challenges [research use, intensive resource requirements, the pace of change, fragmented collection development, etc.] and went so far as to brainstorm the benefits and risks of an all-in, formal association”
  – https://docs.google.com/document/d/1QxwdpUQxzG0vlf3bNZG3G7_Ln8B3OQOK19a7TESIhBM/edit
Steps toward collaboration

• Community Principles for Web Archiving at Scale
  
  “... a lightweight structure by which web archiving institutions can work collectively in order to achieve significant functional goals and operational efficiencies that they are unlikely to achieve individually”

  [https://docs.google.com/document/d/1Qfg1nDdzTuAhtK9NKuooMdpGtP4PKwZCBxS5BqjFMb0/edit](https://docs.google.com/document/d/1Qfg1nDdzTuAhtK9NKuooMdpGtP4PKwZCBxS5BqjFMb0/edit)
The key step

• Recognizing the need to enable centralized, coordinated, and/or local tool development, operation, and collection building

• Defining a comprehensive set of APIs that expose function at critical junctures in nominal workflows

• “Commodity solutions when available, customized solutions when necessary”
Potential architecture
Pursuing collaboration

- IMLS NLG preliminary proposal, February 2015
  - CDL, Columbia, George Washington, Harvard, IA, LOCKSS, MIT, North Texas, NYARC, Stanford, UCLA
  - Environmental scan, community development, technical collaboration
  - Unfortunately, not invited to submit final proposal
  - All partners agree to continue to work together and plan to resubmit in 2016
Pursuing collaboration

• IIPC


    • Collaborative activities: risks DB and assessment tool, and environments DB

  – Preservation working group (PWG) survey results (May 2015)

    • APIs “of interest” to 100% of respondents
    • 94% willing to participate in new IIPC API working group
Summary

• Widespread recognition of the benefits of collaborative approaches
• Willingness to work together to define APIs
• Continue to look for funding opportunities to help facilitate this effort