Collaboration Beyond Crawling

Jefferson Bailey, Internet Archive

ARLIS 2015 | Texas
March 21, 2015
Talk Outline-ish Stuff

• Obligatory Caveats & Intros
  – Beyond “what is web archiving” & beyond answers

• Community Approaches
  – Collaborative acquisition and...

• Techno-sorta-logical
  – R&D 4 You & Me
  – Distributed collaboration

• User Services
  – The great derive

• “The Access Turn”
  – What’s the way(back) forward?
• We are a non-profit Digital Library & Archive
• Founded in 1996 by Brewster Kahle
• Located in San Francisco California
• 20+PB unique data: 10PB web, ~8m text, 2m vid, 2+m aud, 100K soft, etc
• We work in a former church and it’s awesome
• **https://archive.org/web**

• Largest publicly available web archive in existence

• 470,000,000,000+ URLs (that’s billions)

• Like a billion websites, domain agnostic

• Content in 40+ Languages

• Periodic snapshot of the web; 1b+ per week
Archive-It

- Launched in February 2006
- [https://archive-it.org/](https://archive-it.org/)
- Web based application for creating archival collections of web content that includes access and storage
- Institutions can create, manage, access and store collections of web content
- Provides tools for selection, scoping, QA, metadata
- Archived web content includes: html, text, videos, audio, social media, PDF, images, password protected content, static databases and RSS
350+ Partners
2900 public collections
10,200,000,000+ URLs
Web Archiving museums/art libraries

- Hagley
- Arab American National Museum
- The Clark Art Institute
- National Gallery of Canada
- Musée des beaux-arts du Canada
- Shangri-La Center for Islamic Arts and Cultures
- New York Art Resources Consortium
- 9/11 Memorial
- National Museum of Women in the Arts
- Smithsonian Institution Archives
## How is Archive-It different than the General Archive/Global Wayback

<table>
<thead>
<tr>
<th>General Archive:</th>
<th>Archive-It:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• One collection</td>
<td>• Focused collections</td>
</tr>
<tr>
<td>• Periodic snap shot</td>
<td>• Control of scope and frequency</td>
</tr>
<tr>
<td>• Automated crawling</td>
<td>• Indexed for full text search &amp; metadata</td>
</tr>
<tr>
<td>• Automated QA</td>
<td>• Archived data shipped/downloaded</td>
</tr>
<tr>
<td>• URL-specific discovery</td>
<td>• Available 24 hours after captured</td>
</tr>
<tr>
<td>• Public access</td>
<td>• Restricted access options</td>
</tr>
<tr>
<td>• Absolutely free</td>
<td>• Subscription service</td>
</tr>
</tbody>
</table>

The General Archive provides one collection in a periodic snapshot approach, with automated crawling and QA, and allows for public access. Archive-It, on the other hand, focuses on collections with controlled scope and frequency, offers indexed full text search and metadata, and includes options for downloaded archived data, availability 24 hours after capture, restricted access, and a subscription service.
Community Approaches
Community Approaches

• Collaborative Initiatives
  – IIPC Olympics Web Archives
  – 2014 Election Crawling
• Consortial Initiatives
  – NYARC (represented here!)
  – LIPA (group of law libraries)
• Crowdsourced Collections
  – Public URL nominations
• Spontaneous Events Collections
  – Government Shutdown, Boston Marathon, Ukraine Collection
• Education Initiatives
  – K-12 and LIS programs
The exegesis of collaborative collecting

- Selection coordination is a defined model
  - Not necessarily a web-specific model
- Pivot from collaborative collecting to collaborative infrastructure remains challenging
- Collaborative collecting both remediates but can also perpetuate challenges of staffing
- The Long Tail of Crowdsourcing
The potential of collaborative collecting

• Better visibility into collecting activities
  – Seed registries, points of aggregation

• Archives & special collection beyond the medium
  – Better integration with non-web content

• Capitalize on the web’s populist principles
  – Ephemera, colloquia, evidentiary “lived experience”

• Shared policy & assessment
  – Quality assurance, permissions tools, curatorial strata
NDSA 2013 Web Archiving Survey


FIGURE 9: PARTICIPATION IN A COLLABORATIVE WEB ARCHIVE
Techno-sorta-logical
“Technological projects that remain purely technological are like moralists: their hands are clean but they don’t have hands” – Latour, Aramis or The Love of Technology
Lifecycle vs. development cycle
R&D 4 You & Me

- Archiving contemporary composters websites
  - NYU Libraries & Archive-It (via Mellon)
  - Collect primary source web-only materials; involve content creators; make discipline specificity technically expansible; bridge preservation data stores
  - Curatorial, bibliographical, technological, interoperable

- Other collaborative development (IIPC, OSS)
  - Challenges to technical interoperability
    - Few code commits
    - Divergent implementation
    - What’s the “core”
Technical Distribution as Resource Aggregation

- **Modular-ization**
  - Granularity of micro-services per lifecycle stage
- **Interoperability**
  - Systems they are a talkin’
- **Increased modularization enables distribution**
  - Will collaboration allow different institutions to do different pieces of the lifecycle
- **Say hello to the user’s users!**
User & Access Services
Research Services Datasets

WAT Datasets
(Web Archive Transformation)
Key Metadata from Every Resource

LGA Datasets
(Longitudinal Graph Analysis)
What Links to What over Time

WANE Datasets
(Web Archive Named Entities)
Names of People, Places, Organizations

Ferguson Combinatory Dataset

URL Contributors

- Washington University, Internet Archive, MITH, public
- [https://archive-it.org/collections/4783](https://archive-it.org/collections/4783)

Twitter crawl by MITH

- ~15m tweets; IA extracted & crawled 500K URLs ~50K YouTube videos

Forthcoming release of full public dataset
“The Access Turn” for Web Archives
“The Access Turn”

**WEB ARCHIVING IN THE UNITED STATES: A 2013 SURVEY**

**FIGURE 4: YEAR INSTITUTIONS BEGAN ARCHIVING WEB CONTENT**

- Archive
- College or University
- Commercial
- Consortium
- Government: State or Federal
- Museum
- Other (please specify)
- Public Library
“The Access Turn”

• Program maturity and longitudinal breadth
  – How to move from collection building to collection use?
  – How can we think beyond the item/collection record as a synthetic point of convergence?
  – What tools and access models will most benefit user communities and custodial institutions?
  – How can access be a new point of collaboration for collecting, use, and R&D?

• Summarizing
THANKS!

Jefferson Bailey, Internet Archive
jefferson@archive.org | @jefferson_bail
(I have free stickers)

Internet Archive
https://archive.org/

Archive-It
https://archive-it.org/