Archive-It and Web & Data Services
News and Notes!

Jefferson Bailey, Director

July 21, 2021
USAGE STATS: Q2 2021

- Archive-It:
  - 106 TB (+4%)
  - 2.1B URLs (+4%)
- 42K seeds (+2%)
- 53K crawls (+7%); 18K test (+9%)
- 660 new collections (+4%)
- 419 Archive-It new support tickets; 414 resolved

- Managed collecting:
  - 100+ TB
PARTNER STATS: Q2 2021

● 64 new service partners
  ○ Archive-It services: 20
  ○ Community Webs: 44

● 20 additional service requests
  ○ 6 Waybackfills
  ○ 3 Web Snapshots

● Managed services:
  ○ Work with 11 different National Libraries/Archives
    ■ .au, .nz, .lu, .is, .pl, .ca, .de, [uh, .nara?]
ARCHIVE-IT SERVICE UPDATES: Q2 2021

Additional services

- Archive-It Sponsored
- Archive-It Basic, Waybackfill Service, Redirection Service [product sheets]

Engineering updates and roadmap

- Python Wayback in QA (thanks to our beta testers!)
- Quality Assurance and “rules” tools
- Ongoing social media capture development
- Account management automations & CRM integrations
- Preliminary work on “moving seeds”
• Significantly improved replay fidelity for javascript heavy sites
• New system to roll out for most Archive-It users this summer
• Archive-It to complete migrating all Wayback Machine deployments by end of year
• Open Source: https://github.com/internetarchive/rulesengine
• Included in Archive-It’s Python Wayback
• Set policies for replay access control
• Fine-grained control for capture-specific replay fixes
• Maintains support for extensive site-specific replay fixes made by Archive-It and partners
• Will be made available to the Archive.org Wayback Machine
• Archive.org and Archive-It.org replay fixes will be publicly available
ACQuA: Automated Capture Quality Assurance

- Brozzler takes a (real Chrome browser) screenshot on every page it crawls
- Between Archive-It.org and Archive.org crawling we have millions of images showing how replay is supposed to look
- Batch generate screenshots of Python Wayback replay
- Team build training set rating quality of replay screenshot to Brozzler screenshot
- Run battery of automated image difference algorithms against screenshots
- Make predictive model of image diff against archivist training set
STRATEGIC PROGRAMS UPDATE: Q2 2021

Community Webs

- 50+ new public libraries
- DPLA integration in development
- Expanding worldwide and adding services
- Deadline to apply for fall cohort: August 2

Web-Based Art Resources Preservation and Access

- New NEH grant with partners at NYARC
- Recruiting 30+ American art and museum libraries by July 31
- See Session Presentation
  - scholar.archive.org
  - 30+M open-access web-published scholarly articles archived
  - Automated scholarly article extractor from web crawls
  - Can extract and enhance scholarly literature from any web crawl - testers?
  - Forthcoming release of 1.3B citation graph
  - Upcoming work on secondary scholarly outputs (datasets, code, protocols, etc)
  - Multiple in process partnership
SERVICES DEVELOPMENT UPDATE: Q2 2021

Scaling Research Use of Web Archives ("ARCH")

- First research cohorts launched
- Application UI in beta testing
- Partner testing soon

Digital Repository & Preservation Service ("Vault")

- Intro webinar video & Presentation slides
- 1st cohort group up and running!
- Contact us to participate in pilot
• Integrating Archives Unleashed into the Archive-It web application and back-end systems
• On-demand, auto-generated datasets and analytic tools and visualizations for web archive collections
• Port and rewrite of all analytic code
• New infrastructure and open-source tools
• Partner user testing summer; rollout Q3
• Generalizing for any corpus (beyond existing web collections)
Digital Repository & Preservation Service

- Origins
- Good response to our call for interest, 50+ orgs; highlights of needs assessments
- In pilot testing - join us!
- Key features:
  - Content agnostic storage and preservation
  - Allows for geographic distribution of content across up to 3 countries (US, CA, EU)
  - Fixity audit & repair reporting
- Essentials:
  - Cost-effective with flexible feature pricing
  - Ease-of-use & functional simplicity
  - Integration and API first (IA + others)
  - Non-profit and mission-aligned
Q&A

How can we help?


More updates and news

- Help Center and Community Forum
- Blog
- Twitter feed

Email us anytime

- ait@archive.org / webservices@archive.org
- jefferson@archive.org