Archive-It 5.0

Internet Archive

- We are a Digital Library
- Founded in 1996 by Brewster Kahle
- Located in San Francisco, California

Web Archiving Service: Archive-It

Archive-It is a subscription service launched in February 2006

- **Web based application** that allows users to create, manage, access and store collections of digital content
- The service is a **fully hosted solution**, and includes access and storage.
- **Provides tools for selection and scoping** including cataloging with metadata
- Ability to **capture content using 9 different time frequencies**
- Archived content includes: html, text, videos, audio, social media, PDF, images, online newspapers
- **Can browse archived content 24 hours after a capture is complete**; and full text search is available within 7 days
- **Restricted access options** are available

5.0 PHASE 1
USER INTERFACE

New User Interface

- Cleaner layout and aesthetic
- More intuitive
- User friendly
- Highlights important features
- Streamlined work flow
- Dynamic charts and visualizations to summarize and communicate key crawl, collection, and account information

5.0 PHASE 1 – UNDER THE HOOD

How did we build it?

• Druid: a real time data store

• Django: Python framework

• AngularJS: Javascript frontend

Top 10 things about 5.0 reports

1. Quick filter and search reports
2. Infographics
3. New information: who started a crawl
4. New information: Scoping rules in place at the time of the crawl
5. Add individual notes for specific crawls, and search notes
6. Browse hosts by Seed
7. View archived URLs in Wayback directly from the report
8. Analyze and browse crawled documents by File Type
9. Compare the results of two crawls side by side
10. Toggle between All and New Data

5.0 WINTER RELEASE

Harvesting

– Crawls will have the option to to keep data as part of a collection or delete within a set period of time (tentatively 30 days).

– Embargo periods for collections and/or specific seeds (ability to pre- designate a specific length of time, ie. 3 or 6 months)

– Ability to resume crawls that have stopped or hit the time limit

– View a list of URLs being crawled in real time

Collection Management

– Improved import of metadata, including detailed error reports and confirmation of changes/additions

– Export of metadata, including ability to export metadata in a spreadsheet.

Access

- Firefox plugin for proxy mode

5.0 SUMMER RELEASE

Collection Management

– Ability to move and share seeds between collections

– Refined scoping rules, including option to apply only to specific seeds and frequencies

– Customizable user roles

– View detailed seed history information, including crawl history and post crawl reports.

– Ability to archive/export specific file types such as PDFs or images

Access

– Enhanced integration with General Archive at www.archive.org

– Updated Wayback look and feel (banner, calendar page, etc.)

https://archive-it.org/learn-more/archive-it-5-0-release-schedule