



Archive-It

IWAW'06

Gordon Mohr, Internet Archive

September 22, 2006

1

Worldwide Wayback Machine

- Largest publicly available web archive in existence
- Free to the public at www.archive.org
- Over 60 billion pages culled from over 60 million sites
- 60k unique users per day

2

Archiving for Partners

Technology partner for institutions who wish to create focused collections.

- Specific websites
- Customized depth and frequency of crawl
- Monitoring and reporting on collection activities
- Text search of collections
- Hosting of content
- Storage and preservation

3

Curated Crawls

- Since 2002, partnering with Library of Congress, NARA and other National Libraries, including Australia and France, for thematic and national crawls
 - Domain crawls (.au, .fr)
 - Topic Crawls (Iraq War)
 - Library of Congress collections:
 - Iraq War: 700 Million documents and growing
 - U.S. National Elections
 - 2000: 132 Million documents
 - 2004: 88 Million documents
 - Supreme Court Nomination: 100 Million documents

4

Archive-It

- Early 2005, we saw a need to expand archiving services to a broader range of institutions, cost effectively.
- Web-based service to **create, manage, search** and **store** medium-sized web archives through an easy to use web interface
- Technical expertise or infrastructure not required

5

Archive-It Application

- Pilot launched in September 2005
 - (100,000,000 documents were archived)
- 1.0 Release in February 2006
- 1.5 Release in April 2006
- 2.0 Release in July 2006

6

How is Archive-It different than the Worldwide Wayback?

- Paid subscription service.
- Archive-It creates distinct "collections", containing only the content the subscriber is interested in harvesting, at whatever frequency suits their needs.
- All collections are full-text searchable.
- The collections created with Archive-It can be catalogued and managed directly by the subscriber.
- We keep a minimum of two copies of each collection online.
- If a subscriber ends their service, their collections will be moved over to the general archive for public access via the Wayback Machine.

7

Archive-It Partners & Trial Partners

- | | |
|------------------------------------|--------------------------|
| • University of Texas | University of Toronto |
| • University of USC | University of Minnesota |
| • Utah State Library | Library of Virginia |
| • University of Hawaii | Tennessee State Archives |
| • New York State Archives | Michigan State Archives |
| • State Archives of South Dakota | State Archives Alabama |
| • State Archives of North Carolina | University of Indiana |
| • State Archives of Nebraska | University of USC |

8

Access

- Partners view their archived collections online through private web application (member login/password)
- Public can view archived collections online, at the public Archive-It site – www.archiveit.org
- Link from partner institutions' website to the Archive-It website, with a template provided by us.

9

Preservation

- Mirrored copies of the archived collections are stored at IA and feed
- Archive-It provides copies of the archived files on a hard drive that is shipped to partners 1x or 2x a year

10

Partner Subscription

- Standard Subscription is \$10,000 for one (1) year
 - 10,000,000 documents
 - 3 collections archived, text searchable and stored
- Customized Subscription is also available

11

Trials 4 – 6 weeks

- One Collection
- 50 seeds
- Up to 5,000,000 URL's

12

Support Contacts

- Molly Bragg, Partner Specialist, Archive-It
 - Mbragg@archive.org 415 561 6799 x 6
- Kristine Hanna, Program Director, Archive-It
 - Kristine@archive.org 415 561 6799 x 5

13