



international internet preservation consortium

GA 2010

National
Library
Building
Singapore
3-7 May 2010

Web Archiving Software Tools

**Mr. Aaron Binns
Internet Archive**

About the speaker

- Mr. Aaron Binns
- Internet Archive
 - archive.org
 - San Francisco, CA, USA
 - Joined Internet Archive in January 2008
- Lead developer on NutchWAX
- IIPC Program Officer
 - Term began January 2010

Web Archiving Software Tools

- Developed by IIPC members
 - Not vendor products
- Open Source
 - Leverage existing OpenSource projects
- Common needs
 - Crawling and harvesting
 - Display/view web archive
 - Analytics

Web Archiving Software Tools

- No single software architecture or technology
 - Variety is the spice of life
 - Different software solutions to different needs
 - Often differences in institutional goals and policy drive software solutions
- Toolkits vs. tools
 - Configuration, customization, extensibility
- Much more than can be covered in 30 minutes
 - Highlight a few key software projects

WARC: Web Archive Data Format

→ Data

→ Captures everything in HTTP transaction

- DNS lookups
- Full HTTP request and response

→ Metadata

→ Crawl configuration, format conversions, etc.

→ Flexible and extensible

→ Developed by IIPC

→ ISO standard 28500:2009

Software: WARC Tools

- IIPC-funded project
 - Developed by Hanzo Archives
 - Open Source
 - Implemented in C with bindings for Java, Python and other languages
- Read, write, validate, extract WARC records
- Conversion from legacy format to WARC format

Web Harvesting: Heritrix

- Heritrix is the premier open-source, extensible, web-scale, archival-quality web crawler
- Developed by Internet Archive with contributions from IIPC members
- Highly configurable, extensible and scalable
 - Domain crawls exceeding 1,000,000,000 URLs
- Writes WARC format

Heritrix continued

- Implemented in Java
 - Configured operated and managed via web interface and XML files
- Current versions
 - Legacy: 1.14.3
 - Latest: 3.0
- Migration to version 3.0
- Heritrix forums: 2 per year

Viewing web archives: Wayback

→ Wayback Machine

- OpenSource implementation of Internet Archive's popular Wayback Machine

- Java web application

- Scalable: billions of URLs

- Serve entire web archive: HTML, images, audio, video, etc.

- Customizable and extensible

- Many deployments in IIPC member institutions

Integrated Systems

- Define, schedule, run and manage harvests
- Implement harvest policies
 - Access control
 - Scope
- Quality Assurance
 - Test or missing content
 - Feedback for future harvests
- Non-technical audience: librarians, curators.

NetarchiveSuite

- Developed by
 - The Royal Library of Denmark
 - The State and University Library of Denmark
- Event, selective, and domain harvests
- Usable by librarians and curators with a minimum of technical supervision
- Easy setup and automated bit-integrity checks
- Uses Heritrix & Wayback as components

Web Curator Tool

- Developed by
 - National Library of New Zealand
 - British Library
- Primarily for selective web archiving
- Permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata
- Non-technician friendly, librarians, curators
- Uses Heritrix & Wayback as components

Full-text search, analytics, etc.

→ NutchWAX

- Full-text search for web archives

- Based on Apache Nutch project

→ JHOVE

- Digital object format validation

- Images, PDF, HTML, etc.

- JSTOR and Harvard University

→ Analytics: Lewis Crawford (BL) Thursday

- “Big Tools for Big Data”

Future Directions

- Tools adapt to the changing web
 - Web 2.0
 - Flash
 - Streaming media: especially video
- Scaling up
 - Domain harvests: 1,000,000,000+ URLs
- Scaling down
 - Highly selective and thematic harvests

How to get started

- We are here to help!
- Extensive online documentation
- Discussion forums and email lists
- Workshops
- Talk to IIPC members who have similar needs and goals to your own.

References

→ WARC: ISO 28500:2009

→ <http://archive-access.sourceforge.net/warc/>

→ WARC Tools

→ <http://code.google.com/p/warc-tools/>

→ Heritrix, Wayback, NutchWAX

→ <http://crawler.archive.org/>

→ <http://archive-access.sourceforge.net/projects/wayback/>

→ <http://archive-access.sourceforge.net/projects/nutchwax/>

References

→ Netarchive Suite

→ <http://netarchive.dk/suite>

→ Web Curator Tool

→ <http://webcurator.sourceforge.net/>

→ JHOVE

→ <http://hul.harvard.edu/jhove/>