A Busy Hive Creates Better WAX

Archiving the Web from many Perspectives

DLF Fall Forum 2008
Andrea Goethals, Wendy Gogel – Harvard University Library
Today’s Agenda

• An Introduction to WAX
• Dissolving Sticky Challenges
• Still Sticky Challenges
• A Sneak Peek at WAX
An Introduction to WAX

• A busy hive…
Project Goals

- 2.5 year pilot project
- Solve collecting problem for curators, while:
  - assessing technical feasibility;
  - exploring legal terrain
  - gaining experience in the domain
  - quantify the resource requirements including human effort, hardware and software, and financial impact
  - investigate sustainability of a Harvard web archiving service
Who is involved?

• Collection managers (librarians, archivists, faculty)
• Legal counsel (University and external lawyers)
• Technologists (architects, programmer, managers, graphic designers, preservationists)
What is involved?

- Selection
- Rights management
- Acquisition/crawling
- Quality assurance (QA)
- Organization/collection creation
- Metadata
- Archiving/storage
- Presentation
- Searching/indexing
- Preservation
WAX and WAXi Components

3rd Party Software:
- Heritrix
- Hcc
- JBoss
- NutchWA X
- Oracle
- Quartz
- Struts
- Tomcat
- Wayback

Front End
- WAX Management Interface
  - Hollis (Union Catalog)
  - WAX Public Display Interface
- Collection Manager
- Public User

Back End
- Temporary Index
- Back-end Services
- Production Index
- WAX File and Metadata Storage
- DRS (Preservation Repository)
WAX and WAXi Activities

- WAX Management Interface
  - selection, QA, description, collection creation, presentation creation

- Hollis (Union Catalog)
  - discover

- WAX Public Display Interface
  - search, browse, view presentations

- Not supported yet:
  - reporting,
  - access restrictions,
  - IP and legal management,
  - deletions

- WAX File and Metadata Storage
  - storage

- DRS (Preservation Repository)
  - storage, preservation

- WAX File and Metadata Storage
  - acquisition, crawling, importing, indexing, archiving
Dissolving Sticky Challenges

...creates better WAX
Operational Efficiency and Automation

Many steps in the path to curated, archived, viewable, searchable harvests...

... which can we automate?
Operational Efficiency and Automation
Operational Efficiency and Automation

• Automating the back end services
  • Crawl Scheduling
  • Crawling
  • Harvest Archiving
  • Harvest Indexing
Recipe for an Automated Back End Service

Service:
- is always running (or is started up periodically)
- looks for a certain condition to trigger an action
- performs action
- changes a state or triggers another service
Automated Scheduling
Automated Crawling
Operational Efficiency and Automation

• How can we help with the Collection Manager’s manual tasks?
  • QA harvests
    • Remove unwanted pieces
    • Detect missing pieces
  • Refinement of seed scope
  • Select and send content to be preserved
1. Keep URIs
2. Exclude URIs from future crawls
3. Delete URIs from harvest
4. Delete URIs from harvest and exclude them from future crawls
One button push to archive
The Ultimate QA and Archive Shortcut

- Can eliminate QA and/or archive push step on a per seed basis
- Crawl → Preservation Repository
Legal Risks and Mitigations
Copyright Infringement

- Permissions
- Fair Use/Fair Play

SUNDAY, OCTOBER 18, 2008

Bad News
It looks as though I am being umm...laid off, I guess near the end of the year.
The store where I work is closing. Yep..thank you. I have no idea where to go from here. It seems my options are pretty minimal.
Any advice would be welcome.

POSTED BY TIFF AT 3:10:00 PM  10 COMMENTS
LABELS: CRANKITUDE, LIFE, WORK

evelin said...
Oh, boy does that suck. I don't know what to do ... at least you have some time to find a new job if you need to. What do you _want_ to do?
4:48 PM

amy said...
Hey Tiff - Sorry for your news. I also just got laid off from my job at the end of September ... I am not sure what to do at this point either ...
5:39 PM

frenchie said...
Tiff! I'm sorry. That sucks a$$.
7:12 PM

statia said...
I'm sorry Tiff. Hopefully you'll be able to find at least something, even if it's just temporary. And now you can figure out what you want to do with the rest of your life.
4:25 PM
Fair use/Fair Play

- Polite crawling
  - Obey robots.txt
  - Leave crawler information in logs
- Capture surface web only
- Adopt Terms of Use
Fair use/Fair Play

- Don’t compete with or divert traffic from live site
  - Exclude robots from the WAX archive
  - Add transformative content
    - Framing
    - Presentation pages with original intellectual content
  - Embargo display
  - Link to live site
- Employ a liberal *Take Down* policy and process
State Tort Liability

• Civil Damages
  • Invasion of privacy
  • Sensitive personal data
  • Commercial content
  • Defamatory content
State Tort Liability

- Employ a respectful “request frequency” during crawls
- Screen content
- Employ a liberal *Take Down* policy and process
- Polite crawling
  - Obey robots.txt
  - Leave crawler notification in logs
- Offsite Crawler
Statutory Content Restrictions

- U.S. or Foreign Content Regulations
  - Child pornography
  - Obscenity
Statutory Content Restrictions

- Screen content
- Employ a liberal *Take Down* policy and process
Foreign Laws

- Consult foreign or expert lawyers
Archiving and Preservation

Modeling as web harvest objects

METS Profile for Web Harvests
1. Descriptive metadata for seed
2. Seed list
3. Metadata for crawl
4. Crawl configuration

Resource Container 1
Resource Container 2
Resource Container n

IS_PART_OF

IS_LOG_FOR

Crawl log
Archiving and Preservation – Identifying the Challenges

- Requires more pre-processing than other types of content
  - Viruses more likely
  - Misidentified formats
  - Invalid formats
- Proliferation of formats
- What are we preserving?
  - Incomplete acquisition – parts of web pages
- High maintenance delivery
  - Hyperlinked resources
  - Multiple renderers
Still Sticky Challenges
Still Sticky Challenges

• Sustainability
  • Ballooning database and storage
    • Duplication
    • Junk
    • Metadata efficiencies
  • Keeping up with evolving software, standards and laws

• Immature software
  • Capture and Display Problems
  • Foreign language support
A Sneak Peek at WAX Pre-release

December 1, 2008

http://wax.lib.harvard.edu
Revisiting Project Goals

- Technical feasibility
- Legal terrain
- Experience
- Resource requirements
- Sustainability

Partial
Partial
Yes!
In process?

1BSUJBM:FT*OQSPDFT
Image Credits

• Bees on honeycomb:

• Beeswax:
  http://beeyondthehive.com/store/media/thumbnails/BeesWaxBar

• Man with bees:
  http://www.johnbgrimes.com/blog/bees.jpg

• Bear w/jar:
  http://i.dailymail.co.uk/i/pix/2008/07/31/article-0-021F28380000578-978_468x286.jpg

• Bee on flower:

• Cartoon bee:
Questions?