HARVARD UNIVERSITY LIBRARY

Web Archive Collection Service

A Busy Hive Creates Better WAX

Archiving the Web from many Perspectives

DLF Fall Forum 2008

Andrea Goethals, Wendy Gogel – Harvard University Library

Harvard University Library

Web Archive Collection Service

Today's Agenda

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

Harvard University Library

Web Archive Collection Service

An Introduction to WAX

- A busy hive…

Harvard University Library
Web Archive Collection Service

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

Harvard University Library
Web Archive Collection Service

Who is involved?

- Collection managers
  (librarians, archivists, faculty)
- Legal counsel
  (University and external lawyers)
- Technologists
  (architects, programmer, managers,
  graphic designers, preservationists)

Harvard University Library
Web Archive Collection Service

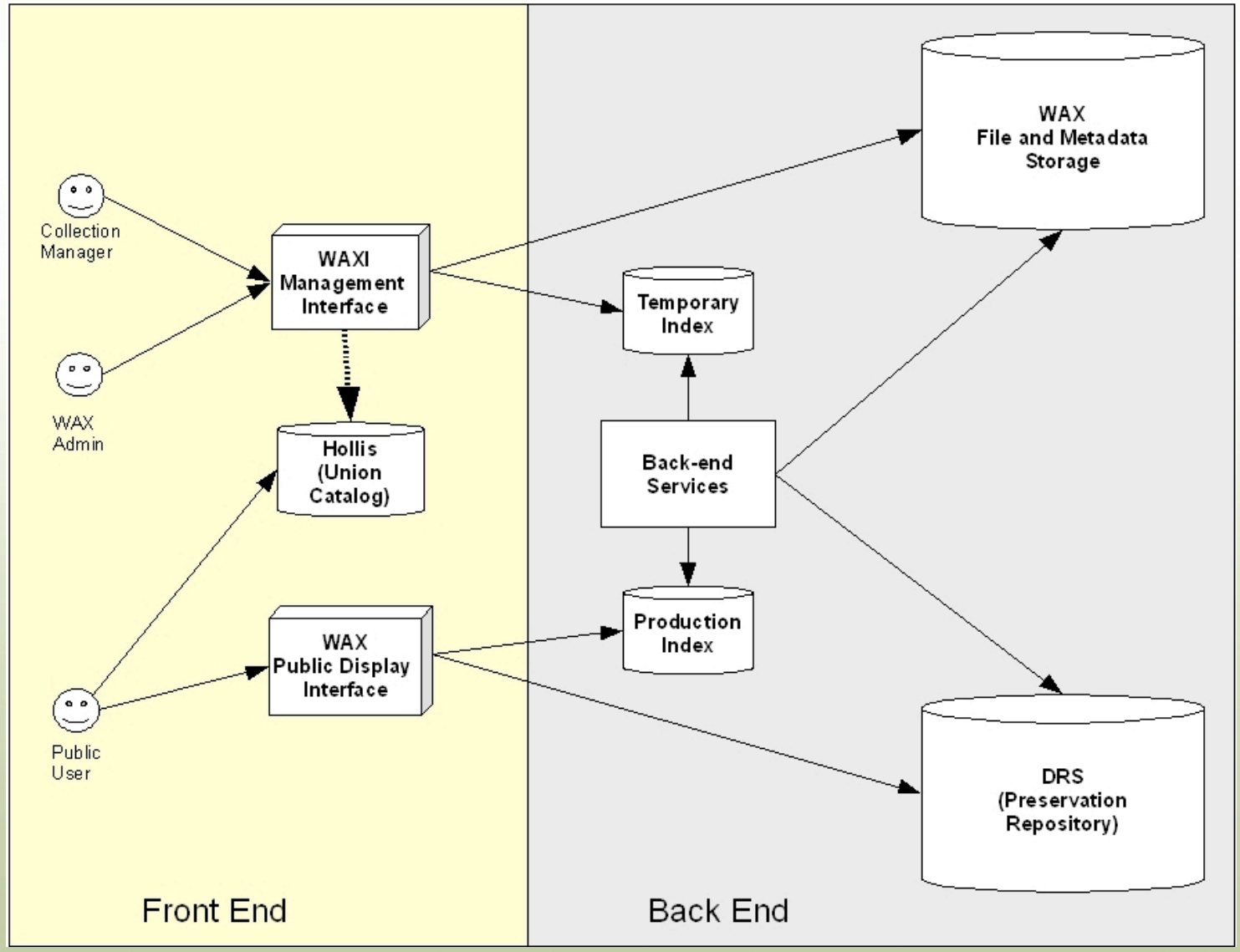
What is involved?

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

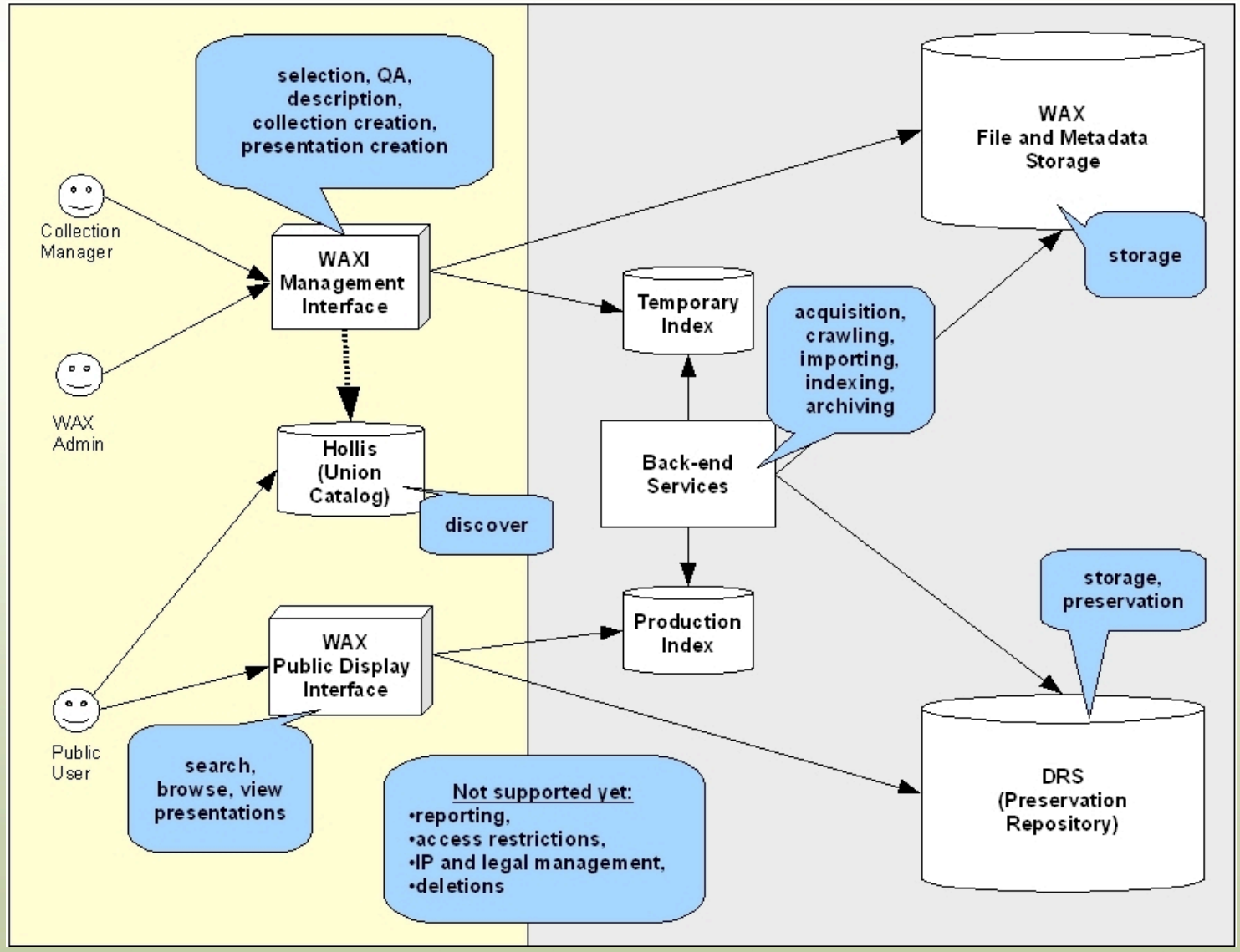
WAX and WAXi Components

3rd Party Software:

- **Heritrix**
- **Hcc**
- **JBoss**
- **NutchWAX**
- **Oracle**
- **Quartz**
- **Struts**
- **Tomcat**
- **Wayback**



WAX and WAXi Activities



Harvard University Library

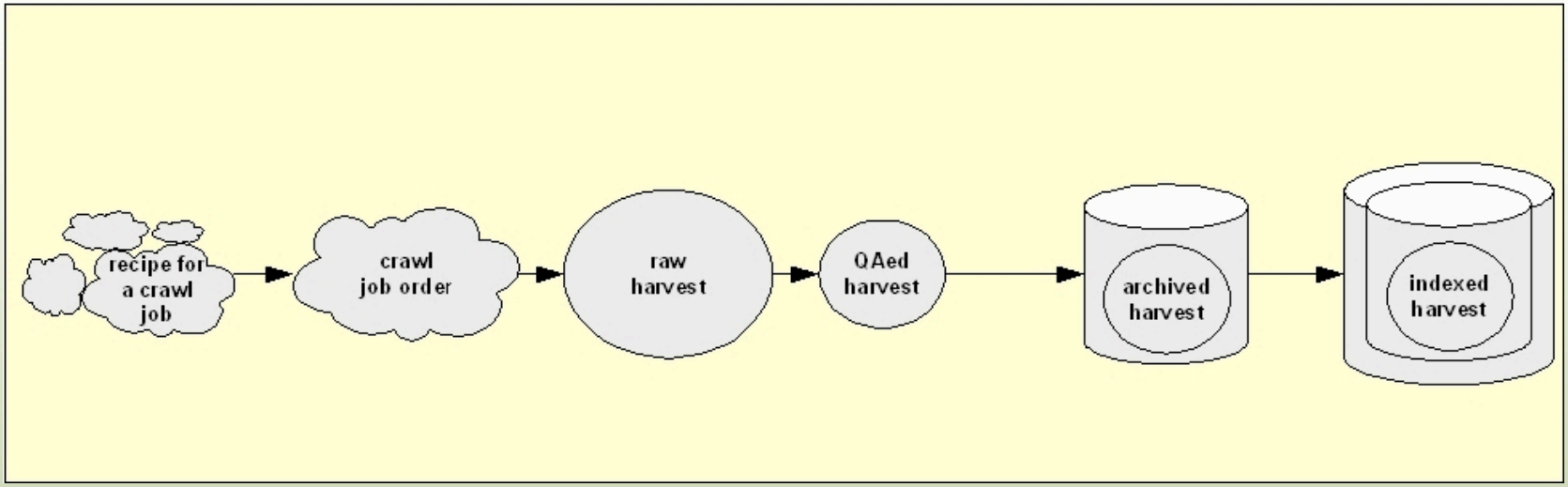
Web Archive Collection Service

Dissolving Sticky Challenges

…creates better WAX

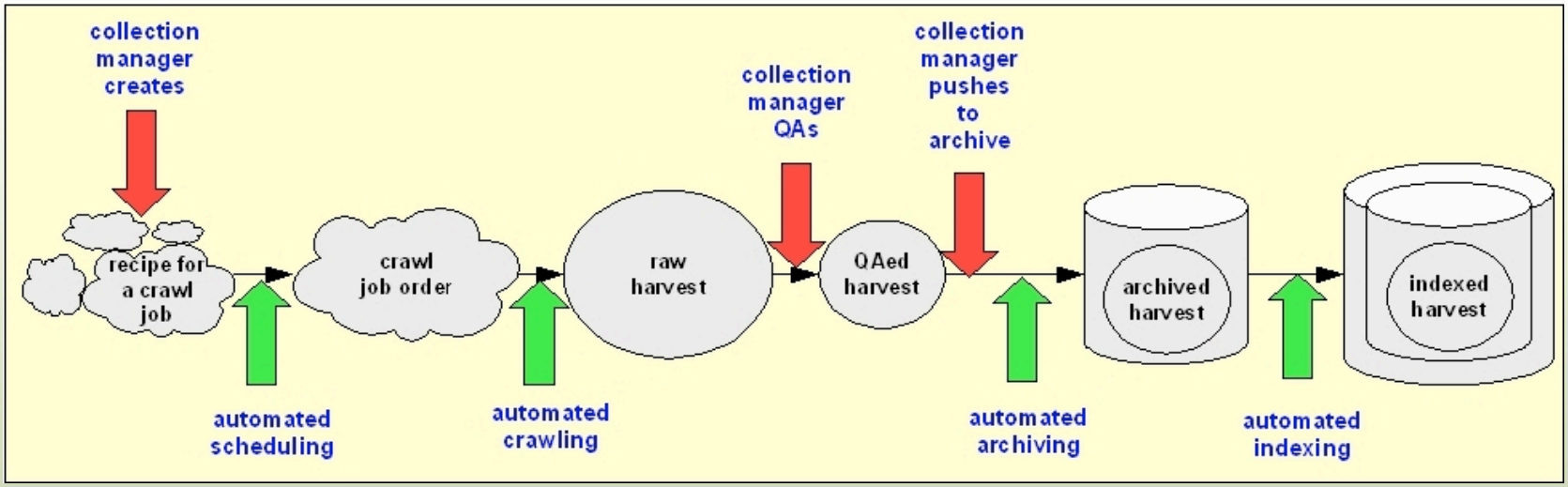
Operational Efficiency and Automation

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



… which can we automate?

Operational Efficiency and Automation



Operational Efficiency and Automation

Harvard University Library
Web Archive Collection Service

- Automating the back end services
  - Crawl Scheduling
  - Crawling
  - Harvest Archiving
  - Harvest Indexing

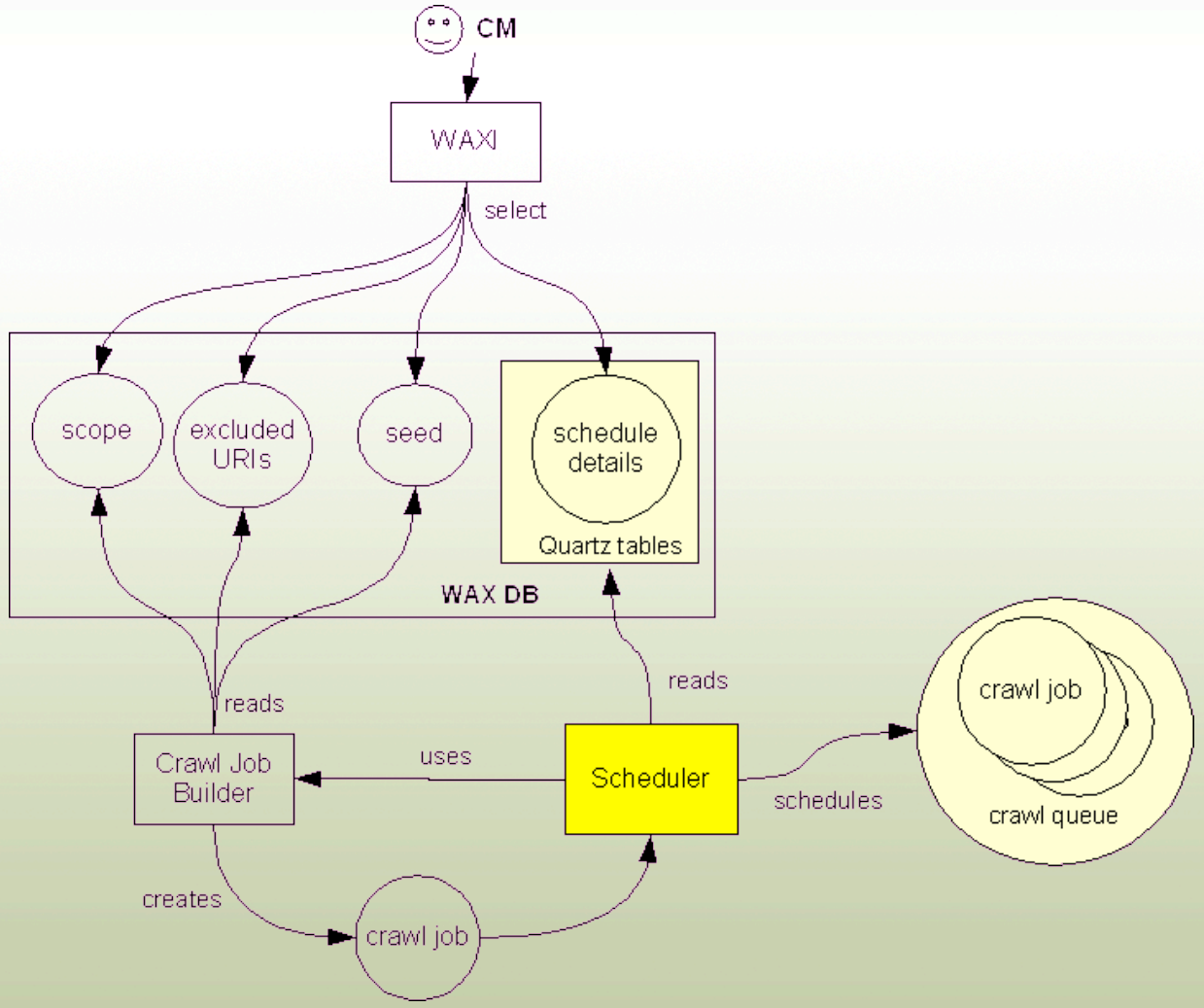
Recipe for an Automated Back End Service

Harvard University Library
Web Archive Collection Service

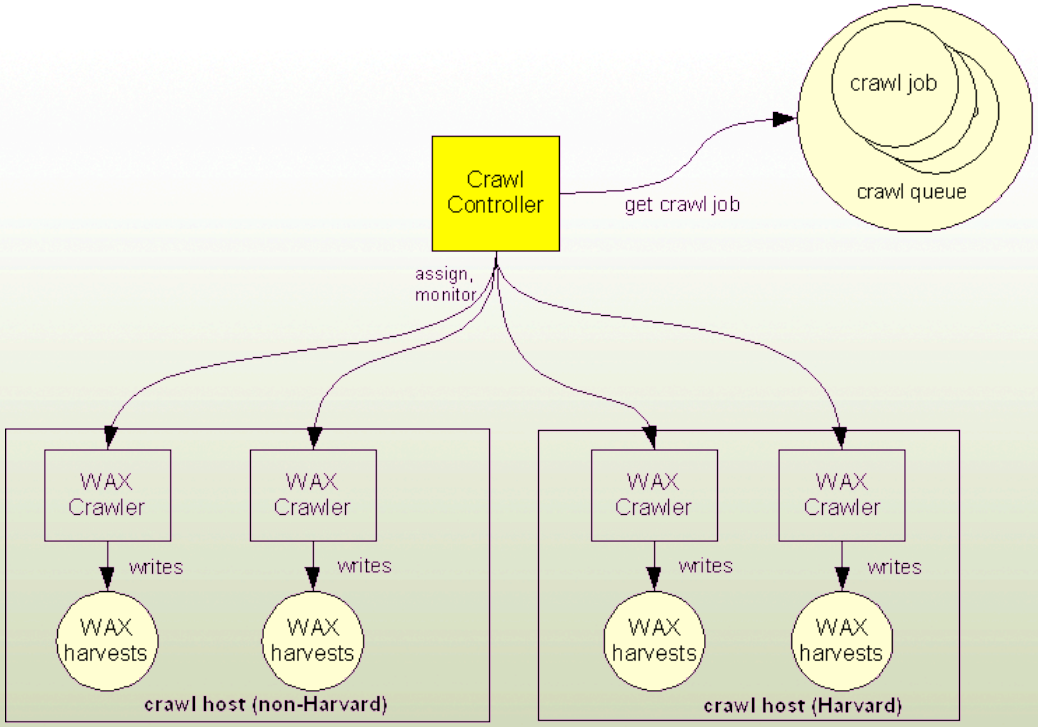
Service:

- 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

Harvard University Library
Web Archive Collection Service

- 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

accounts | collections | seeds | URIs | harvests | rights | WAX | help

Andrea Goethals
Role: WAX Administrator
log out (Can't log out? Try this.)
Logged in since: 01:45 PM (11/06/08)

Harvests

Select | Manage | Reports

show controls
yes / no

choose scope
http://www.thesultanas.com/ (01/25/2007 17:27:33)
refresh display

choose display
show QA screen
tree display:
show excluded URIs
show deleted URIs
preview:
include live sites

Harvest Info:

Seed URI: http://www.thesultanas.com/
Harvest ID: 3
Crawl start date: Thu Jan 25 16:54:36 EST 2007
Crawl end date: Thu Jan 25 17:27:33 EST 2007
Crawl duration: 32m57s675ms
Crawl scope: Unknown
Crawler host: chaz.hul.harvard.edu
Crawler software: heritrix/1.8.0
Current status:
QA status: Needs QA
Mark QA as done | Delete harvest
Archive status: Not ready to archive

save | cancel

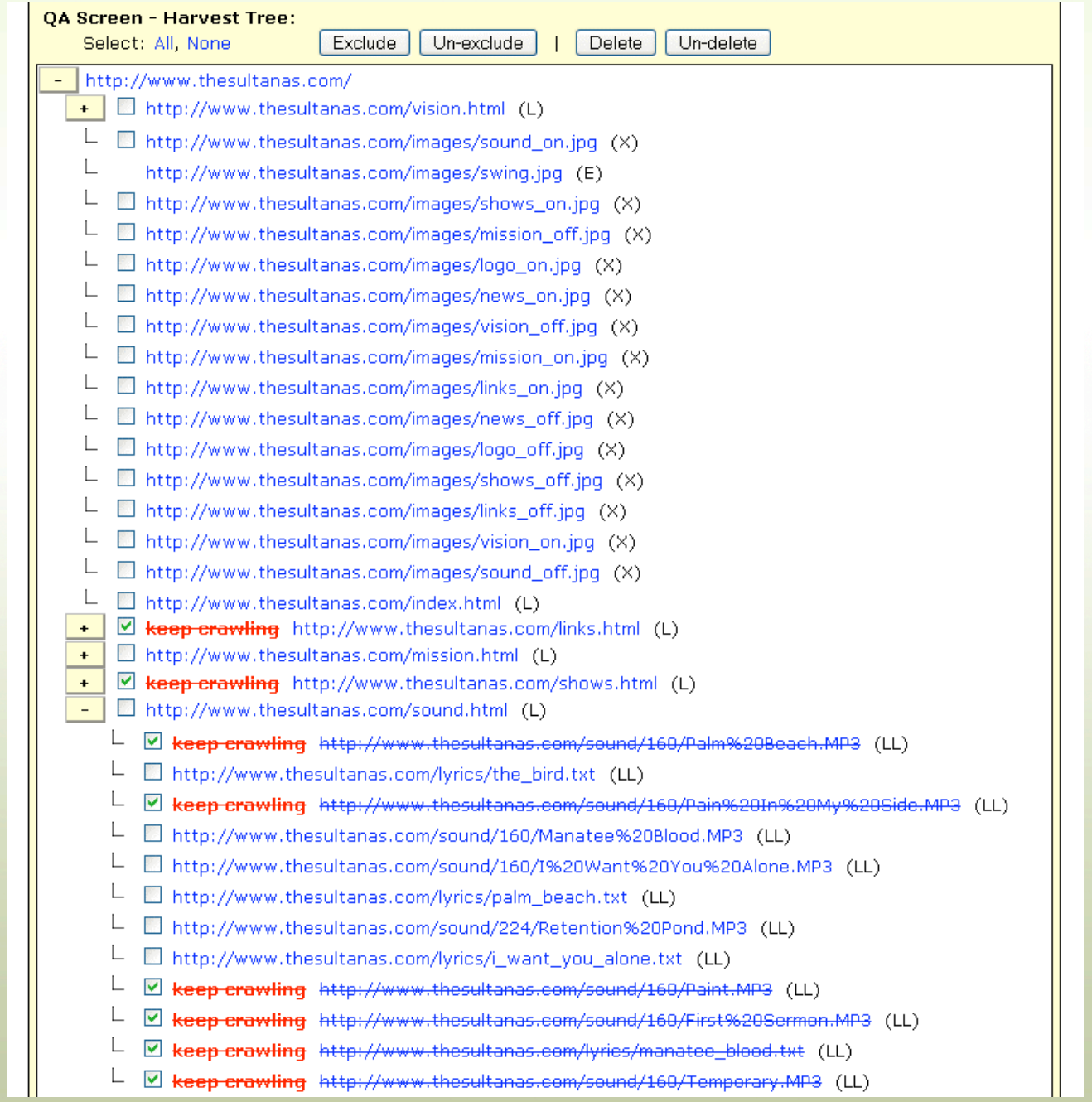
Notes:

QA Screen - Harvest Tree:

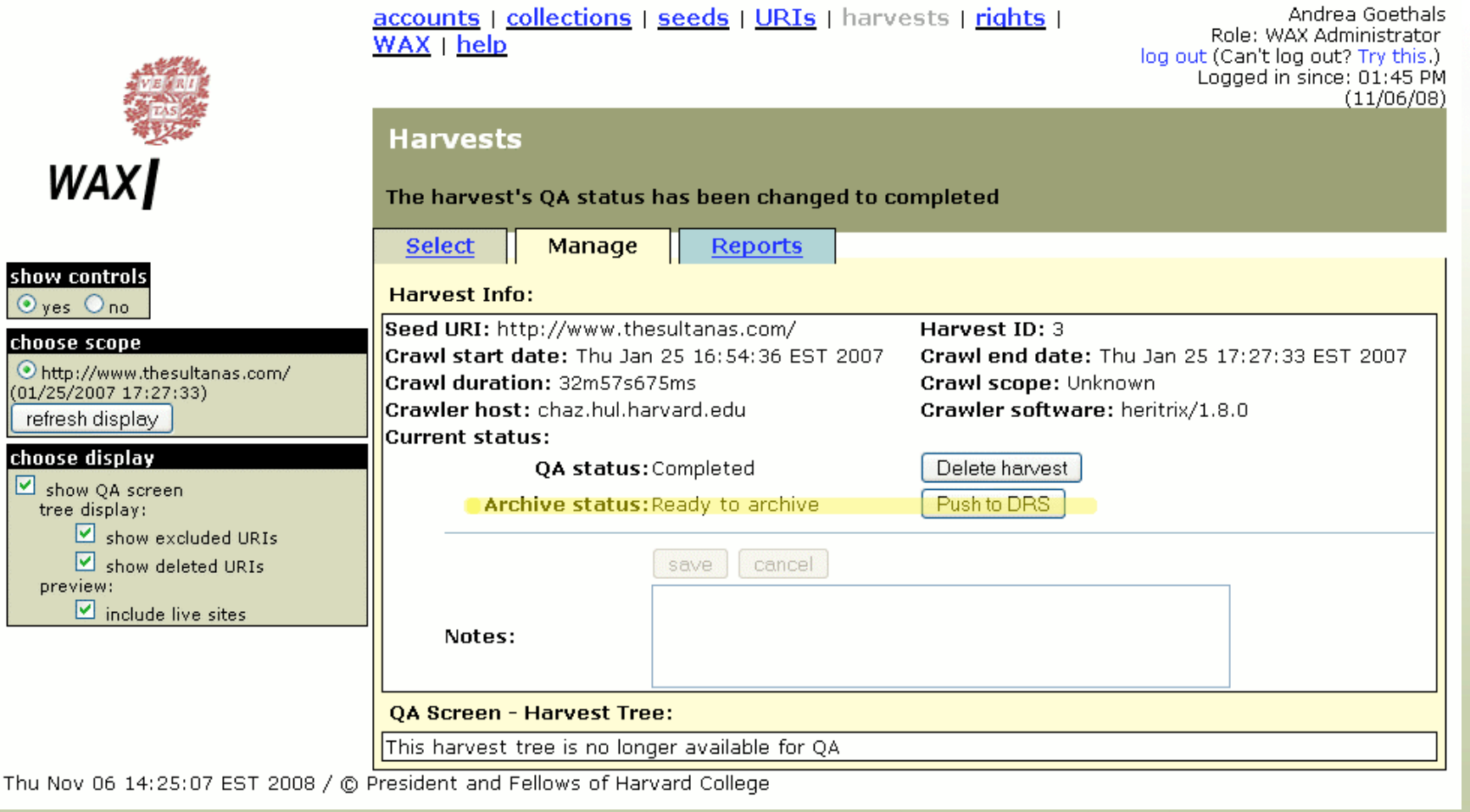
Select: All, None
Exclude | Un-exclude | Delete | Un-delete

- http://www.thesultanas.com/
  - + http://www.thesultanas.com/vision.html (L)
  - http://www.thesultanas.com/images/sound_on.jpg (X)
  - http://www.thesultanas.com/images/swing.jpg (E)
  - http://www.thesultanas.com/images/shows_on.jpg (X)
  - http://www.thesultanas.com/images/mission_off.jpg (X)
  - http://www.thesultanas.com/images/logo_on.jpg (X)
  - http://www.thesultanas.com/images/news_on.jpg (X)
  - http://www.thesultanas.com/images/vision_off.jpg (X)
  - http://www.thesultanas.com/images/mission_on.jpg (X)

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



Harvard University Library
Web Archive Collection Service

The Ultimate QA and Archive Shortcut

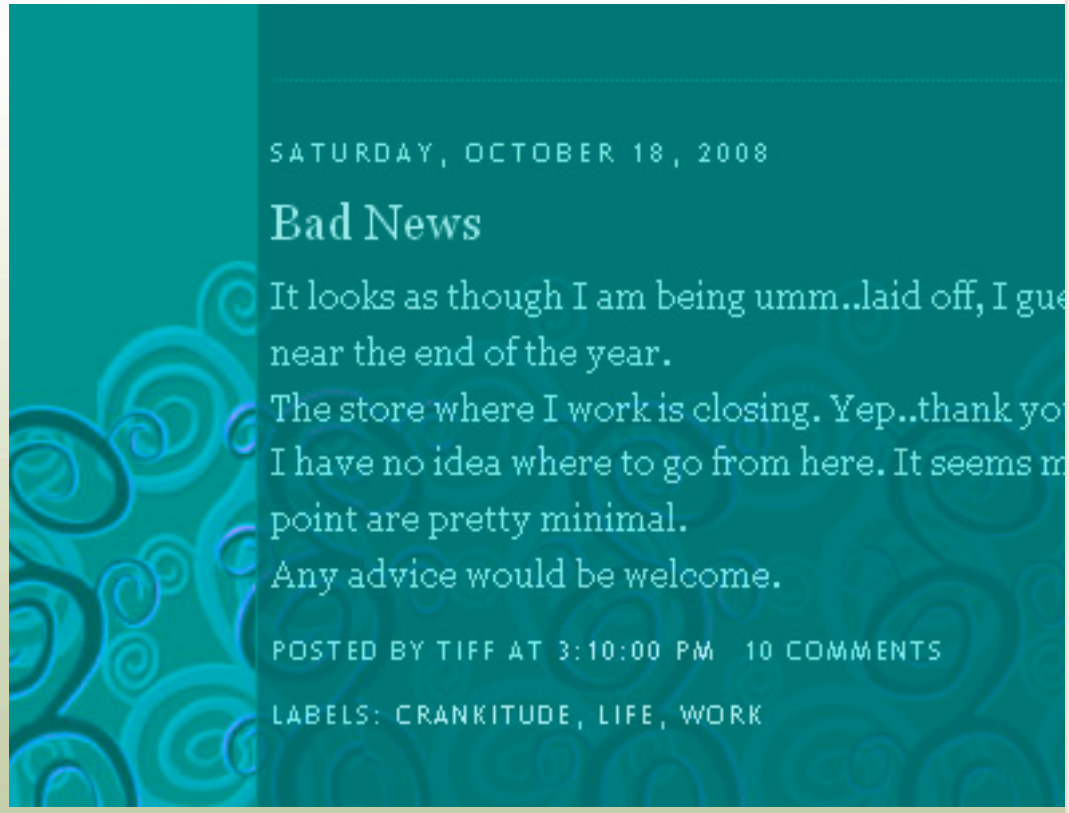
- Can eliminate QA and/or archive push step on a per seed basis
- Crawl → Preservation Repository

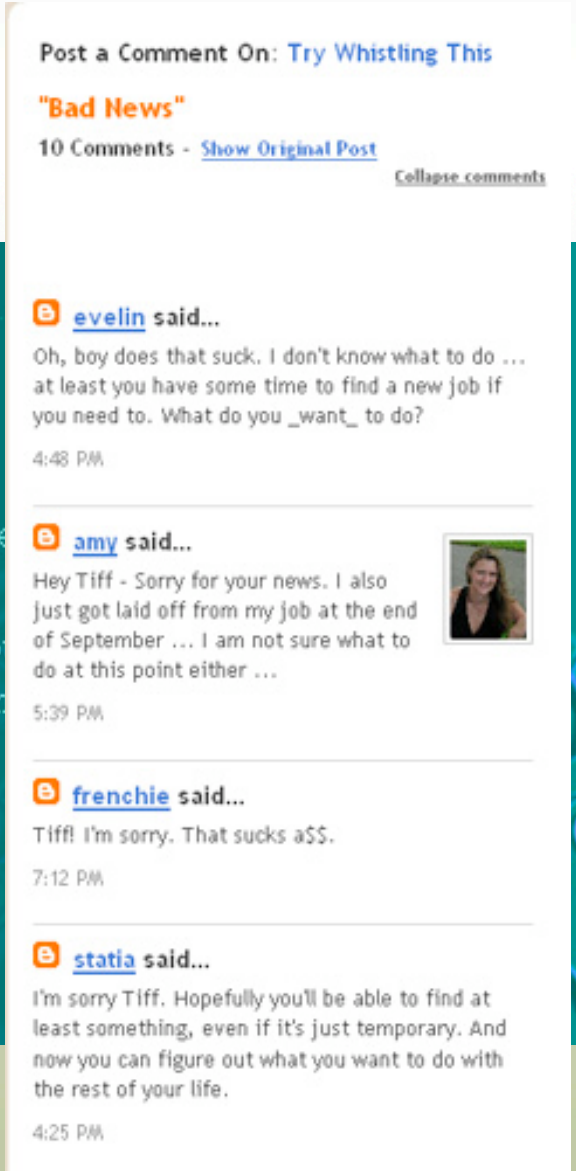
Legal Risks and Mitigations

Harvard University Library
Web Archive Collection Service

Copyright Infringement

- Permissions
- Fair Use/Fair Play





Harvard University Library
Web Archive Collection Service

Fair use/Fair Play

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

Harvard University Library
Web Archive Collection Service

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

Harvard University Library
Web Archive Collection Service

State Tort Liability

- Civil Damages
  - Invasion of privacy
  - Sensitive personal data
  - Commercial content
  - Defamatory content

Harvard University Library
Web Archive Collection Service

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

Harvard University Library
Web Archive Collection Service

Statutory Content Restrictions

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

Harvard University Library
Web Archive Collection Service

Statutory Content Restrictions

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

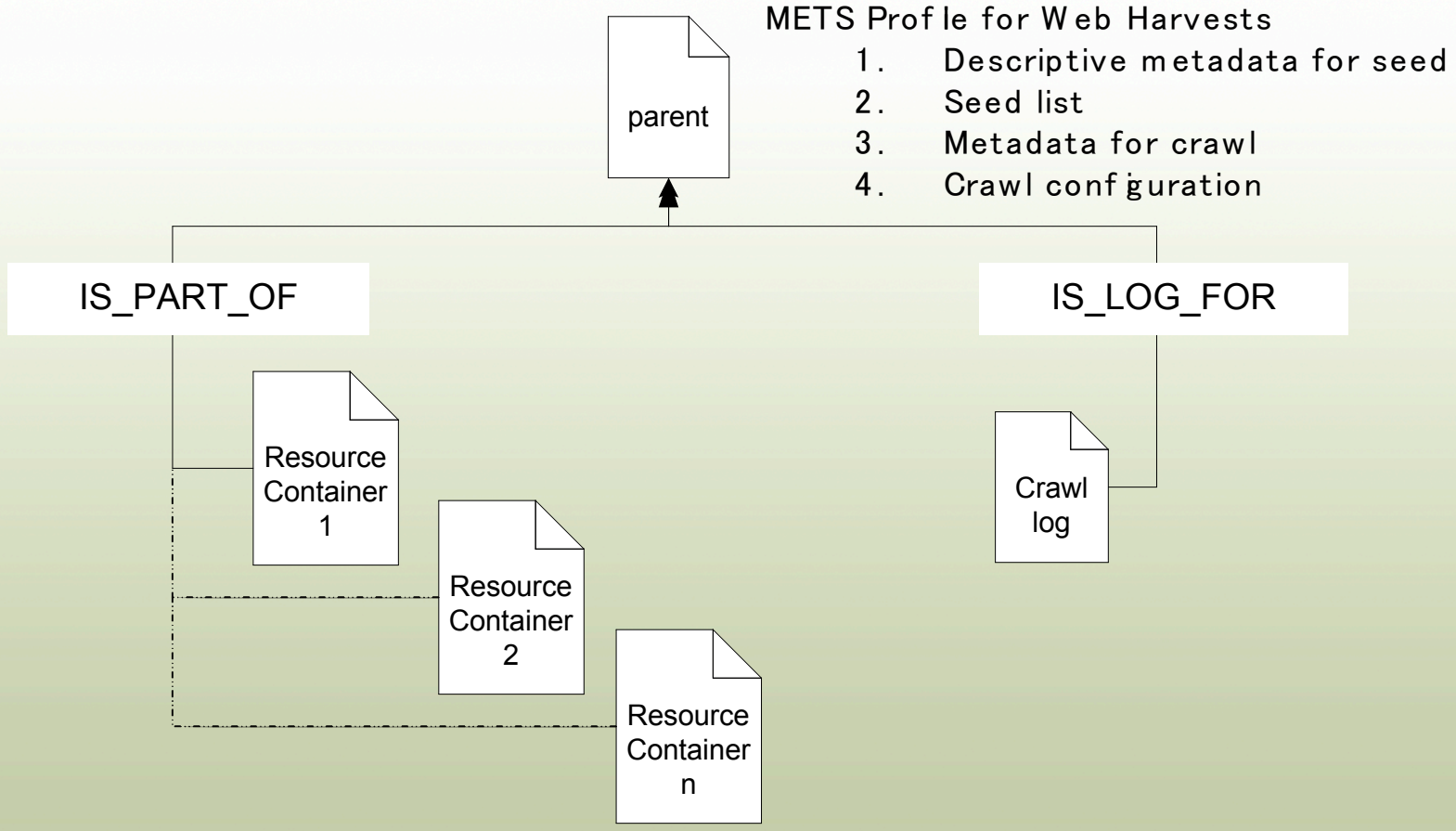
Harvard University Library
Web Archive Collection Service

Foreign Laws

- Consult foreign or expert lawyers

Archiving and Preservation

Modeling as web harvest objects



Archiving and Preservation – Identifying the Challenges

Harvard University Library
Web Archive Collection Service

- Requires more pre-processing then 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

Harvard University Library
Web Archive Collection Service

Harvard University Library
Web Archive Collection Service

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

Harvard University Library

Web Archive Collection Service

December 1, 2008

http://wax.lib.harvard.edu

Revisiting Project Goals

Harvard University Library
Web Archive Collection Service

Partial	• Technical feasibility
Partial	• Legal terrain
Yes!	• Experience
In proces s ?	• Resource requirements • Sustainability

Image Credits

- Bees on honeycomb:
  http://morningnoonandnight.files.wordpress.com/2007/09/beehive.jpg
- 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-0-021F283800000578-978_468x286.jpg
- Bee on flower:
  http://www.mitzenmacher.net/blog/wp-content/uploads/2007/03/bee.jpg
- Cartoon bee:
  http://www.how-to-draw-funny-cartoons.com/cartoon-bee.html

Questions?

Harvard University Library

Web Archive Collection Service