



# 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)



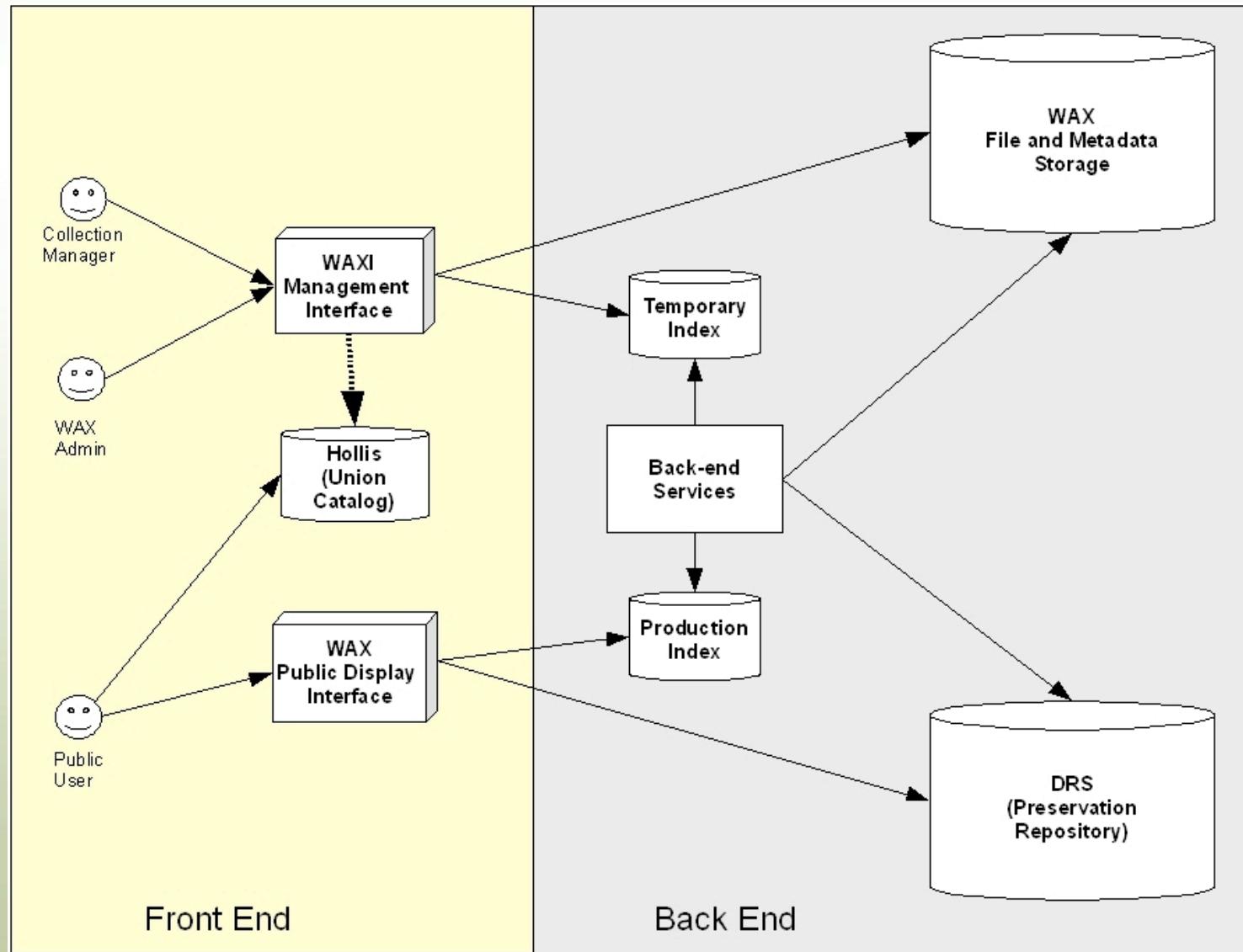
# What is involved?

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

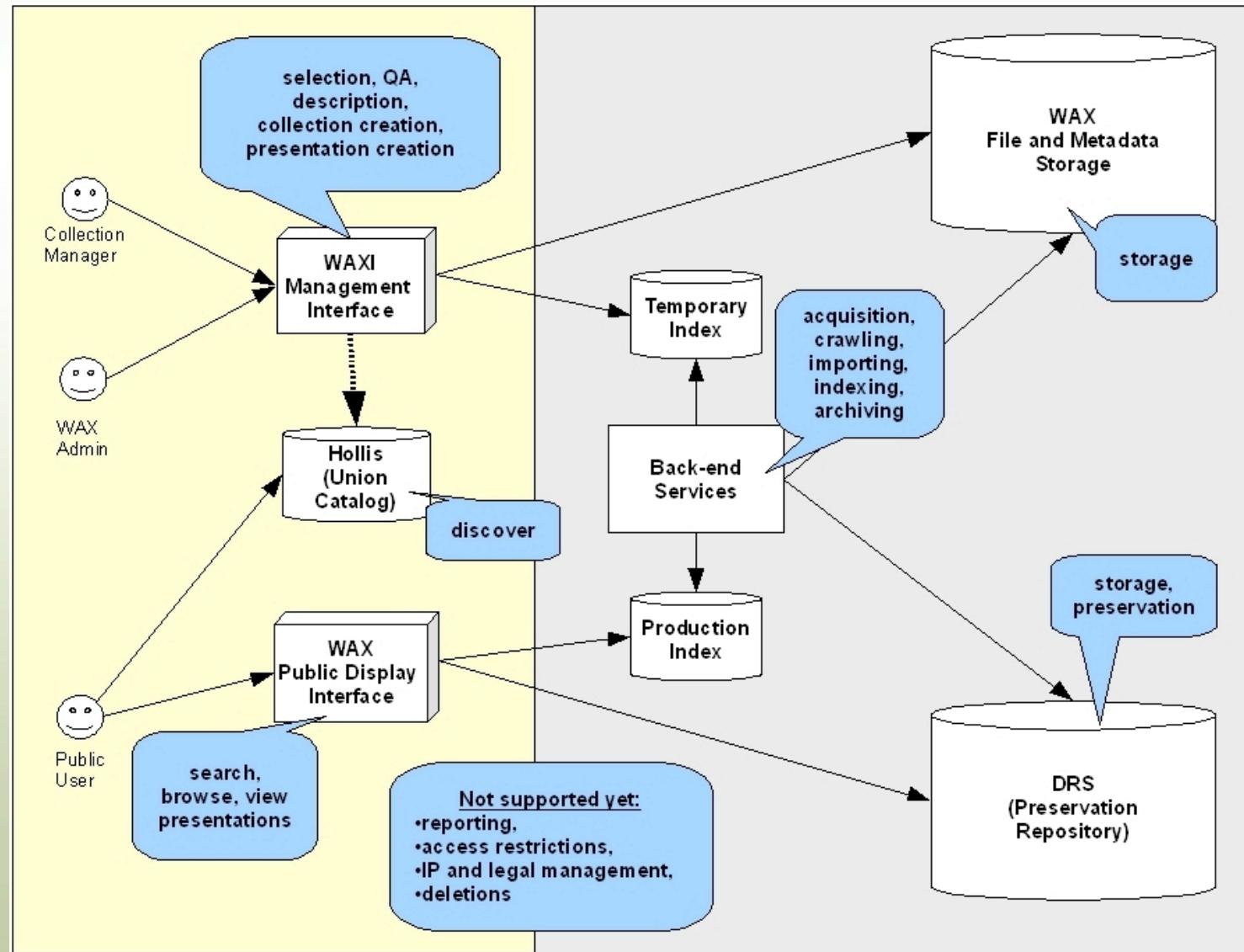
# WAX and WAXi Components

**3rd Party Software:**

- Heritrix**
- Hcc**
- JBoss**
- NutchWA X**
- Oracle**
- Quartz**
- Struts**
- Tomcat**
- Wayback**



# WAX and WAXi Activities





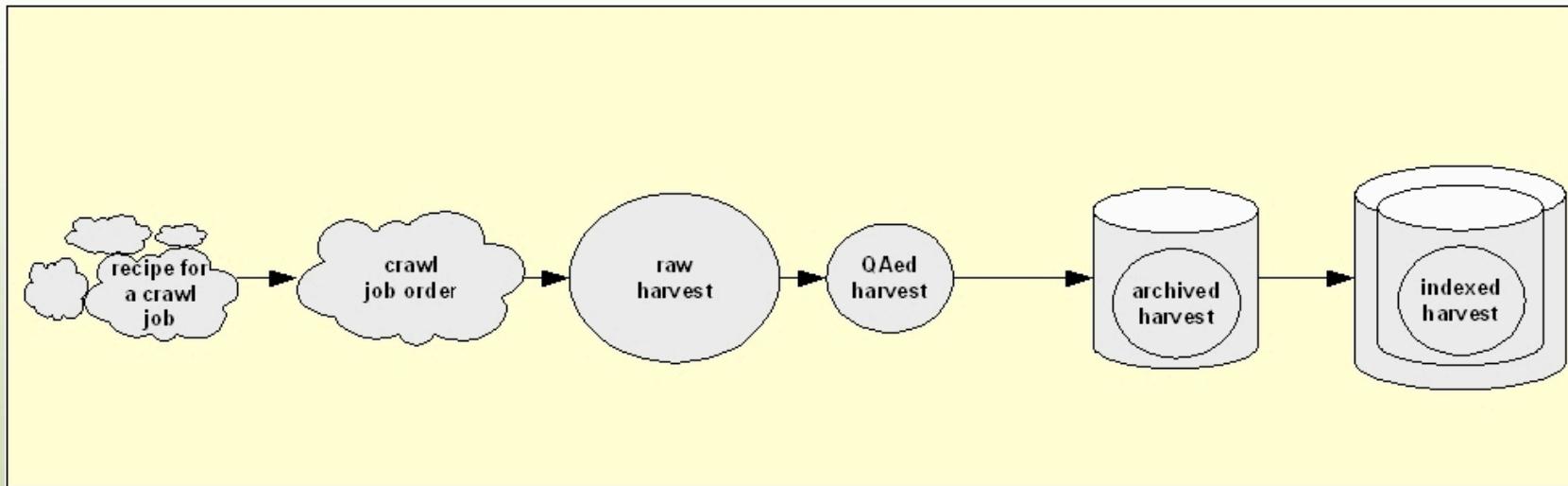
# *Dissolving Sticky Challenges*

...creates better WAX



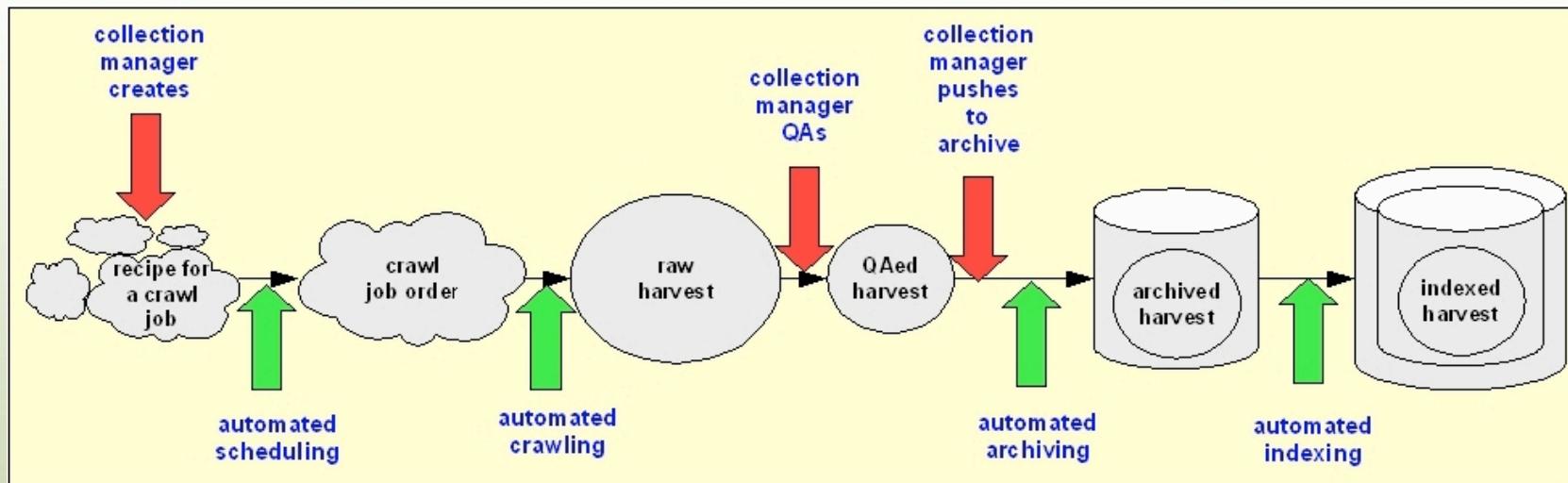
# Operational Efficiency and Automation

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

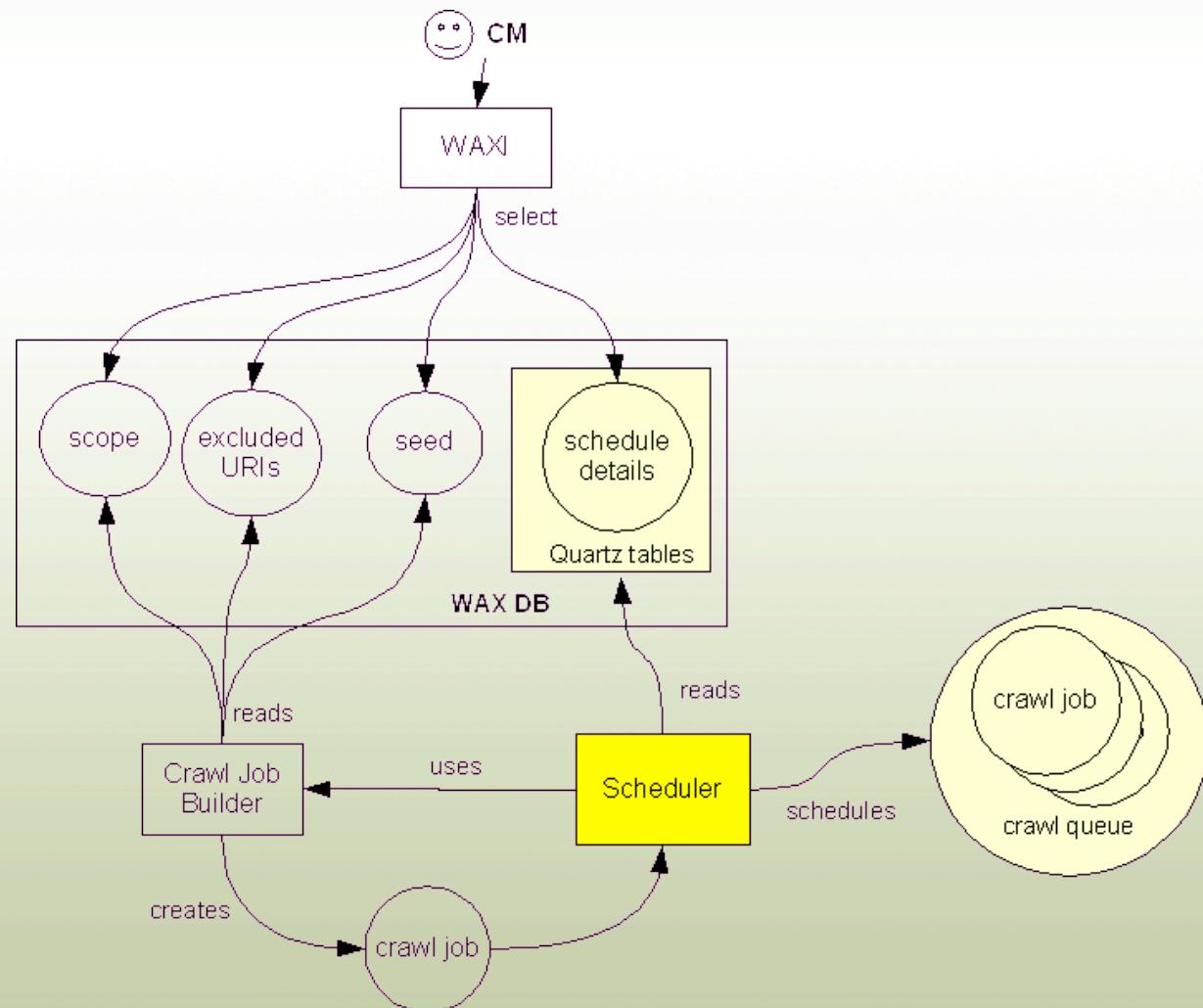


# Recipe for an Automated Back End Service

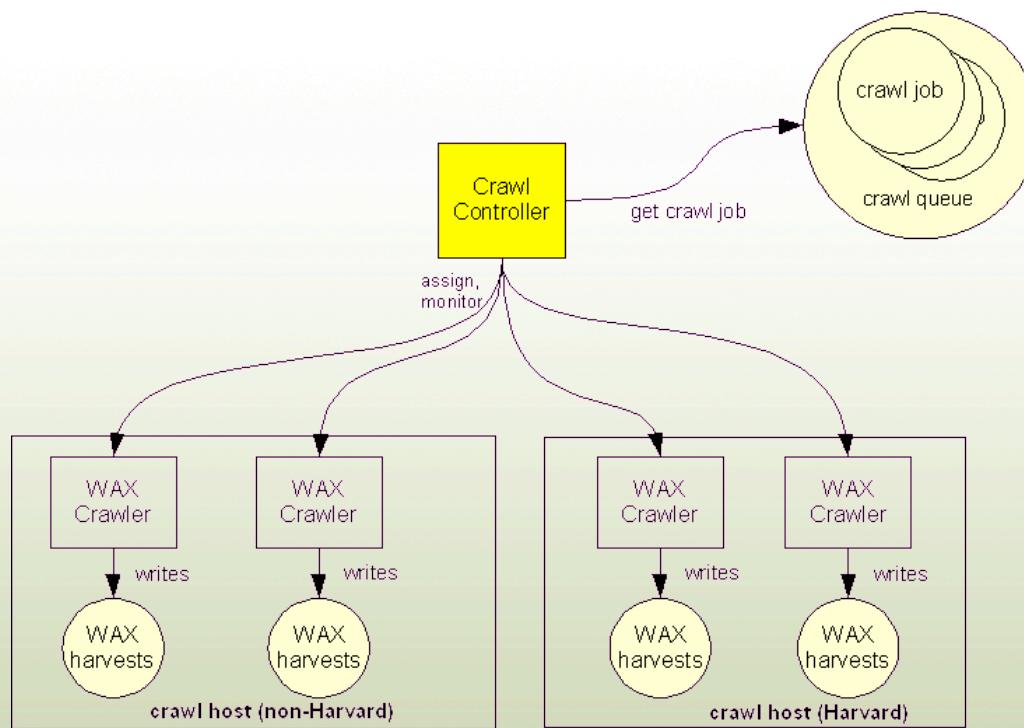
## Service:

- 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



# WAX

[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](#)

### Harvest Info:

**Seed URI:** <http://www.thesultanash.com/>  
**Crawl start date:** Thu Jan 25 16:54:36 EST 2007  
**Crawl duration:** 32m57s675ms  
**Crawler host:** chaz.hul.harvard.edu

**Harvest ID:** 3  
**Crawl end date:** Thu Jan 25 17:27:33 EST 2007  
**Crawl scope:** Unknown  
**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.thesultanash.com/>
- + <http://www.thesultanash.com/vision.html> (L)
  - [http://www.thesultanash.com/images/sound\\_on.jpg](http://www.thesultanash.com/images/sound_on.jpg) (X)
  - <http://www.thesultanash.com/images/swing.jpg> (E)
  - [http://www.thesultanash.com/images/shows\\_on.jpg](http://www.thesultanash.com/images/shows_on.jpg) (X)
  - [http://www.thesultanash.com/images/mission\\_off.jpg](http://www.thesultanash.com/images/mission_off.jpg) (X)
  - [http://www.thesultanash.com/images/logo\\_on.jpg](http://www.thesultanash.com/images/logo_on.jpg) (X)
  - [http://www.thesultanash.com/images/news\\_on.jpg](http://www.thesultanash.com/images/news_on.jpg) (X)
  - [http://www.thesultanash.com/images/vision\\_off.jpg](http://www.thesultanash.com/images/vision_off.jpg) (X)
  - [http://www.thesultanash.com/images/mission\\_on.jpg](http://www.thesultanash.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**

QA Screen - Harvest Tree:

Select: All, None    [Exclude](#) [Un-exclude](#) | [Delete](#) [Un-delete](#)

- http://www.thesultanas.com/
  - + http://www.thesultanas.com/vision.html (L)
  - L http://www.thesultanas.com/images/sound\_on.jpg (X)
  - L http://www.thesultanas.com/images/swing.jpg (E)
  - L http://www.thesultanas.com/images/shows\_on.jpg (X)
  - L http://www.thesultanas.com/images/mission\_off.jpg (X)
  - L http://www.thesultanas.com/images/logo\_on.jpg (X)
  - L http://www.thesultanas.com/images/news\_on.jpg (X)
  - L http://www.thesultanas.com/images/vision\_off.jpg (X)
  - L http://www.thesultanas.com/images/mission\_on.jpg (X)
  - L http://www.thesultanas.com/images/links\_on.jpg (X)
  - L http://www.thesultanas.com/images/news\_off.jpg (X)
  - L http://www.thesultanas.com/images/logo\_off.jpg (X)
  - L http://www.thesultanas.com/images/shows\_off.jpg (X)
  - L http://www.thesultanas.com/images/links\_off.jpg (X)
  - L http://www.thesultanas.com/images/vision\_on.jpg (X)
  - L http://www.thesultanas.com/images/sound\_off.jpg (X)
  - L http://www.thesultanas.com/index.html (L)
  - + keep-crawling http://www.thesultanas.com/links.html (L)
  - + http://www.thesultanas.com/mission.html (L)
  - + keep-crawling http://www.thesultanas.com/shows.html (L)
  - http://www.thesultanas.com/sound.html (L)
    - L keep-crawling http://www.thesultanas.com/sound/160/Palm%20Beach.MP3 (LL)
    - L http://www.thesultanas.com/lyrics/the\_bird.txt (LL)
    - L keep-crawling http://www.thesultanas.com/sound/160/Pain%20In%20My%20Side.MP3 (LL)
    - L http://www.thesultanas.com/sound/160/Manatee%20Blood.MP3 (LL)
    - L http://www.thesultanas.com/sound/160/I%20Want%20You%20Alone.MP3 (LL)
    - L http://www.thesultanas.com/lyrics/palm\_beach.txt (LL)
    - L http://www.thesultanas.com/sound/224/Retention%20Pond.MP3 (LL)
    - L http://www.thesultanas.com/lyrics/i\_want\_you\_alone.txt (LL)
    - L keep-crawling http://www.thesultanas.com/sound/160/Paint.MP3 (LL)
    - L keep-crawling http://www.thesultanas.com/sound/160/First%20Sermon.MP3 (LL)
    - L keep-crawling http://www.thesultanas.com/lyrics/manatee\_blood.txt (LL)
    - L keep-crawling http://www.thesultanas.com/sound/160/Temporary.MP3 (LL)

# One button push to archive

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**

The harvest's QA status has been changed to completed

[Select](#)   [Manage](#)   [Reports](#)

**Harvest Info:**

<b>Seed URI:</b> http://www.thesultanas.com/	<b>Harvest ID:</b> 3
<b>Crawl start date:</b> Thu Jan 25 16:54:36 EST 2007	<b>Crawl end date:</b> Thu Jan 25 17:27:33 EST 2007
<b>Crawl duration:</b> 32m57s675ms	<b>Crawl scope:</b> Unknown
<b>Crawler host:</b> chaz.hul.harvard.edu	<b>Crawler software:</b> heritrix/1.8.0

**Current status:**

QA status: Completed      Delete harvest

Archive status: Ready to archive      Push to DRS

save   cancel

**Notes:**

**QA Screen - Harvest Tree:**  
This harvest tree is no longer available for QA

Thu Nov 06 14:25:07 EST 2008 / © President and Fellows of Harvard College



# The Ultimate QA and Archive Shortcut

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

URN authority path:	NOUGEST
Cataloging method:	wax
Requires QA:	yes
Requires archive push:	yes
Crawler preference:	non-harvard
Access flag:	P

save  
cancel



# Legal Risks and Mitigations



# Copyright Infringement

- Permissions
- Fair Use/Fair Play

SATURDAY, 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 point are pretty minimal.

Any advice would be welcome.

POSTED BY TIFF AT 3:10:00 PM 10 COMMENTS

LABELS: CRANKITUDE, LIFE, WORK

Post a Comment On: [Try Whistling This](#)

"Bad News"

10 Comments - [Show Original Post](#)

[Collapse comments](#)

 [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

 [stacia](#) 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

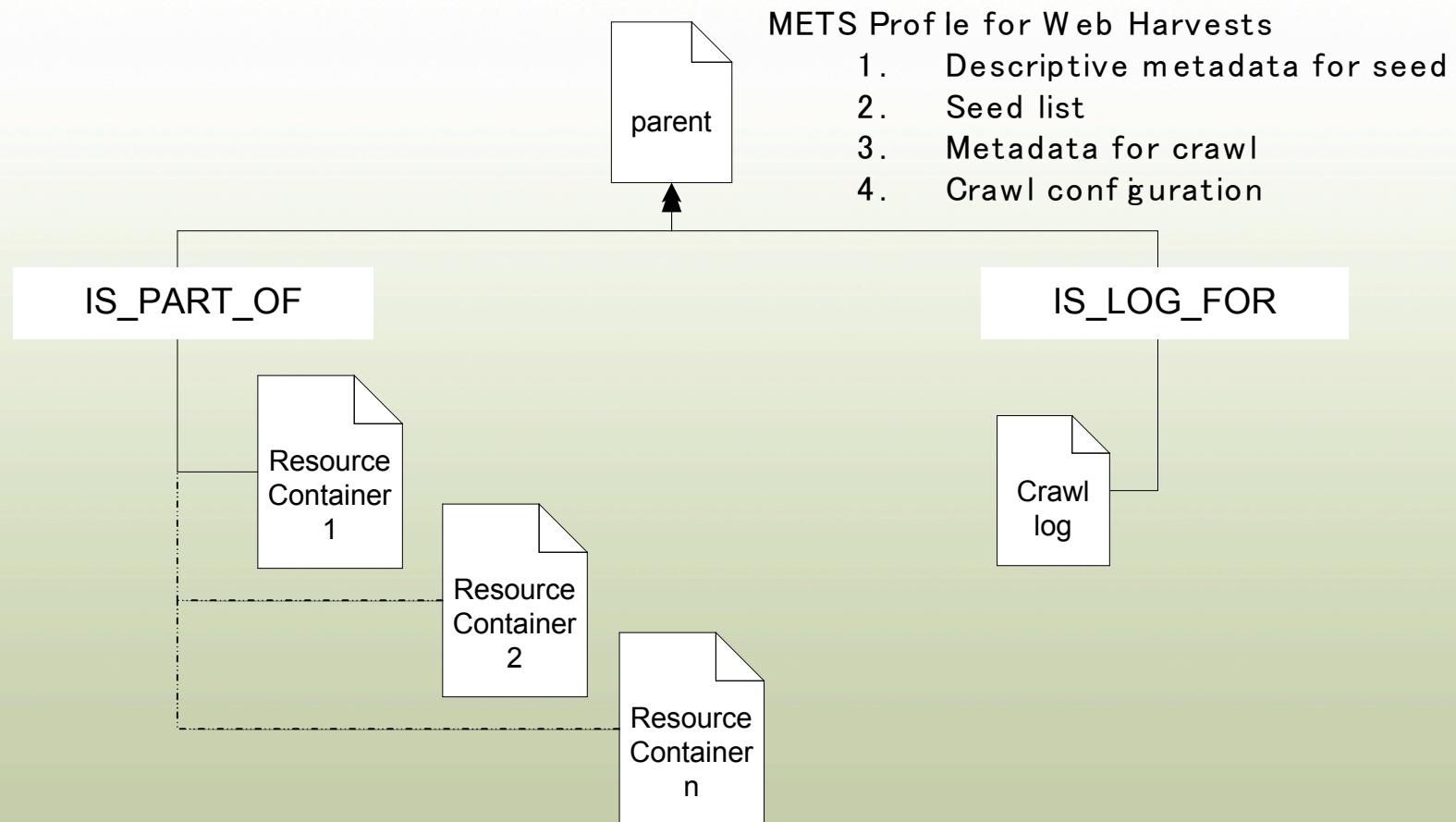


# Foreign Laws

- Consult foreign or expert lawyers

# Archiving and Preservation

## Modeling as web harvest objects





# 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

- Partial** • Technical feasibility
- Partial** • Legal terrain
- Yes!** • Experience
- In** • Resource requirements
- process** • Sustainability
- s**
- ?**

# 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](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?

