Deep Web Content and Internet Discovery: Exposing Harvard University Library's Digital Resources to Search Engines
OK, we built it...


...did they come?
"At Harvard, we collect library materials comprehensively and globally in traditional as well as digital forms. Our long and remarkable history of worldwide collecting results from generations of sustained support from alumni/ae, friends, corporations, and foundations. But in the digital age, we are taking new and additional steps that can make those global collections accessible to the world beyond Harvard Yard."

--- Sidney Verba, Carl H. Pforzheimer University Professor and former Director of the University Library, in "Harvard Libraries 2004."
Library Digital Initiative

Mission: to create an infrastructure to support the "collecting" of digital resources at Harvard

Initiated in 1998

Comprises both “back-end” systems and “front-end” applications

Internally funded
Client machines were slow
Netscape “owned” 90% of the market
IE v3, MS’s first useable browser
What’s a baud rate?
Back-end servers: expensive and slow
The web environment in 1998 (cont’d)

How one discovered new web sites:

– Usenet: comp.infosystems.www.announce
– Links from other web sites
– Links from on-line news sources
– Articles in journals
– "Yet Another Hierarchical Officious Oracle"
LDI: Some “front end” applications

• **OASIS**: Online Archival Search Information System
  Union catalog of archival and manuscript finding aids

• **VIA**: Visual Information Access
  Union catalog of visual materials at Harvard

• **PDS**: Page Delivery Service
  Delivers scanned page images from multi-paged documents
LDI: Some “front end” applications

- TED: TEemplated Databases
  Online home for new specialized collections catalogs

- VC: Virtual Collections
  Provides a unified special collection view across systems

- Open Collections
  Based on VC, integrates the unified view with other content
LDI: Some “front end” applications

Altogether: 400,000+

– page-turned objects,
– high-quality images,
– sound clips, and
– other digital objects
“Portal” orientation

Search VIA

The Visual Information Access (VIA) system is a union catalog of visual resources at Harvard, focusing on artistic and cultural materials.

VIA includes catalog records for objects or images owned, held or licensed by Harvard. Access to the catalog is open to the general public; all catalog records and thumbnail images are available to everyone. Access to higher resolution images is usually available to the Harvard community, is always determined by an individual repository, and is often dependent on copyright.

Access to original object or image is determined by the individual repositories. Restrictions on access may be noted in the VIA record.

For more detailed information, see About VIA.

Search

Search for: [field] in Anywhere

and [field] in Anywhere

and [field] in Anywhere

Limit search to records that have digital images

Limit search to records that have originals at Harvard

Limit search to records with dates from YYYY to YYYY

Limit repository to: [field]

Search

Hints:
Use * as a wildcard. Examples: cat*, *yperable and *itics*

Display/Sort Preferences:

- Image grid size: Small - 3 rows x 5 columns
- Sort by: None
- Result sets greater than 2000 will not be sorted
The Problem
The Problem
The Problem
The Problem
INTRODUCTION.

THE SUGAR-BEET CROP AND ITS HAND WORKERS.

The beet-sugar industry in the United States is of comparatively recent development; but its growth during the last 20 or 25 years has been so rapid that its importance both as a manufacturing and an agricultural industry is fully established. In 1896 there were but 7 factories in the country, producing 37,536 tons of beet sugar; 10 years later the number of factories had increased to 63 and the sugar tonnage to 483,612.1 In 1920 there were 98 factories with a total output of 1,090,021 tons.2

The increase in sugar-beet acreage has kept pace with the growth in the manufacture of beet sugar. In 1920, 872,376 acres of beets were harvested3, an increase of almost 700 per cent over the acreage in 1899.4

Beet-growing areas are located all the way from Ohio to California, from New England to the Pacific Coast. The middle area, of which...
The Problem

Drawbacks for crawlers:

- Session based
- Form (POST) driven
- Items frequently displayed with frames
- Unique semantic content of individual pages lost in noise of repetitive text
- The HTML coding is frequently non W3C compliant
- URLs have too many parameters in the query string
The Paradigm Shifts...
Considerations

- Time needed to re-engineer
- Server load
- Getting crawlers to crawl the pages
- Getting the pages indexed
- Attracting users
- Providing context
Re-engineering Applications

- Apache Tomcat Server
- Java/JSP/STRUTS
- Dynamically generated everything
- Database-Driven
  - Tamino (native XML database)
  - Oracle (relational database)
- Uses Harvard’s Digital Repository Service
Addressing Server Load

- Use “robot” meta-tags to control which pages get (and don’t get) crawled

  ```html
  <meta name="robots" content="noindex,follow" />
  <meta name="robots" content="noindex,nofollow" />
  <meta name="robots" content="index,follow" />
  
  • Respond to the “If-Modified-Since” request
  Return Status 304 if the page being requested hasn’t changed
  
  • Slow down crawlers where possible
Being Crawler-Friendly

- Provide a “site map”
  Static html index pages, linked to from our “home” page
- Use “robot” meta-tagging to direct the flow
- Change our URL structure for “deep” pages
  Replace parameters with a different structure, then use rewrite rules on the server side:
  
  http://oasis.lib.harvard.edu/oasis/deliver/~ajp00002
  vs.
  
  http://oasis.lib.harvard.edu/oasis/deliver/deepLink?_collection=oasis&uniqueId=ajp00002
Page-turned objects provide their own challenges – especially with OCR’d text!

We provide a non-framed version of each text page with

– Link back to Frames version
– Unique title for each text page
– Links to previous and next text page
INTRODUCTION.

THE SUGAR-BEET CROP AND ITS HAND WORKERS.

The beet-sugar industry in the United States is of comparatively recent development, but its growth during the last 20 or 25 years has been so rapid that its importance both as a manufacturing and an agricultural industry is fully established. In 1896 there were but 7 factories in the country--producing 37,536 tons of beet sugar. 10 years later the number of factories had increased to 63 and the sugar tonnage to 483,621 lbs in 1906 there were 98 factories with a total output of 1,090,621 lbs.

The increase in sugar-beet acreage has kept pace with the growth in the manufacture of beet sugar. In 1900, 872,376 acres of beets were harvested, an increase of almost 700 per cent over the acreage in 1899.

Best-growing areas are located all the way from Ohio to California, but are concentrated in three sections: The middle western, of which the most important States are Michigan, Ohio and Wisconsin; the western mountain section, with Colorado, Utah, and Idaho leading in beet production; and the Pacific coast section in which California is the only important beet-growing State. Table I shows the relative importance of the beet-growing States in 1920.

1Letter from the Secretary of Agriculture, Seventy-first congress, First Session, Senate Document 22, pp. 3, 14.
2U.S. Department of Agriculture, Monthly crop Reporter, April, 1921, P. 38.
3Thid.
Getting Crawlers to Index Our Content

• W3C/Accessibility Conformant HTML
  – Meaningful “alt” tags on images
  – Meaningful, unique <title> values
  – Correct HTML markup enables crawlers to “read” entire page

• Providing metadata
  – Meta tagging with Dublin Core
  – RDFa tagging with Dublin Core
Producing Meaningful Search Results

• Meaningful titles, most specific info first
• As much information as possible to be indexed, and therefore searched and presented in the summary

*Curtain design: [theatrical mask with brown curly wig], VIA ...*  
Inscription: Front: "Design in aplique over the background of the velvet selected for the curtain [...] Fringe of silk tassels, colour-- same as the fruit ...  
via.lib.harvard.edu/via/deliver/deepcontent?recordId=olwork246060 - 9k -  
Cached - Similar pages - Note this
OK. I’ve landed on your page. Now What?

- Where am I?
- Why am I here?
- What else can I do?
- Where else can I go?
Title: Curtain design: [theatrical mask with brown curly wig]
Work Type: drawings
Creator: Komisarevsky, Theodore (1882-1954), painter (artist)
Date: n.d.
Description: Shows large gray-blue or black theatrical mask with brown curly wig against a background of pink with leaves, fruit, and flowers. Design is for a cinema curtain.
Dimensions: 26.7 x 36.8 cm. (10 7/8 x 14 1/2 in.)
Topics: curtains; set design; drawings
Materials/Techniques: Ink and watercolor on paper
Note: General: Signature: Tkomy
   Inscription: Front: 'Design in applique over the background of the velvet selected for the curtain [...] Fringe of silk tassels, colour-- same as the fruit above. [...]" [On reverse:] 'Design in applique for curtain of Cinema designed by Komisarevsky, London area.'
HIC 6,633
Designs - by artist name
Record Identifier: clvwork246050
Related Item

For further information, please see HOLLIS number 10096794

A Harvard University Library Virtual Collection,
Copyright 2008 by the President and Fellows of Harvard College

This record is part of the Chinese Rubbings Collection, which contains more than 2000 digital images of Chinese rubbings that capture Buddhist and Daoist scriptural texts dating from the Qin Dynasty (221-207 BCE) to the Ming Dynasty (1368-1644 CE) that were carved on stone slabs, cave walls, bronze vessels, jade, ceramics, roof tiles, and other materials. The rubbings themselves date from the Ming Dynasty to about 1940 and are highly accurate, often unique sources for scholars of Chinese history, epigraphy, and related disciplines.
Summary

Common Re-engineering Elements

- W3C-compliant HTML code
- Robots meta-tagging
- Support “If-Modified-Since”
- Unique titles, most specific info first
- As much meta-data as possible
- Non-frames presentation
- Links back to full presentation
- Periodic generation of static index files
Results

Google Referrals

- TED
- VC
- OASIS/10
- PDS
Significant Increase in Off-Campus Users

Virtual Collections

Released mid-June, 2008

Unique visitors

<table>
<thead>
<tr>
<th>Month</th>
<th>Visitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 2008</td>
<td>1714</td>
</tr>
<tr>
<td>Jun 2008</td>
<td>1419</td>
</tr>
<tr>
<td>Jul 2008</td>
<td>3199</td>
</tr>
<tr>
<td>Aug 2008</td>
<td>3678</td>
</tr>
<tr>
<td>Sep 2008</td>
<td>3778</td>
</tr>
<tr>
<td>Oct 2008</td>
<td>4363</td>
</tr>
</tbody>
</table>
Significant Increase in Off-Campus Users

VIA Collections

Ongoing phased release from September, 2008

Unique Visitors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique Visitors</td>
<td>2645</td>
<td>2439</td>
<td>1886</td>
</tr>
</tbody>
</table>
Conclusions

• Conformant code
• Frames-free navigable site, and/or site map
• Disambiguate pages as much as possible
• Consider how you want your pages presented in search results listings
• Understand crawler conventions
• When designing new systems: incorporate “crawlability” into your design!!
Questions?

Bobbi Fox
Digital Library Software Engineer
bobbi_fox@harvard.edu
Office for Information Systems
Harvard University Library