



# Deep Web Content and Internet Discovery: Exposing Harvard University Library's Digital Resources to Search Engines

DLF Fall Forum 2008: Providence, RI  
Roberta Fox, Michael Vandermillen, and Spencer McEwen  
Office for Information Systems  
Harvard University Library

Copyright © 2008 by the President and Fellows of Harvard College



# OK, we built it...



*Laying rail on the Mohawk Division c.1932. Industrial Life Photograph Collection. Baker Library Historical Collections. Harvard Business School. .*

## ...did they come?



# Harvard's Library Digital Initiative

*"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



# The web environment in 1998

- 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 Official 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: TEmplated 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



Harvard University Library

## Visual Information Access

Quick search  [GO](#)



[About VIA](#)

[Search](#)

[Browse](#)

[Search History](#)

[Portfolios](#)

[Help](#)

### Search VIA

[? Tips: On search queries](#) [On display preferences](#)

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 repository. Restrictions on access may be noted in the VIA record.

For more detailed information, see [About VIA](#).

Search for:  in [Anywhere](#)

and  in [Anywhere](#)

and  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  to

Limit repository to: [All](#)

#### Search

#### Hints:

Use \* as a wildcard. Examples: cat\*, \*operable and \*itics\*

#### Display/Sort Preferences:

Image grid size: [Small - 3 rows x 5 columns](#)

Sort by: [None](#) [None](#) [None](#)

Result sets greater than 2000 will not be sorted



# The Problem

## MCZ Ernst Mayr Library Artwork Collection

Watercolors, drawings, and other images  
from the Museum of Comparative Zoology  
Library's Special Collections

[Home](#)  
[About MCZ Collection](#)  
[Search](#)  
[Browse Indexes](#)  
[Search History](#)  
[Get Selected Records](#)  
[Help](#)

[MCZ Ernst Mayr Library](#)  
[Museum of Comparative Zoology](#)  
[MCZ Fish Collection Database](#)  
[OASIS Ernst Mayr Library](#)  
[Finding Aids](#)

Record 1 of 1 Jump to record #

[Back to Results Display](#)

**Search Terms:** arc 209-201 in (Anywhere) and sorted by (Physical Piece ID)

**Number of Hits:** 1

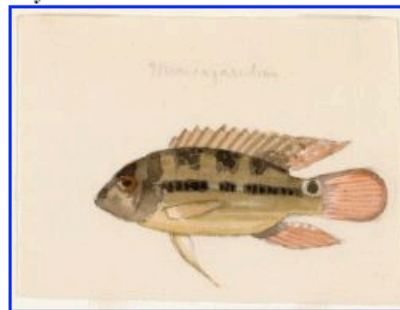
[Expand All](#) [Collapse All](#)

To save a record check the parts of the record you want to keep and click save.

**Record Number:** 1210

- ☐ — MCZ Artwork
  - — Artwork

- **Physical Piece Id:** **ARC 209-201**



- **Title:** Aequidens pallidus
  - **Image Creator:** Burkhardt, Jacques
  - **Work Type:** drawings
  - **Materials/Techniques:** watercolor
  - **Support:** paper
  - **Dimensions:** 8 x 10 cm.
  - **Image Description:** 1 fish drawing (5 x 8 cm.)
  - **Expedition Name:** Thayer Expedition
  - — Annotation



# The Problem

[Home](#)

[Search](#)

[Browse Indexes](#)

[Search History](#)

[Help](#)

[HOLLIS Catalog](#)

[Rubel Chinese Rubbings  
Collection](#)

## Full Record Display

Record 1 of 4 for search: **portrait of confucius in (Anywhere)** .

Jump to record #

### Title

Portrait of Confucius  
Kongzi  
Zhi sheng xian shi xiang  
Kong Qiu xiang  
Kong zi xiang

### Name/Creator

Confucius, 551 BC-479 BC, China, China, associated names

### VIA ID

[olwwork279752](#)

### Digital Object



### Location

Harvard Fine Arts Library, Special Collections; F-31

### Creation Date

19th. century?

### Description

145 x 111 cm

### Form/Genre

still image

rubbings



# The Problem

Harvard University Library Visual Information Access

Record 1 of 1

[Bookmark this item](#)



Image caption: Screen size





# The Problem

Harvard University Library

**OASIS: Online Archival Search Information System**

Questions or comments Copyright Statement OASIS Search


Full Finding Aid Summary: Context Digital Content Bring main window to front Easy print view Help Save ...

Dickinson family. Dickinson family photographs, ca. 1840-1940: Guide. (MS Am 1118.99b ) [Persistent ID: [nrs.harvard.edu/urn-3:FHCL:Hough:hough00049](https://nrs.harvard.edu/urn-3:FHCL:Hough:hough00049)]  
Houghton Library, Harvard College Library, Harvard University

### Table of Contents


Expand All Collapse All

- [Descriptive Summary](#)
- [Processing Information](#)
- [Acquisition Information](#)
- [Use Restrictions](#)
- [Arrangement](#)
- [Scope and Content](#)
- [Related Material](#)
- [Container List](#)
  - [I. Individuals and groups.](#)
  - [II. Places.](#)




[Click for larger view](#)

- ◊ (28) Dickinson, Lavinia Norcross, 1833-1899. Portraits, 1852 and 1896. 4 items: cabinet photographs and others.  
Item 1 image is the same as daguerreotype in no. (27) above; photographed by J.L. Lovell of Amherst, Mass. Items 2-4 are an image of LD with a cat (1896).
- ◊ (28a) Dickinson, Lavinia Norcross, 1833-1899. Portrait, undated. 1 item: photograph.  
Item 1 image is the same as daguerreotype in no. (27) above (without frame). Annotated on verso: Lavinia Dickinson Amherst, Mass. \*46M-290.
- ◊ (29.1) **Shelved in box 5** Dickinson, Susan Huntington, 1830-1913. Portrait, undated. 1 item: daguerreotype in case.



[Click for larger view](#)

- ◊ (29.2) **Shelved in box 5** Dickinson, Susan Huntington, 1830-1913. Portrait, undated. 1 item: daguerreotype in case.





# The Problem

Harvard University Library  
Page Delivery Service

*Child labor and the work of mothers in the beet fields of Colorado and Michigan. Washington: G.P.O., 1923.*

SearchView TextPrintable VersionRelated LinksHelpCopyright

Page: 1Go toSequence: 10Go toImage Size: □ □ ■ □

Expand AllCollapse All

☒ Child labor and the work of moth  
Front Matter [pp. unnumbered pa.  
Contents [pp. iii-iv (seq. 4-5)]  
**Body** [pp. v-122 (seq. 6-137)]

## OF COLORADO AND MICHIGAN.

---

### 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.<sup>1</sup> In 1920 there were 98 factories with a total output of 1,090,021 tons.<sup>2</sup>

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 harvested<sup>3</sup>, an increase of almost 700 per cent over the acreage in 1899.<sup>4</sup>

Beet-growing areas are located all the way from Ohio to California,

Frames



# 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

```
<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`



# Being Crawler-Friendly - PDS

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



# Being Crawler-Friendly - PDS



Harvard University Library  
Page Delivery Service

*Child labor and the work of mothers in the beet fields of Colorado and Michigan. Washington: G.P.O., 1923.*

Sequence 10 of 137 (Page 1)

[View framed version of this document](#)

CHILD LABOR AND THE WORK OF MOTHERS IN THE BEET FIELDS  
OF COLORADO AND MICHIGAN.

## 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.<sup>1</sup> in 1920 there were 98 factories with a total output of 1,090,021 tons.<sup>2</sup>

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 harvested ~, an increase of almost 700 per cent over the acreage in 1899.~

Beet-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.

<sup>1</sup>Letter from the Secretary of Agriculture, Sixty-first congress, First Session, Senate Document 22, pp. 3, 14.

<sup>2</sup>U. S. Department of Agriculture, Monthly crop Reporter, April, 1921, P. 38.

<sup>3</sup>Ibid.

<sup>4</sup>Thirteenth census of the United States, 1910, Vol. V, Agriculture, p. 691. Washington, 1913.

1



[Table of Contents](#)

[View Image](#)





# 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 applique 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](http://via.lib.harvard.edu/via/deliver/deepcontent?recordId=olwork246060) - 9k -

[Cached](#) - [Similar pages](#) - [Note this](#)



# Context

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?



# A VIA Page Reached via Search Engine

VIA (Visual Information Access) is a growing online union catalog documenting the arts, material culture, and social history.



Refer to the [Full Record View](#) of this work for full VIA functionality

Harvard University Library **Visual Information Access**

Components: 2 Images

## Work



Screen size



Detailed

**Title:** Curtain design: [theatrical mask with brown curly wig]  
**Work Type:** drawings  
**Creator:** Komisarjevsky, Theodore (1882-1954), painter (artist)  
**Date:** n.d.  
**Description:** Shows large grey-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 Komisarjevsky, London area."  
*Provenance:* Gift of Ernestine Stodelle Komisarjevsky Chamberlain, 1957.  
**Repository:** Harvard Theatre Collection. Part of LDI project: Russian Theatrical Designs in the Harvard Theatre Collection.  
HTC 6,633  
Designs - by artist name  
**Record Identifier:** olvwork246060



# A VIA page from the VIA Portal



Harvard University Library  
**Visual Information Access**

Quick search

GO

About VIA

Search

Browse

Search History

Portfolios

Help

Full Record

Grid View

Save selected items or Bookmark ...

Search Results

[Show Selected](#) [Show only digital images](#)

Record 1 of 1

Click on image to select. Double click for large image. Click on title to view data. Pop-ups must be enabled.

## Work

**Title:** Curtain design: [theatrical mask with brown curly wig]  
**Creator:** Komisarjevsky, Theodore  
**Date:** n.d.

**Components:** 1 to 2 of 2 for record 1 from the search **Curtain design: [theatrical mask with brown curly wig] in (Anywhere)**



Screen size



Detailed

**Components:** 1 to 2 of 2 for record 1 from the search **Curtain design: [theatrical mask with brown curly wig] in (Anywhere)**

[Harvard Libraries Home](#) | [Copyright and Permissions](#) | [Questions or Comments](#) | For email updates, [join viainfo list](#)

About VIA

Search

Browse

Search History

Portfolios

Help



# Context paragraph in a VC page

<http://vc.lib.harvard.edu/vc/deliver/~rubblings/olvwork352661>

## 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.*



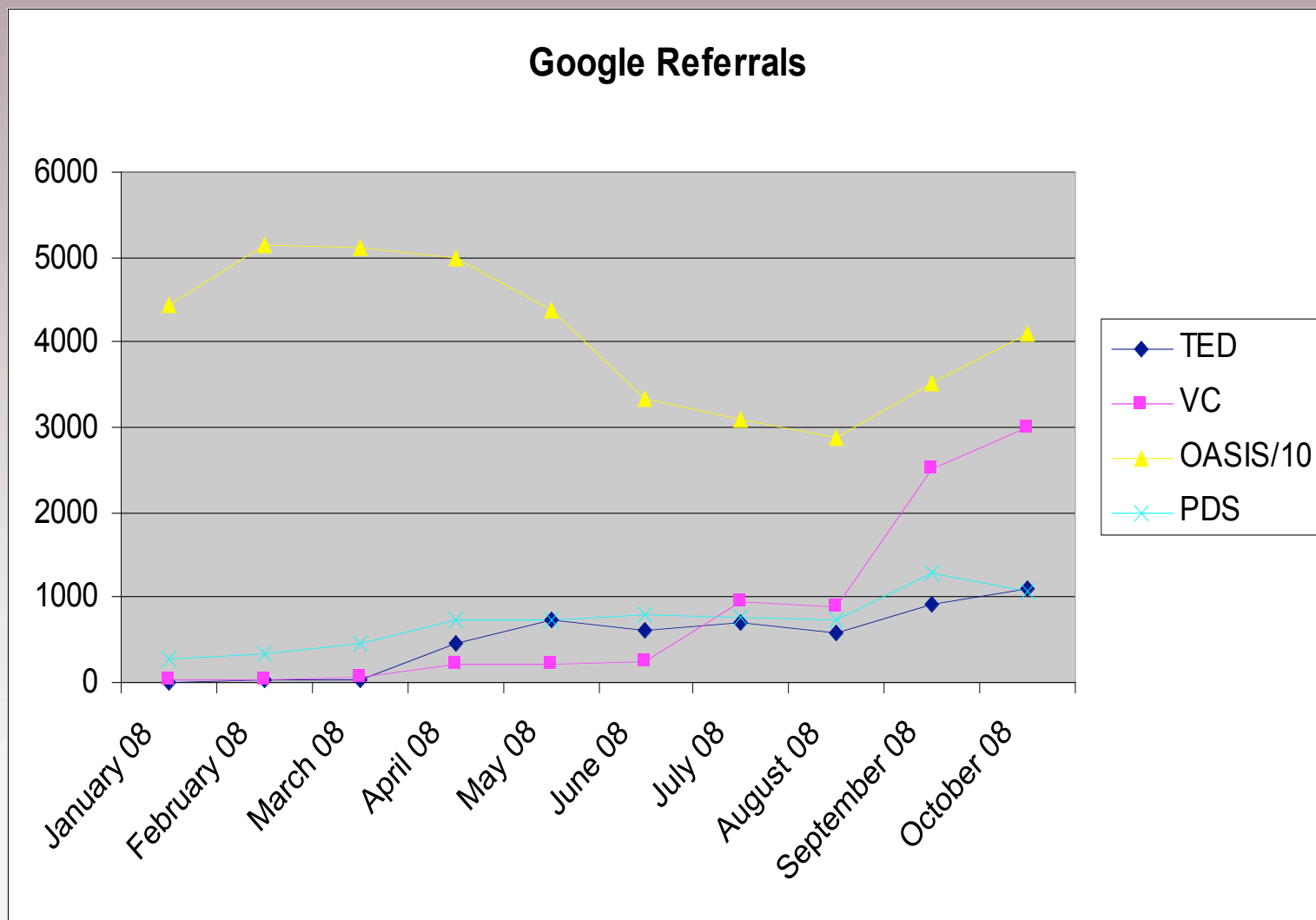
# Summary

## 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

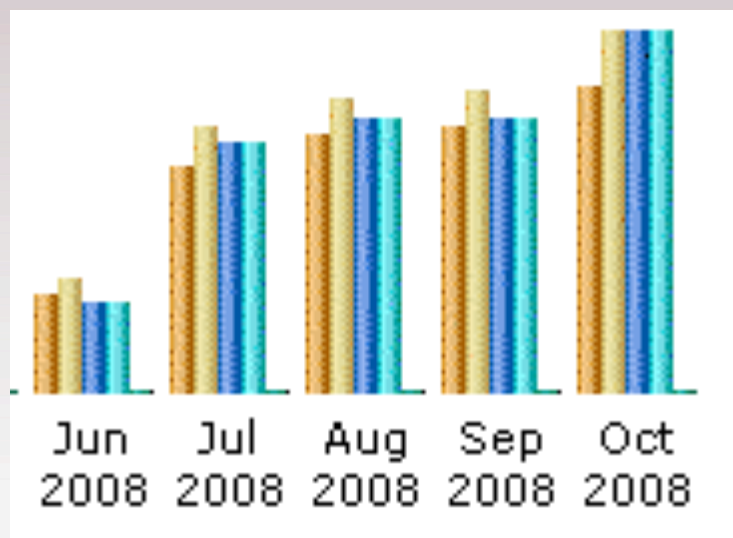




# Significant Increase in Off-Campus Users

## Virtual Collections

Released mid-June, 2008



Oct 2008	4363
Sep 2008	3778
Aug 2008	3678
Jul 2008	3199
Jun 2008	1419
May 2008	1714

Unique visitors

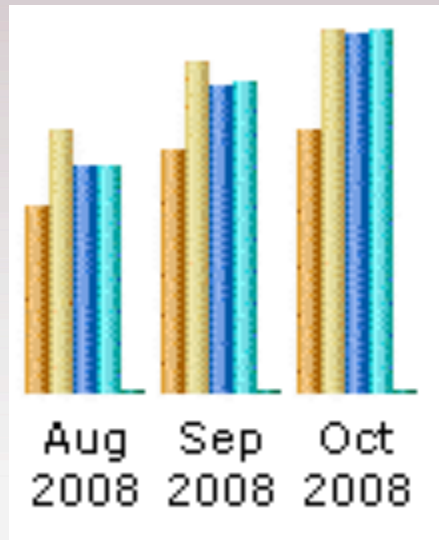




# Significant Increase in Off-Campus Users

## VIA Collections

Ongoing phased release from September, 2008



Oct 2008	2645
Sep 2008	2439
Aug 2008	1886

Unique Visitors



# 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](mailto:bobbi_fox@harvard.edu)

Office for Information Systems

Harvard University Library