SEARCH ENGINE OPTIMIZATION
FOR DIGITAL COLLECTIONS

Kenning Arlitsch
Patrick OBrien
Sandra McIntyre
Agenda

- Assessment
- Phase 1: Start feedback loop
- Phase 2: Get indexed
- Phase 3: Increase visibility (future)
- Wrap-up
Context and history at Utah

- Large digital library programs
  - Mountain West Digital Library
  - Utah Digital Newspapers
  - Western Soundscape Archive
  - Western Waters Digital Library
- Digital collections are “Deep Web”
- Google indexing diminished recently
  - Ceased OAI harvest in August 2008
  - Average as low as 8% in spring 2010
Initial Repositories Survey

- Surveyed 13 repositories of the MWDL in July
  - 10 CONTENTdm
  - 1 Digital Commons
  - 1 ArchivalWare
  - 1 home grown (HEAL)
- Randomly selected 50 objects from each (650)
- Searched by title in Google and Google Images
  - 38% find rate in Google
  - Almost 0% in Google Images
MWDL Repositories Survey

% w/ Indirect URL

- Utah Digital Newspapers Repository
- University of Nevada, Reno
- University of Utah
- Southern Utah University
- Brigham Young University
- Utah State University
- Utah State Archives
- Utah State University
- Utah Valley University
- Weber State University
- Health Education Assets Library
- University of Nevada, Las Vegas
- Utah State Library

0% 25% 50% 75% 100%
Discoverability of digital resources

- Priority Collections
  - Institutional Repository (USpace)
  - Special Collections EAD finding aids
  - University Press

- Discoverability is important for
  - Faculty (contributors and users)
  - Donors
  - Students
Where College Students Begin Searching

Note: Only electronic resources with usage rates of 1 percent or more are represented on this graph.
comScore Releases September 2010 U.S. Search Engine Rankings

RESTON, VA, October 13, 2010 – comScore, Inc. (NASDAQ: SCOR), a leader in measuring the digital world, today released its monthly comScore qSearch analysis of the U.S. search marketplace. Google Sites led the explicit core search market in September with 66.1 percent of searches conducted, an increase of 0.7 share points from August 2010.

The September 2010 qSearch figures reflect the impact of Google Instant Search, Google's new feature that delivers results in real-time while users type their query. To learn more about how comScore is measuring search activity as users engage with Google Instant Search, please read our recent blog post on the subject:
http://blog.comscore.com/2010/10/comscore_september_qsearch.html

U.S. Explicit Core Search

Google Sites led the U.S. explicit core search market in September with 66.1 percent market share, followed by Yahoo! Sites with 16.7 percent and Microsoft sites with 11.2 percent. Ask Network captured 7.7 percent of explicit core searches, followed by AOL LLC Network with 2.3 percent.

<table>
<thead>
<tr>
<th>Core Search Entity</th>
<th>Explicit Core Search Share (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Explicit Core Search</td>
<td>Aug-10</td>
</tr>
<tr>
<td>Total</td>
<td>100.0%</td>
</tr>
<tr>
<td>Google Sites</td>
<td>66.1%</td>
</tr>
<tr>
<td>Yahoo! Sites</td>
<td>16.7%</td>
</tr>
<tr>
<td>Microsoft Sites</td>
<td>11.2%</td>
</tr>
<tr>
<td>Ask Network</td>
<td>7.7%</td>
</tr>
<tr>
<td>AOL LLC Network</td>
<td>2.3%</td>
</tr>
</tbody>
</table>

**“Explicit Core Search” excludes contextually driven searches that do not reflect specific user intent to interact with the search results.**

More than 16.0 billion explicit core searches were conducted in September. Google Sites ranked first with 10.6 billion searches, followed by Yahoo! Sites in second with 2.7 billion and Microsoft Sites in third with 1.8 billion. Ask Network accounted for 583 million explicit core searches followed by AOL LLC Network with 362 million.
Literature Review

Literature Lessons

- Most are dated
- Most deal with general websites
- “Black hat” techniques get you banned
- Few deal with digital collections in db’s
- Some suggest duplicating the content outside the database
Problems evident on several levels

- Web server
  - robots.txt
  - Crawler errors
- Application layer (repository software)
  - URL redirects
  - Many URLs for same objects
- Presentation layer
  - HTML and Graphic design
- Metadata issues
External Influence: Search Engine Policies

- Rules and enforcement levels change
  - OAI harvesting
  - Sitemaps
- Requirements & standards adoption
  - W3C, Highwire, etc.
- Insensitive to standards valued by librarians
  - “Use Dublin Core tags (e.g., DC.Title) as a last resort”*

* Google Scholar Inclusion Guidelines for Webmasters
Agenda

- Assessment
- Phase 1: Start feedback loop
- Phase 2: Get indexed
- Phase 3: Increase visibility (future)
- Wrap-up
Google and Digital Assets Management

- 2008: Google announced it would no longer crawl Open Archives Initiative (OAI) streams
- Many digital collections have been slowly “disappearing” from Google since then
- What’s going on?
- What’s needed instead?
Phase 1:
Learning about Web Crawlers

Queue of URLs → Pick one URL → Fetch page → Parse page → Add URLs to queue → Index and store data
Phase 1: Notifying crawlers about dynamic pages

- Digital asset management systems construct pages in HTML on the fly
  - Header
  - Record retrieved from database and formatted
  - Footer

Item Page
Phase 1:  
Notifying crawlers about dynamic pages

- Have to tell crawler how to assemble it (with URL)
Google Sitemaps

- Sitemap file for each collection
  
  “Here is a list of the URLs of the dynamic pages that I want you to crawl, one for each item.”

- Sitemap Index file to list all the Sitemaps
  
  “Here is a list of all the Sitemap files.”

- Protocol: [http://www.sitemaps.org](http://www.sitemaps.org)
Start the feedback loop

- Create Sitemaps, one for each collection, and Sitemap Index.
- Register with Google Webmaster Tools.
- Inform Google about the location of Sitemap Index.
  - In the robots.txt file at the root on the server
- Monitor crawler results in Webmaster Tools.
Initial experiments and theories: Presentation layer

- Compound objects – frameset
- Page titles
- Putting metadata up in head as <meta> tags
Monitor crawler results

- Webmaster Tools
  - Top search queries
  - Links to your site
  - Keywords
  - Internal links
  - Crawl errors
  - Crawl stats
  - HTML suggestions
Phase 1 results:
Feedback loop is in place

- Webmaster Tools shows us results
  - Incomplete indexing
  - Lots of crawler errors
  - Inconsistencies across collections
  - Low ranking on search engine listings
Cross-departmental collaboration

- Search Engine Optimization (SEO) Team
  - Associate Director for IT Services
    - Server administrators
    - Programmers
    - Digital Initiatives Librarian
  - Collection managers and other metadata experts

- SEO consultant volunteered services: Patrick OBrien of RevX Corp.
Agenda

- Assessment
- Phase 1: Start feedback loop
- Phase 2: Get indexed
- Phase 3: Increase visibility (future)
- Wrap-up
Know your customers and what they value.

- **Faculty**
  - High Value
  - Publication Page Views
  - Publication Downloads
  - Requests for Information
  - Publication Citations

- **Collection Donors**
  - High Value
  - Digital Collection Pages Indexed
  - Digital Collection Page Views
  - Digital Collection Visitors
  - Requests for More Info
  - Physical Collection Visitors
  - Reproductions Ordered
Phase 2 goals and results

- Increase the number of Digital Collection web pages in the Google search engine.
- Develop a program to maximize a collections visibility and reach.

<table>
<thead>
<tr>
<th>Goals</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

EAD Finding Aids

Google URL Index Ratio

- Baseline
- Pilot
Phase 2 goals and results

Goals

- Increase the number of Digital Collection web pages in the Google search engine.
- Develop a program to maximize a collections visibility and reach

Results

- IR Articles*
  - Baseline: 0
  - Pilot: 100

* [site:content.lib.utah.edu as of April 24, 2010](http://site:content.lib.utah.edu) as of November 01, 2010
Why can’t the public find our content?

- Are you worthy enough for their customer (i.e Index)?
- How much will their customer value the introduction (i.e, Visibility)?
The Digital Collection environment is complex and very difficult for robots to index.

- Multiple Web Server Technologies
- Complex Application Platforms
- Different Metadata Organization, Context, and process
- Constantly changing Search Engine Requirements

Crawl errors = 1,000+ per Day
Are you worthy enough for their customers (i.e. Index)

- Reduce Google Crawl Errors
- Developed efficient Google Crawler path
- Reconfigured the environment to meet Google’s requirements
Check the Crawl Errors in Google Webmaster

- Page Forbidden (401 errors)
- User Not Authorized (403 errors)
- Network Unreachable (5xx errors)
- Page Not Found (404 errors)
Eliminate sitemap & robots.txt conflicts

Robots.txt
User-agent: *
Disallow: /dmscripts/
Disallow: /cdm4/admin/
Disallow: /cdm4/client/
Disallow: /cdm4/cqr/
Disallow: /cdm4/images/
Disallow: /cdm4/includes/
Disallow: /cdm4/jscripts/
Disallow: /cdm-diagnostics/
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /u/

Sitemap
http://content.lib.utah.edu/cgi-bin/browseresults.exe?CISOROOT=/DC_Beckwith
Address errors and don’t leave their customers stranded!

Low Trust Example
403 Error

Directory Listing Denied
This Virtual Directory does not allow contents to be listed.

<table>
<thead>
<tr>
<th>How to Fix It</th>
<th>Example</th>
</tr>
</thead>
</table>
| Inform the Client Browser     | <title>HTTP 403 Error</title>
|                               |   <meta HTTP-EQUIV = "Refresh" CONTENT = "8; URL ="/>
|                               |   <meta NAME="robots" CONTENT="NOINDEX,NOFOLLOW">                     |
| Inform the Search Engine      | <?php
|                               |   header("HTTP/1.1 403 Forbidden");
|                               |   header("Location: http://content.lib.utah.edu/");
|                               |   ?>                                                                   |
| Inform Their Customer         | <p>The page you requested is no longer available or has been moved. </p>
|                               |   <p>You will be taken to our opening home page within the next 5 seconds. </p> |
A Papermaking Pilgrimage to Japan, Korea and China
Provide path with context using simple URLs

Agenda

- Assessment
- Phase 1: Start feedback loop
- Phase 2: Get indexed
- Phase 3: Increase visibility (in progress)
- Wrap-up
Multiple Dynamic URLs pointing to a single URI

- Example: same content had 2+ URLs
  - http://content.lib.utah.edu/u/?/ir-main,5239
  - content.lib.utah.edu/cdm4/document.php?
    CISOROOT=/ir-main&CISOPTR=370&CISOSHOW=5239

- Implemented Canonical Link Element to clarify 500+ URL Parameters
"Use Dublin Core tags (e.g., DC.title) as a last resort - they work poorly for journal papers...

- Google Scholar Inclusion Guidelines for Webmasters
Embed bibliographic metadata in HTML & full text PDF files

- Mapped Dublin Core to a Google supported HTML meta tag
  - Highwire Press (e.g., citation_title)
- Extended Dublin Core fields
  - Journal Title
  - Journal Volume
  - Journal Issue
  - Starting Page Number
  - Ending Page Number
- Link directly to existing Full Text PDF
Link data to establish context and improve visibility

- Apply Taxonomy Schemas
  - Glossary
  - Acronyms
- External Linking
  - Authors
  - Organizations
  - External Feeds
- Target Audience Segments with Declared Ontology’s
Agenda

- Assessment
- Phase 1: Start feedback loop
- Phase 2: Get indexed
- Phase 3: Increase visibility (future)
- Wrap-up
Lessons Learned

- Search engines want to send users to content that solves users’ problem, not just to metadata
- Establish trust
- Linking strategies enormously important
  - Chicken and egg problem
- Ensure metadata is unique and descriptive
  - Dublin Core too ambiguous
  - Different audiences use different vocabularies
- Accessibility standards good for SEO
Managing expectations

- SEO-SEM is a long-term strategy that requires constant monitoring
- Build a good site that is useful to people and engines will find it
- Search engine is the customer
- Influence vendors to add SEO features into products
Q&A

- Kenning Arlitsch
  - Associate Director for IT Services, Univ of Utah
  - kenning.arlitsch@utah.edu

- Sandra McIntyre
  - Program Director, Mountain West Digital Library
  - sandra.mcintyre@utah.edu

- Patrick O’Brien
  - Principal, RevX Corporation
  - patrick@revxcorp.com
Google Sitemap – example

http://content.lib.utah.edu/sitemaps/sitemap_ir-main-001.xml

<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9 http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd"
  xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <lastmod>2010-01-08T11:19:13-07:00</lastmod>
  </url>
  <url>
    <lastmod>2010-01-08T11:19:13-07:00</lastmod>
  </url>
  <url>
    <lastmod>2010-01-08T11:19:13-07:00</lastmod>
  </url>
  <url>
    <loc>http://content.lib.utah.edu/ehd4/item_viewer.php?CISOROOT=/ir-main&amp;CISOPTR=34199</loc>
    <lastmod>2010-01-08T11:19:13-07:00</lastmod>
  </url>
  <url>
    <loc>http://content.lib.utah.edu/ehd4/item_viewer.php?CISOROOT=/ir-main&amp;CISOPTR=34208</loc>
    <lastmod>2010-01-08T11:19:13-07:00</lastmod>
  </url>
  <url>
Sitemap Index - example

http://content.lib.utah.edu/cdm4/autositemap/sitemapindex.xml

<?xml version="1.0" encoding="UTF-8"?>
<sitemapindex xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9 http://www.sitemaps.org/schemas/sitemap/0.9/siteindex.xsd"
xmllns="http://www.sitemaps.org/schemas/sitemap/0.9">
    <sitemap>
        <loc>http://content.lib.utah.edu/sitemaps/sitemap_altaav-001.xml</loc>
        <lastmod>2010-01-08T11:50:41-07:00</lastmod>
    </sitemap>

    <sitemap>
        <loc>http://content.lib.utah.edu/sitemaps/sitemap_uw-001.xml</loc>
        <lastmod>2010-01-08T11:50:41-07:00</lastmod>
    </sitemap>

    <sitemap>
        <loc>http://content.lib.utah.edu/sitemaps/sitemap_uuappp-001.xml</loc>
        <lastmod>2010-01-08T11:50:41-07:00</lastmod>
    </sitemap>

    <sitemap>
        <loc>http://content.lib.utah.edu/sitemaps/sitemap_Aztec-001.xml</loc>
        <lastmod>2010-01-08T11:50:41-07:00</lastmod>
    </sitemap>

    <sitemap>
        <loc>http://content.lib.utah.edu/sitemaps/sitemap_Caesaris-001.xml</loc>
        <lastmod>2010-01-08T11:50:41-07:00</lastmod>
    </sitemap>

    <sitemap>
        <loc>http://content.lib.utah.edu/sitemaps/sitemap_coa-001.xml</loc>
        <lastmod>2010-01-08T11:50:41-07:00</lastmod>
    </sitemap>

    <sitemap>
        <loc>http://content.lib.utah.edu/sitemaps/sitemap_utleaven-001.xml</loc>
        <lastmod>2010-01-08T11:50:41-07:00</lastmod>
    </sitemap>
</sitemapindex>
Step 1: Create Sitemaps and Index

- According to the protocol at http://www.sitemaps.org:
  - Create a Sitemap file for each collection.
  - Create a Sitemap Index file.
Step 2: Webmaster Tools Registration

- Register (free) with Google Webmaster Tools at http://www.google.com/webmasters/tools
Step 2: Webmaster Tools Registration
Step 3: Inform Google

- **Step 3A:** Submit the address of Sitemap Index file on Webmaster Tools.
Step 3: Inform Google

- Step 3B: Modify the robots.txt file at the root of your CONTENTdm server to specify the location of the Sitemaps Index.

```plaintext
User-agent: *
Disallow: /dmscripts/
Disallow: /cdm4/admin/
Disallow: /cdm4/client/
Disallow: /cdm4/cqr/
Disallow: /cdm4/images/
Disallow: /cdm4/includes/
Disallow: /cdm4/javascript/
Disallow: /cdm-diagnostics/
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /w/
Sitemap: http://content.lib.utah.edu/cdm4/autositemap/sitemapindex.xml
```