SEO for Digital Repositories

Kenning Arlitsch & Patrick OBrien
April 5, 2011
CNI Spring Forum, San Diego, CA
Today’s Objectives

- Understand
  - Issues and Opportunity
  - Goals and Framework
  - Key Steps and Questions
SEO Repository Goals

- Digital repositories vs general websites
  - Millions of objects in databases
  - Include IR
- Goal 1 – Increase Reach
  - Get objects indexed in search engines
- Goal 2 – Increase Visibility
  - Provide robust descriptive content
College Students Begin Research - 2005

Note: Only electronic resources with usage rates of 1 percent or more are represented on this graph.
College Students Begin Research - 2010

### Where college students begin their information search

<table>
<thead>
<tr>
<th>Information Source</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search engine</td>
<td>83%</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>7%</td>
</tr>
<tr>
<td>Social networking site</td>
<td>2%</td>
</tr>
<tr>
<td>E-mail</td>
<td>1%</td>
</tr>
<tr>
<td>E-mail subscription/alert</td>
<td>1%</td>
</tr>
<tr>
<td>Online database</td>
<td>1%</td>
</tr>
<tr>
<td>Ask-an-expert site</td>
<td>0%</td>
</tr>
<tr>
<td>Library Web site</td>
<td>0%</td>
</tr>
<tr>
<td>Online bookstore</td>
<td>0%</td>
</tr>
<tr>
<td>Topic-specific Web site</td>
<td>0%</td>
</tr>
</tbody>
</table>

Start with the 800 pound gorilla – Google.

Press Release

comScore Releases February 2011 U.S. Search Engine Rankings

RESTON, VA, March 11, 2011 – 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 February with 65.4 percent of searches conducted.

U.S. Explicit Core Search

Google Sites led the U.S. explicit core search market in February with 65.4 percent and Microsoft sites accounted for 16.1 percent, followed by Ask Network and AOL, Inc.

<table>
<thead>
<tr>
<th>Core Search Entity</th>
<th>Explicit Core Search Share (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jan-11</td>
</tr>
<tr>
<td>Total Explicit Core Search</td>
<td>100.0%</td>
</tr>
<tr>
<td>Google Sites</td>
<td>65.6%</td>
</tr>
<tr>
<td>Yahoo! Sites</td>
<td>16.1%</td>
</tr>
<tr>
<td>Microsoft Sites</td>
<td>13.1%</td>
</tr>
<tr>
<td>Ask Network</td>
<td>3.4%</td>
</tr>
<tr>
<td>AOL, Inc.</td>
<td>1.7%</td>
</tr>
</tbody>
</table>
Management Experiences

- Large digital collections built over a decade
  - 1.3+ million items
- Why weren’t we getting indexed?
  - Harvesting/indexing rates as low as 8%
  - Poor IR showing in Google Scholar
- Sitemaps generated for Google
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%
MWDL Repositories Survey

% w/ Direct URL

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

- Most are dated
- Most deal with general websites
- Few deal with digital collections in db’s
- Some suggest duplicating the content outside the database
Know your stakeholders 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
What do the search engines value?

1) Are you worthy enough for their customer (i.e, Index)?

2) How much will their customer value the introduction (i.e, Visibility)?
Relate risk to organizational functions

Major Barriers

- Administrative/Organizational issues
- Descriptive metadata uniqueness and structure
- Search engines policies and practices
- Server configuration and performance

Organizational Risk Areas

Framework

Descriptive Metadata
Presentation Layer
Application
Web Server
Setup Google Webmaster Tools and ask questions.

- Reduce Google Crawl Errors

- Improve Server Performance
Set goals and establish a baseline ...

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

<table>
<thead>
<tr>
<th>Goals</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilots</td>
<td>EAD Finding Aids</td>
</tr>
<tr>
<td></td>
<td>Google URL Index Ratio</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
</tr>
<tr>
<td></td>
<td>2%</td>
</tr>
</tbody>
</table>

75 pages indexed / 3,221 pages submitted as of April 24, 2010
... with objective performance criteria.

Goals

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

Results

<table>
<thead>
<tr>
<th>EAD Finding Aids</th>
<th>Baseline</th>
<th>Current</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2%</td>
<td>69%</td>
</tr>
</tbody>
</table>

Google URL Index Ratio

- Baseline
- Current

2,239 pages indexed / 3,235 pages submitted as of January 14, 2010
Goal 1: Initial focus was to make it easier for Google to index.

**Initial Priorities**

- Reduce Google Crawl Errors
- Developed efficient Google Crawler path
- Reconfigure the environment to meet Google’s key requirements
Collection Google Index Ratios have increased across the board.

**Collection Google Index Ratio**

*Google Index Ratio = URLs submitted / URLs Indexed by Google*

**Collections with over 500 URLs submitted to Google**
Metadata is a major driver in Google Index Ratio variance

- **Unique Page Titles**
- **Robust Page Descriptions**
- **Defined Ontology / Taxonomy**
- **Relevant outbound links**
Be ready with overwhelming evidence.

... there's a good chance that many of your papers aren't included at all, because documents with the same title are often considered duplicates.

- Google Scholar Inclusion Guidelines for Webmasters

“... incorrect identification of references could lead to exclusion of your papers from Google Scholar or to low ranking of your papers in the search results.”

- Google Scholar Inclusion Guidelines for Webmasters

“...the most common cause of indexing problems is incorrect extraction of bibliographic data by the automated parser software.

- Google Scholar Inclusion Guidelines for Webmasters
Ensure your staff understand the strategic importance of your SEO efforts.

Marriott Strategy

Exploit the Digital and Networked Environments

Marriott Goal

Digitize Collection and share in many venues where users go

Marriott Activity

• Develop strategies, priorities, and procedures for building our digital collections.
• Work to ensure library collections are well placed in search results listings

Elevate our position and impact on campus

Be a model and recognized for our work

• Communicate our work and results more widely in professional journals and conferences
• Tell our story on campus in many venues and opportunities

Diversify and increase the financial base

Obtain more grants for experimentation and projects

• Identify and leverage strategic opportunities and partnerships
Search Engine Policies and Practices

- Rules and enforcement levels change
  - OAI harvesting
  - Sitemaps
- Insensitive to standards valued by librarians
  - “Use Dublin Core tags (e.g., DC.Title) as a last resort”*
- Scholar wants Highwire Press, PRISM, Be Press, Eprints metadata schema

* Google Scholar Inclusion Guidelines for Webmasters
Promote the “Right way” and set policy to prevent the wrong way for SEO.

- Recent Black Hat news stories
  - JC Penney
  - Overstock

- Staff must know the difference, and that black hat techniques can get you banned
  - Establish policies
Administrative Issues

- Not just about the technology
  - Cross-departmental staff work together
    - Sitemaps vs. robots.txt
  - Develop skill sets
  - Staff can become self-directed if they understand the goals

- Relevance
  - Metrics must support organizational goals
Questions & Contacts?

Kenning Arlitsch  
kenning.arlitsch@utah.edu

Patrick OBrien  
www.RevXcorp.com