SEO for Digital Repositories

Kenning Arlitsch and Patrick OBrien
OCLC TAI CHI Webinar
March 16, 2012
Agenda

- Brief discussion of two years of research
  - Priorities & Results
  - Issues & Opportunity
  - Key Steps & Questions
  - Google Scholar
Marriott Library Management Experiences

- Large digital collections built over a decade
  - 1.5+ million items
- Why weren’t we getting indexed?
  - Google index ratios as low as 2%
  - Poor IR showing in Google Scholar
Marriott Library SEO program priorities

- Digital repositories vs. general websites
  - Millions of objects in databases
  - Include IR

- Priority 1 – Increase Reach
  - Get objects indexed in search engines

- Priority 2 – Increase Visibility
  - Provide robust descriptive content
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
Crucial Findings

- Technology is only a piece of the search engine optimization (SEO) puzzle
- SEO has to be addressed from a combination of leadership, management, and communication
- Only some of the stakeholders and some of the problems are based in IT
- Search engine crawlers must:
  - Easily access content
  - Interpret content as machine-readable data

* For full text of “Authors' response Re: Google Scholar discoverability of repository content”
See https://groups.google.com/a/arl.org/group/sparc-ir/browse_thread/thread/77d7b3a1aabee4f5?pli=1
Collection Google Index Ratios have increased across the board...

Google Index Ratio - All Collections*

- Average:
  - 07/05/10: 12%
  - 04/04/11: 51%
  - 11/30/11: 79%

- High**:
  - 07/05/10: 37%
  - 04/04/11: 87%
  - 11/30/11: 100%

* Google Index Ratio = URLs submitted / URLs Indexed by Google for about 150 collections containing ~170,000 URLs
**Highest index ratio achieved for Collections with over 500 URLs submitted to Google
...increasing Google referrals by 200% and total visitors by 79%.

<table>
<thead>
<tr>
<th>Initial Referrer</th>
<th>Custom View: 2/6/11 - 4/30/11 Visits</th>
<th>Custom View: 2/7/10 - 5/1/10 Visits</th>
<th>% Change Visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>google.com</td>
<td>28,841</td>
<td>9,564</td>
<td>201.56%</td>
</tr>
<tr>
<td>utah.edu</td>
<td>26,440</td>
<td>14,783</td>
<td>78.85%</td>
</tr>
<tr>
<td>Direct Traffic</td>
<td>22,810</td>
<td>16,010</td>
<td>42.47%</td>
</tr>
<tr>
<td>bing.com</td>
<td>3,574</td>
<td>4,376</td>
<td>-18.33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Visitor Summary</th>
<th>Custom View: 2/6/11 - 4/30/11</th>
<th>Custom View: 2/7/10 - 5/1/10</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visitors</td>
<td>89,124</td>
<td>49,774</td>
<td>79.06%</td>
</tr>
<tr>
<td>Visitors Who Visited Once</td>
<td>82,310</td>
<td>44,553</td>
<td>84.75%</td>
</tr>
<tr>
<td>Visitors Who Visited More Than Once</td>
<td>6,814</td>
<td>5,221</td>
<td>30.51%</td>
</tr>
</tbody>
</table>
Agenda

- Brief discussion of two years of research
  - Priorities & Results
  - Issues & Opportunity
  - Key Steps & Questions
  - Google Scholar
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

- Search engine: 83%
- Wikipedia: 7%
- Social networking site: 2%
- E-mail: 1%
- E-mail subscription/alert: 1%
- Online database: 1%
- Ask-an-expert site: 0%
- Library Web site: 0%
- Online bookstore: 0%
- Topic-specific Web site: 0%

Start with the 800 pound gorilla

<table>
<thead>
<tr>
<th>Core Search Entity</th>
<th>Explicit Core Search Share (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jan-12</td>
</tr>
<tr>
<td>Total Explicit Core Search</td>
<td>100.0%</td>
</tr>
<tr>
<td>Google Sites</td>
<td>66.2%</td>
</tr>
<tr>
<td>Microsoft Sites</td>
<td>15.2%</td>
</tr>
<tr>
<td>Yahoo! Sites</td>
<td>14.1%</td>
</tr>
<tr>
<td>Ask Network</td>
<td>3.0%</td>
</tr>
<tr>
<td>AOL, Inc.</td>
<td>1.6%</td>
</tr>
</tbody>
</table>
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

Average = 38%

October 2010
Example of indirect URL

<table>
<thead>
<tr>
<th>Image</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Topaz Times 1942-05-30 Want to Get Married?</td>
</tr>
<tr>
<td>2.</td>
<td>Topaz Times 1942-08-15 Cupid Snares First Center Pair; Wedding on Sunday</td>
</tr>
<tr>
<td>3.</td>
<td>Topaz Times 1942-10-21 Marriage</td>
</tr>
<tr>
<td>4.</td>
<td>Topaz Times 1942-10-21 Marriage</td>
</tr>
</tbody>
</table>
Users and search engines want links directly to relevant content.
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

Average = 13%

October 2010
Agenda

- Brief discussion of two years of research
  - Priorities & Results
  - Issues & Opportunity
  - Key Steps & Questions
  - Google Scholar
Know your stakeholders and what they value

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

- **Archivists**
  - Digital Collection Pages Indexed
  - Digital Collection Page Views
  - Digital Collection Visitors
  - Requests for More Info
  - Physical Collection Visitors
  - Reproductions Ordered
  - Value: High
Set repository 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

Goals

Pilot Results

EAD Finding Aids

Google URL Index Ratio

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

**Pilot Results**

**EAD Finding Aids**

- 80%

**Google URL Index Ratio**

- Baseline
- 03/13/12

2,675 pages indexed / 3,332 pages submitted as of March 13, 2012
Ensure your staff understand the strategic importance of your SEO efforts

<table>
<thead>
<tr>
<th>Library Strategic Plan</th>
<th>SEO Program Activities</th>
</tr>
</thead>
</table>
| **Exploit the Digital and Networked Environments** | • Optimize Collections to Improve Visibility  
  • Descriptions  
  • Link Popularity  
  • Page Elements  
  • Metadata  
  • Calls to Action  
  • Develop Metrics Dashboard to monitor and improve efforts |
| **Elevate our position and impact on campus** | • Improve IR Visibility and Citeability  
  • Present program results at national and regional forums  
  • Collaborate with the Library’s Campus Stakeholders |
| **Diversify and increase the financial base** | • Apply for IMLS Grant  
  • Improve Press website visibility and reach |
Educate your staff on what 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)?
...and their 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”*
  - Google Scholar wants Highwire Press, PRISM, Be Press, Eprints metadata schema

* Google Scholar Inclusion Guidelines for Webmasters
Set up Google Webmaster Tools and ask questions

- Reduce Google Crawl Errors

- Improve Server Performance
Promote the “Right way” and set policy to prevent the wrong way for SEO

- Recent Black Hat news stories
  - JC Penney
  - Overstock
  - Google Chrome

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

- Brief discussion of two years of research
  - Priorities & Results
  - Issues & Opportunity
  - Key Steps & Questions
  - Google Scholar
Why does Google Scholar Matter?

- “researchers find Google and Google Scholar to be amazingly effective” and accept the results as “good enough in many cases” (Kroll & Forsman 2010)
- “broader awareness of specialized Google tools such as Google Scholar and Google Book among faculty members and graduate students” (Rieger 2009)
- “the amount of qualified scholarly content has increased considerably in Google Scholar since it was launched in 2004 (Mikki 2009)
- 4% - 27% use increase in four-year U Miss study (Herrera 2010)
USpace IR Google Index Ratios baseline

Google Index Ratio

- ETD 1: 12%
- ETD 2: 0%
- UScholar Works: 23%
- Board of Regents: 4%

*Weighted Average Google Index Ratio = 18.33% (1,188/6,482)
*Weighted Average Google Index Ratio = 18.33% (1,188/6,482)
Google wants the right metadata tags used consistently and accurately

"Use Dublin Core tags (e.g., DC.title) as a last resort—they work poorly for journal papers...”

- Google Scholar Inclusion Guidelines for Webmasters

...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
Low GS indexing ratios of “primary links” cut across institutions

Google Scholar Indexing Ratio for Selected Institutional and Disciplinary Repositories October 2011

Average = 30%
Ensure Google Scholar can easily find, crawl, and understand your content

<table>
<thead>
<tr>
<th>Indexing Ratio</th>
<th>Cornell</th>
<th>Oregon</th>
<th>Cal Tech</th>
<th>Texas A&amp;M Faculty</th>
<th>UW Aquatic Tech Reports</th>
<th>Columbia</th>
<th>Rochester</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>98%</td>
<td>88%</td>
<td>88%</td>
<td>48%</td>
<td>46%</td>
<td>45%</td>
<td>38%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Software</th>
<th>Digital Commons</th>
<th>DSpace</th>
<th>ePrints</th>
<th>DSpace</th>
<th>DSpace</th>
<th>Fedora/Blacklight</th>
<th>IR+</th>
</tr>
</thead>
</table>

| Titles Available/Captured | Unknown /1,421 | 4,067/1,463 | 24,146/1,306 | 763/757 | 563/539 | 3,819/1,432 | 1,562/926 |

<table>
<thead>
<tr>
<th>Crawling Guidelines</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Browse by Date</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Recently added</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>10 clicks from home page</td>
<td>Yes</td>
<td>Yes, Not in root</td>
<td>Yes</td>
<td>No, only first 200</td>
<td>No only first 200</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Robots.txt</td>
<td>Yes</td>
<td>Yes, Not in root</td>
<td>Yes</td>
<td>Yes, Disallows browse by date</td>
<td>Yes, Disallows browse by date</td>
<td>Yes, not configured</td>
<td>Yes</td>
</tr>
<tr>
<td>Sitemap Index</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indexing Guidelines</th>
<th>Meta Tag Schema in HTML headers</th>
<th>BePress</th>
<th>DC</th>
<th>ePrints &amp; DC</th>
<th>None</th>
<th>DC &amp; DCTERMS</th>
<th>None</th>
<th>None</th>
</tr>
</thead>
</table>

This study examines whether the changing social and economic characteristics of women who give birth out of wedlock have led to higher family incomes. Using Current Population Survey data collected between 1982 and 2002, we find that never-married mothers remain poor. They have made modest economic gains, but these have disproportionately occurred at the top of the income distribution. Yet there is no evidence of a burgeoning class of "Murphy Browns" middle-class professional women who give birth out of wedlock. Surprisingly, never-married mothers' incomes have stagnated in spite of impressive gains in education and other personal and vocational characteristics that should have resulted in greater economic progress than has been the case. These gains cast doubt on various stereotypes about women who give birth out of wedlock.
... parse each citation into HTML meta tags Google Scholar can read

```html
1  <meta name="citation_title" content="Thanks for nothing: changes in income and la"
2  <meta name="citation_author" content="Wolfinger, Nicholas H." />
3  <meta name="citation_author" content="McKeever, Matthew" />
4  <meta name="citation_date" content="2006-07-26" />
5  <meta name="citation_firstpage" content="1" />
6  <meta name="citation_lastpage" content="42" />
7  <meta name="citation_keywords" content="Motherhood; Single Mothers; Income; Popul;
8  <meta name="citation_technical_report_institution" content="Institute of Public &
9  <meta name="citation_technical_report_number" content="2006-07-04" />
10 <meta name="citation_language" content="en" />
11 <meta name="citation_conference_title" content="101st American Sociological Assoc:
12 <meta name="citation_pdf_url" content="http://cdm6gs.lib.utah.edu/utils/getfile/c"
First step was to begin aligning our Dublin Core fields with Highwire Press and ... 

First step was to begin aligning our Dublin Core fields with Highwire Press and ...

<table>
<thead>
<tr>
<th>Highwire Press Tags</th>
<th>Dublin Core Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>citation_author</td>
<td>DC.creator</td>
</tr>
<tr>
<td>citation_date</td>
<td>DC.issued</td>
</tr>
<tr>
<td>citation_title</td>
<td>DC.title</td>
</tr>
<tr>
<td>citation_publisher</td>
<td>DC.publisher</td>
</tr>
<tr>
<td>citation_journal_title</td>
<td>DC.relation.ispartof</td>
</tr>
<tr>
<td>citation_volume</td>
<td>DC.citation.volume</td>
</tr>
<tr>
<td>citation_issue</td>
<td>DC.citation.issue</td>
</tr>
<tr>
<td>citation_firstpage</td>
<td>DC.citation.spage</td>
</tr>
<tr>
<td>citation_lastpage</td>
<td>DC.citation.epage</td>
</tr>
<tr>
<td>citation_isssn</td>
<td>n/a</td>
</tr>
<tr>
<td>citation_isbn</td>
<td>n/a</td>
</tr>
<tr>
<td>citation_keywords</td>
<td>DC.subject</td>
</tr>
<tr>
<td>citation_dissertation_institution</td>
<td>DC.publisher</td>
</tr>
<tr>
<td>citation_technical_report_institution</td>
<td>DC.publisher</td>
</tr>
<tr>
<td>citation_technical_report_number</td>
<td>n/a</td>
</tr>
<tr>
<td>citation_language</td>
<td>DC.Language</td>
</tr>
<tr>
<td>citation_conference_title</td>
<td>DC.publisher</td>
</tr>
<tr>
<td>citation_pdf_url</td>
<td>DC.identifier</td>
</tr>
</tbody>
</table>
Google Scholar Pilot 1 tested importance of Metadata model

- 6,482 URLs in Sitemaps submitted via Google Webmaster Tools.
- Errors generated during Google crawls were analyzed and addressed.
- Updated & corrected metadata for 20 pilot articles
  - Ensured full-text PDF met GS inclusion guideline requirements.
  - Provided a “landing page” per GS inclusion guidelines, containing links to the 20 IR pilot papers that was within a few clicks of the home page.
USpace IR Google Index Ratios increased

*October 16, 2011 Weighted Average Google Index Ratio = 97.82% (10,306/10,536).
October 16, 2011 Weighted Average Google Index Ratio = 97.82% (10,306/10,536).
GS Pilot 2 Utilized OCLC’s relationship with Google Scholar

- 19 Papers in GS Pilot 2
  - 6 of 7 GS paper types represented
  - 19 Full Text PDFs

- Augmented CONTENTdm v.6
  - Highwire Press Meta tags
  - Browse By Year
  - Recently Added
  - College & Department
GS Pilot 2 Utilized OCLC’s relationship with Google Scholar

- 19 Papers in GS Pilot 2
- 6 of 7 GS paper types represented
- 19 Full Text PDFs
- Augmented CONTENTdm v.6
- Highwire press Meta tags
- Browse By Year
- Recently Added
- College & Department

Google Scholar Index Ratio

62%
GS Pilot 3 Increased Sample Size to 56

- 56 Papers in GS Pilot 3
  - USpace IR items not in GS Index
  - 6 of 7 GS paper types represented
- 100% Server Up-time
GS Pilot 3 Increased Sample Size to 56

- 56 Papers in GS Pilot 3
- Google Scholar Index Ratio
- 100% Server Uptime
- 6 of 7 GS paper types represented
- USpace items not in GS Index
# A Pre-Print Author Manuscript is not the Journal Article.

<table>
<thead>
<tr>
<th>Meta Tag</th>
<th>Pre-Print</th>
<th>Journal Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - citation_author</td>
<td>Maloney, Krisellen; Antelman, Kristin; Arlitsch, Kenning; Butler, John</td>
<td>Maloney, Krisellen; Antelman, Kristin; Arlitsch, Kenning; Butler, John</td>
</tr>
<tr>
<td>2 - citation_date</td>
<td>2009</td>
<td>2010</td>
</tr>
<tr>
<td>3 - citation_title</td>
<td>Future leaders' views on organizational culture</td>
<td>Future leaders' views on organizational culture</td>
</tr>
<tr>
<td>4 - citation_publisher</td>
<td>N/A</td>
<td>Association of College &amp; Research Libraries</td>
</tr>
<tr>
<td>5 - citation_journal_title</td>
<td>N/A</td>
<td>College and Research Libraries</td>
</tr>
<tr>
<td>6 - citation_volume</td>
<td>N/A</td>
<td>71</td>
</tr>
<tr>
<td>7 - citation_issue</td>
<td>N/A</td>
<td>4</td>
</tr>
<tr>
<td>8 - citation_firstpage</td>
<td>1</td>
<td>322</td>
</tr>
<tr>
<td>9 - citation_lastpage</td>
<td>56</td>
<td>347</td>
</tr>
<tr>
<td>10 - citation_doi</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>11 - citation_isiss</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>12 - citation_isbn</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>13 - citation_keywords</td>
<td>Organizational culture</td>
<td>Organizational culture</td>
</tr>
<tr>
<td>14 - citation_dissertation_institution</td>
<td>Uspace Institutional Repository, University of Utah</td>
<td>N/A</td>
</tr>
<tr>
<td>15 - citation_dissertation_name</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>16 - citation_technical_report_institution</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>17 - citation_technical_report_number</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>18 - citation_language</td>
<td>en</td>
<td>en</td>
</tr>
</tbody>
</table>

**Not Relevant**
- 14 - citation_dissertation_institution
- 15 - citation_dissertation_name
- 19 - citation_conference_title
- 20 - citation_inbook_title
IMLS Grant Deliverables – October 2014

- Expand research
- Publish Toolkit
  - SEO recommendations
  - Metadata transformation mechanisms
  - Tools for monitoring and reporting
- Disseminate findings
  - Papers
  - Conference presentations
  - Webinar training sessions
Summary – What You Can Do

- Establish baseline data
  - Configure Google Analytics
  - Set up Webmaster Tools
- Submit sitemaps/configure robots.txt file
- Monitor/address errors
- Inform staff/assign ownership
  - Find out what your staff know
- Clean up metadata
- Upgrade repository software
Questions?

Kenning Arlitsch
Associate Dean for IT Services
kenning.arlitsch@utah.edu

Patrick O'Brien
SEO Research Manager
Patrick.O'Brien@utah.edu