SEO to SWI (a progression)

Improving the visibility of libraries and their collections

Kenning Arlitsch, Dean of the Library

OCLC-RLP, London, UK
Theme

Achieving machine comprehension of library products, services, and the organizations, themselves

– SEO = Search Engine Optimization
– SWI = Semantic Web Identity
A Short Poll, Please

What is SEO?

Does your library have a formal SEO program?

Is your SEO program driven by administration?
Why SEO/SWI Matters

Institutional reputation

Student enrollment

Faculty citation rates

University rankings

Research funding
If a digital library is not indexed by Internet search engines then it will suffer from low visitation and use.

A digital library that is machine comprehensible is by definition also accessible to disabled users.
The dominance of search engines

• Americans submit 18 billion queries to SE each month*
  – 12 billion to Google sites (67%)
  – 3.5 billion to Microsoft sites (19%)
  – 1.8 billion to Yahoo! Sites (10%)

• Google market share in EU is 90+% **


2005 OCLC Study
Where College Students Begin Research

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

SEO Research Inspiration

• Decade building digital library @ University of Utah
  – Mountain West Digital Library
  – Utah Digital Newspapers
  – Western Waters Digital Library
  – Western Soundscape Archive

• Were they being used...?
Well, not really…

• University of Utah in 2010
  – Only 12% of digital collections were indexed by Google
  – 0.5% of Utah’s IR scholarly papers were indexed by Google Scholar

• Spot checks revealed similar problems in most academic libraries

Patrick OBrien
Basic SEO improved indexing ratio in Google...

Google Index Ratio - All Collections*

* Google Index Ratio = URLs submitted / URLs Indexed by Google
** ~150 collections containing ~170,000 URLs (07/2010) and ~170 collections containing ~282,000 URLs (12/2013)
...resulting in more referrals and visitors

12 week comparison 2010 vs. 2012

<table>
<thead>
<tr>
<th>Domain</th>
<th>Custom View: 2/6/12 4/29/12 Visits</th>
<th>Custom View: 2/1/10 4/25/10 Visits</th>
<th>% Change Visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>google.com</td>
<td>51,694</td>
<td>8,959</td>
<td>477.01%</td>
</tr>
<tr>
<td>google.co.uk</td>
<td>1,284</td>
<td>182</td>
<td>605.49%</td>
</tr>
<tr>
<td>google.ca</td>
<td>1,203</td>
<td>415</td>
<td>189.88%</td>
</tr>
<tr>
<td>google.it</td>
<td>670</td>
<td>38</td>
<td>1,063.16%</td>
</tr>
<tr>
<td>google.co.in</td>
<td>602</td>
<td>68</td>
<td>785.29%</td>
</tr>
<tr>
<td>google.fr</td>
<td>475</td>
<td>35</td>
<td>1,257.14%</td>
</tr>
<tr>
<td>google.es</td>
<td>466</td>
<td>26</td>
<td>1,692.31%</td>
</tr>
<tr>
<td>google.com.au</td>
<td>463</td>
<td>95</td>
<td>387.37%</td>
</tr>
<tr>
<td>google.de</td>
<td>441</td>
<td>88</td>
<td>401.14%</td>
</tr>
<tr>
<td>google.com.br</td>
<td>408</td>
<td>29</td>
<td>1300.93%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>63,637</strong></td>
<td><strong>10,559</strong></td>
<td><strong>502.68%</strong></td>
</tr>
</tbody>
</table>

Increase: 502.68%
...and significant increases in the average number of page views per day.
THE PROBLEM OF INSTITUTIONAL REPOSITORIES
Almost 100% USpace IR content indexed in Google

*October 16, 2011 Weighted Average Google Index Ratio = 97.82% (10,306/10,536).
Structured data SE can identify, parse and digest

**Human Readable Citation**

SEO Deficiencies in Libraries

BOTH organizational and technical issues

SEO Deficiency Themes (Organizational)

• Administration and Strategy:
  – SEO is rarely driven from the top of the organization; usually considered a technical issue and is left to IT with little consideration of strategy, goals or reporting.

• Communication:
  – Administrators don’t communicate the reasons for an SEO program and its impact to the rest of the organization. Communication among the staff involved in SEO programs can also be poor.

• Ineffective Analytics Reporting:
  – Web Analytics software is often incorrectly configured, diminishing the ability to report use of a digital library or monitor the effects of change to the repository.
SEO Deficiency Themes (Technical)

• Website Design
  – Graphics
  – Confusing site hierarchies and paths

• Poor experience for search engine customers
  – Slow servers
  – Incorrect use of redirects

• CMS/DAM often lack canonical links

• Metadata
  – Schema not understood by search engines
  – Not unique
  – Inconsistent/inaccurate
SEO Building Block Priority

1. **Increase Reach**
   - Get objects indexed by search engines

2. **Increase Visibility in SERP**
   - Provide robust descriptive content

3. **Get Relevant**
   - Increase click-through rates (CTR)
“Getting Found” SEO Cookbook

• Measure digital library performance
  – Monitor and diagnose problems

• Structure
  – Five video tutorials
  – Supporting documentation
    • Case studies
    • Deep dives

• http://www.clir.org/pubs/reports/pub165
Getting Found: SEO Cookbook

by Patrick O'Brien and Kenning Arlitsch

May 2015
CLIR pub 165

This is a web-only report—it is not available in print

At a time when Internet search engines have become the default discovery layer for most users, libraries need to report that their websites and digital repositories are discoverable through those search engines as well. The Getting Found (GF) Cookbook provides a step-by-step video guide to help libraries measure and monitor the search engine optimization (SEO) performance of their digital repositories. The Cookbook includes everything necessary to implement a preconfigured Google Analytics dashboard that continuously monitors SEO performance metrics relevant to digital repositories.

The Cookbook was supported by a grant from the Institute of Museum and Library Services.

Phase I: Institutionalizing SEO
1. Justifying the Need—Video
2. Strategic Planning—Video
   ▪ Review Case Study: MSU Library Strategic Planning
   ▪ Read Recommended References

Phase II: Prep Work
1. Prepping for Inventory—Video
   ▪ Identify Stakeholder Roles and Conduct Survey of Web Properties
   ▪ Use Web Metrics and Analytics—Survey Template
   ▪ Review Analytics Stakeholders—Graphic Overview
We are gathering and consolidating information about the library’s website homepages and accounts used to gather web metrics or communicate with users through social media outlets. This information is needed to improve the library’s Website(s) search engine optimization (SEO) and web analytics. Please provide as much information as you can about the entire library’s web presence.

1. What is the Library’s public homepage web address? (e.g. www.lib.montana.edu)

2. List ALL the public “home page(s)” for accessing your library’s services, content, and information via the Internet. Please separate responses containing multiple URLs with commas. See link for “homepage examples” or click on the section titles below.

   - Digital Collections
   - Institutional Repository (IR)
   - Library Services

3. Does your institution have any other domains or sub-domains not listed above? (e.g. scholarworks.montana.edu or etc.live.montana.edu) If so, please list them below.

4. Please list ALL the web address(es) for your library’s social media sites? (Please cut and paste the addresses and separate multiple URLs with commas).

   - Facebook
   - Google+
   - Twitter
   - LinkedIn
   - YouTube
   - Pinterest
   - Instagram
   - Flickr
   - Tumblr

Contact Information

List the name, title, and email address for the person currently performing each of the roles below. See roles description for more information. Please provide the single most pertinent contact for each role.

8. IT Manager Role
   - Name
   - Title
   - Email Address

9. System Administrator Role
   - Name
   - Title
   - Email Address

10. Applications Administrator Role
    - Name
    - Title
    - Email Address

11. Google Analytics Administrator Role
    - Name
    - Title
    - Email Address

12. Administration Contact Role (e.g. Dean of the Library)
    - Name
    - Title
    - Email Address
Monitor SE crawlers
Search engine comprehension (of organizations)

SEMANTIC WEB IDENTITY (SWI)
SWI Research Began in 2012
From which sources does Google gather facts?
Semantic Web

• Latest extension of the WWW (Web 3.0)
  – “Data and information that can be processed automatically”*
  – “Computers must have access to structured collection of information”*
  – Every entity must have a Universal Resource Identifier (address)

• Implications for search engines
  – Change: matching strings of text to matching entities and their relationships
  – Promise more accurate and relevant results
  – Answers rather than lists of websites where answers might reside

Semantic Web Identity (SWI)

• The condition in which Internet search engines recognize the existence and nature of entities
  – Important for semantic technologies such as mapping applications
• Characterized as a search engine having gathered enough verifiable facts about an entity for a formal display of that entity in SERP
• Display of KC is an indicator of SWI
Research Goals

• What is the state of SWI among Association of Research Libraries (ARL) members (and other academic organizations)?

• Can SWI be established or improved by engaging with certain knowledge bases on the Semantic Web?
RESEARCH METHODS
ARL: The problem of Names

• 125 ARL member libraries
• Every library has a primary (official) name
• 94 libraries also have alternate names

• Example:
  – Yale University Library = primary name
  – Sterling Memorial Library = alternate name
Data Collection

• Searched Google for evidence of KC for ARL libraries
  – 125 primary names of libraries
  – 94 alternate names of libraries
  – Total names = 219

• Searched 5 knowledge bases for evidence of records for all 219 names
  – Google My Business
  – Google+ (recorded verified/unverified profiles)
  – Wikipedia (recorded articles w/infobox, w/o infobox)
  – DBpedia
  – Wikidata
Data Analysis

• 1,400+ screen capture files
• Excel spreadsheet
• R statistical computing software for analysis

• Data set available in MSU ScholarWorks repository
  – http://doi.org/10.15788/M2F590
Case Studies

• Documented intervention to improve SWI of three organizations
  – Montana State University Library
  – McMaster University Library
  – Coalition for Networked Information (CNI)
RQ1: What is the current state of SWI of ARL libraries, as indicated by the presence of accurate KC in Google search results when the primary and alternate names of those libraries are searched?

FINDINGS
Table plot showing that ARL library alternate names (column 1, orange rows) were more likely to display an accurate KC (column 2, green rows)
Table plot showing that 82% of ARL libraries displayed an accurate KC (Column 1, yellow rows), but that many of the KC were not the same for the primary and alternate names of the libraries (Column 2, purple rows).
RQ2: Are records or profiles present for ARL primary and alternate library names in the five knowledge bases?

FINDINGS
Records for ARL members in Knowledge Bases

<table>
<thead>
<tr>
<th>Knowledge Base</th>
<th>Primary (% of 125)</th>
<th>Alternate (% of 94)</th>
<th>Total (% of 219)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google My Business</td>
<td>22%</td>
<td>43%</td>
<td>31%</td>
</tr>
<tr>
<td>Google Plus (verified)</td>
<td>18%</td>
<td>20%</td>
<td>19%</td>
</tr>
<tr>
<td>Wikipedia (w/infobox)</td>
<td>24%</td>
<td>28%</td>
<td>26%</td>
</tr>
<tr>
<td>DBpedia</td>
<td>24%</td>
<td>41%</td>
<td>32%</td>
</tr>
<tr>
<td>Wikidata</td>
<td>21%</td>
<td>39%</td>
<td>29%</td>
</tr>
</tbody>
</table>
RQ2, Sub-question 1

Is an accurate KC likely to display in search results if the library organization has not been claimed and verified in Google My Business?

Table plot showing libraries that had claimed and verified their businesses in GMB (column 1, dark red rows) were more likely to display accurate KC (column 2, green rows)
RQ2, Sub-question 2:

Is a KC likely to display a description field if no Wikipedia article exists for the primary or alternate name of the library?

Table plot showing that Wikipedia articles (column 1, yellow rows) tend to result in descriptions (column 2, pink rows) on accurate KC (column 3, green rows).
FINDINGS – GOOGLE+
Basic SWI Process

- Claim and verify business in GMB
  - Generates Google+ profile

- Write and publish article in Wikipedia
  - Generates DBpedia record
  - Generates Wikidata record

- Populate Wikidata record
<table>
<thead>
<tr>
<th>Type of Organization</th>
<th>Description</th>
<th>Website/Directions buttons</th>
<th>Address, Hours, Phone</th>
<th>Logo/Image and Map</th>
</tr>
</thead>
</table>

**Montana State University Library**

The Montana State University Library is the academic library of Montana State University, Montana's land-grant university, in Bozeman, Montana, United States. It is the flagship library for all of the Montana State University System's campuses. [Wikipedia](https://en.wikipedia.org/wiki/Montana_State_University_Library)

**Address:** Renne Library, Bozeman, MT 59717

**Hours:** Open today - 10AM–5PM

**Phone:** (406) 994-3139

**Founded:** 1934

**Staff:** 56

**Parent organization:** Montana State University
Results of SWI Study

• Lack of records in knowledge bases corresponds to lack of KC
• Proposed process to improve SWI is successful
• Library-related concepts and entities are poorly defined on SW
<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbo:thumbnail</td>
<td>wik-commons:Special:FilePath/Melk_-<em>Abbey</em>-_Library.jpg?width=300</td>
</tr>
</tbody>
</table>
  http://www.gutenberg.org/etext/15327  
  http://www.hse.ed.ac.uk/chb/  
  http://www.ibiblio.org/librariesfaq/ |
| dbo:wikiPageID | 17727 [xsd:integer] |
| dbo:wikiPageRevisionID | 708089041 [xsd:integer] |

**dict:subject**
- dbo:Book_promotion
- dbo:Libraries
- dbo:Library_science

**rdf:type**
- owl:Thing
  - yago:Area102735688
  - yago:Artifact100021939
  - yago:Library103660909
  - yago:Object100002684
  - yago:PhysicalEntity100001930
  - yago:Room1041105683
  - yago:Structure104341686
  - yago:Whole100003553
  - yago:YagoGeoEntity
  - yago:YagoPermanentlyLocatedEntity
  - yago:WikicatResearchLibraries
CONCLUSIONS
Librarians struggle with SEO/SWI

- Lack of awareness/formal strategies
- Traditional outreach/marketing practices don't work on the SW
  - Inconsistent use of names
  - Lack of explicit “same as” declarations for machine comprehension
- Disregard/disdain for Semantic Web opportunities
  - Are not proactively creating and maintaining records in knowledge bases
  - Absence of formal strategy results in independent efforts by employees
    - Creates confusion for search engines
Broader Impact

– More than the tactical approach
– Opportunities
  • Develop cohesive marketing strategies and consistent processes
  • Expand skill sets of library faculty and staff
  • **Offer SWI services to campus constituents**
References


