SEO for Digital Repositories

Kenning Arlitsch & Patrick OBrien

April 5, 2011

CNI Spring Forum, San Diego, CA

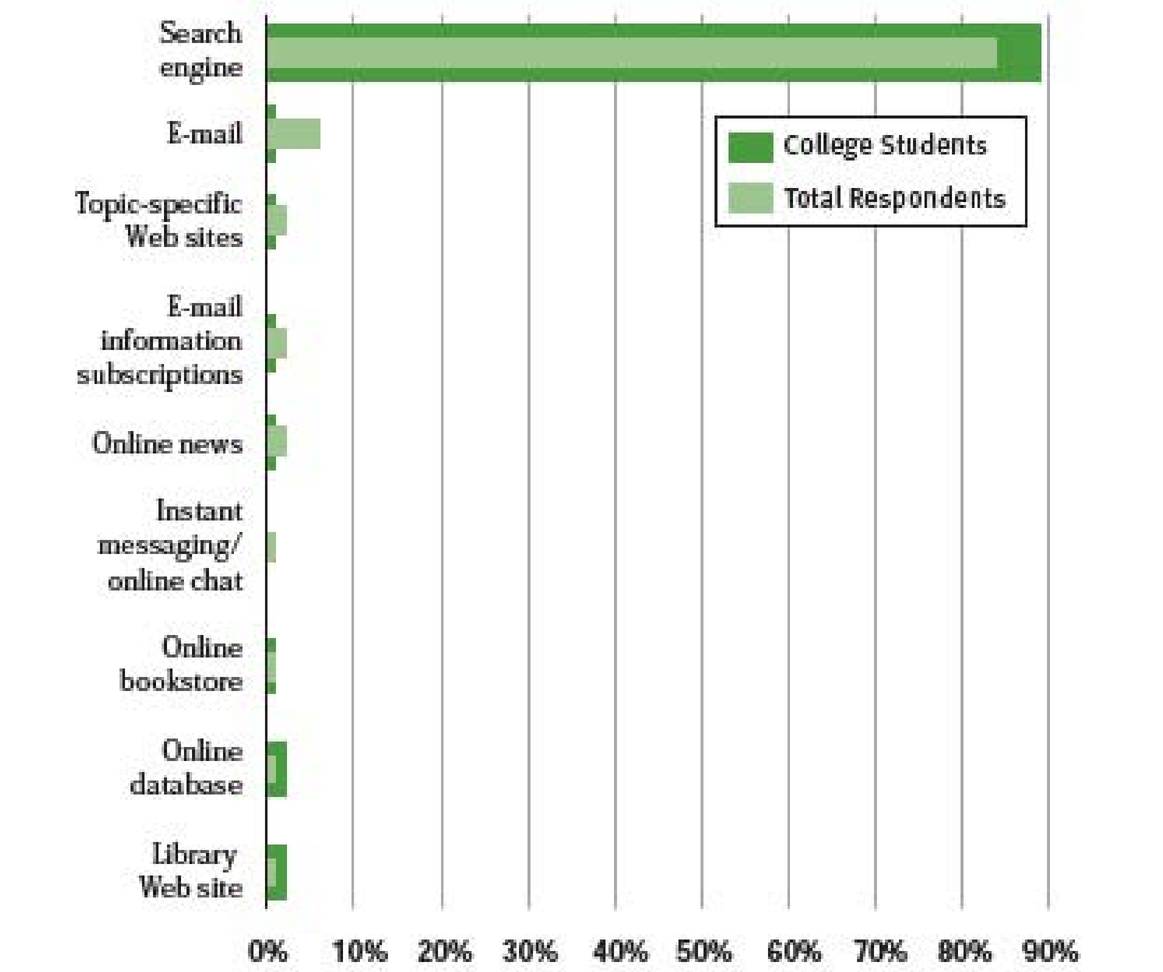
Today's Objectives

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

SEO Repository Goals

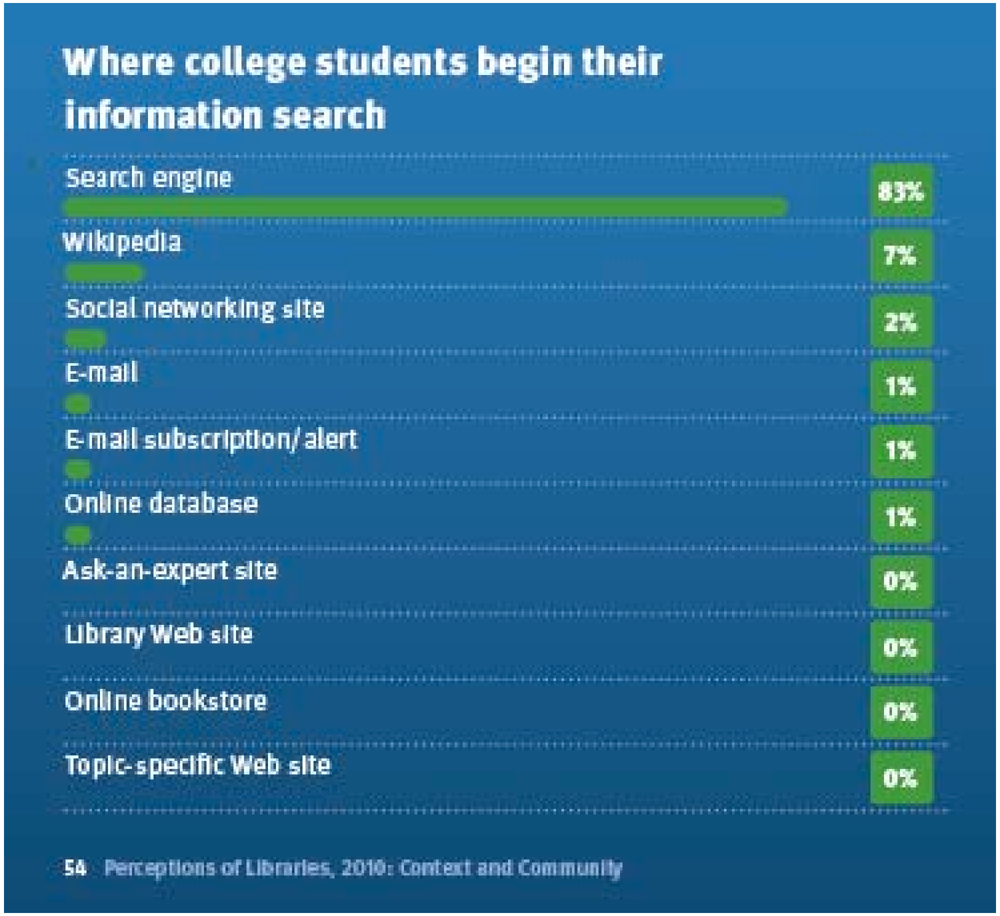
- Digital repositories vs general websites
  - Millions of objects in databases
  - Include IR
- Goal 1 – Increase Reach
  - Get objects indexed in search engines
- Goal 2 – Increase Visibility
  - Provide robust descriptive content

College Students Begin Research - 2005



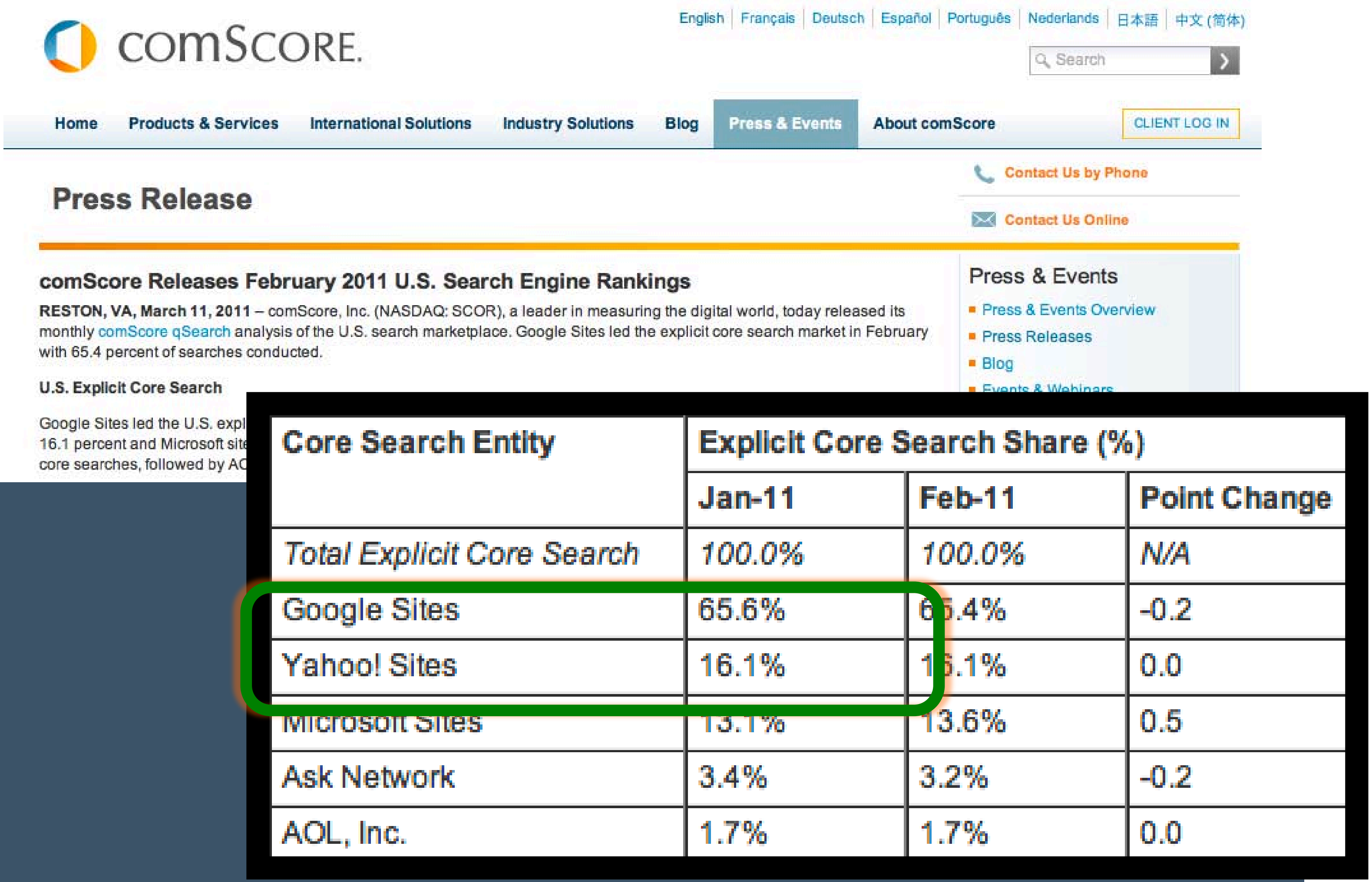
Source: *Perceptions of Libraries and Information Resources*, OCLC, 2005, question 520.
Note: Only electronic resources with usage rates of 1 percent or more are represented on this graph.

College Students Begin Research - 2010



DeRosa, Cathy, et al. “Perceptions of Libraries, 2010: Context and Community: A Report to the OCLC Membership”, OCLC, 2010.

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

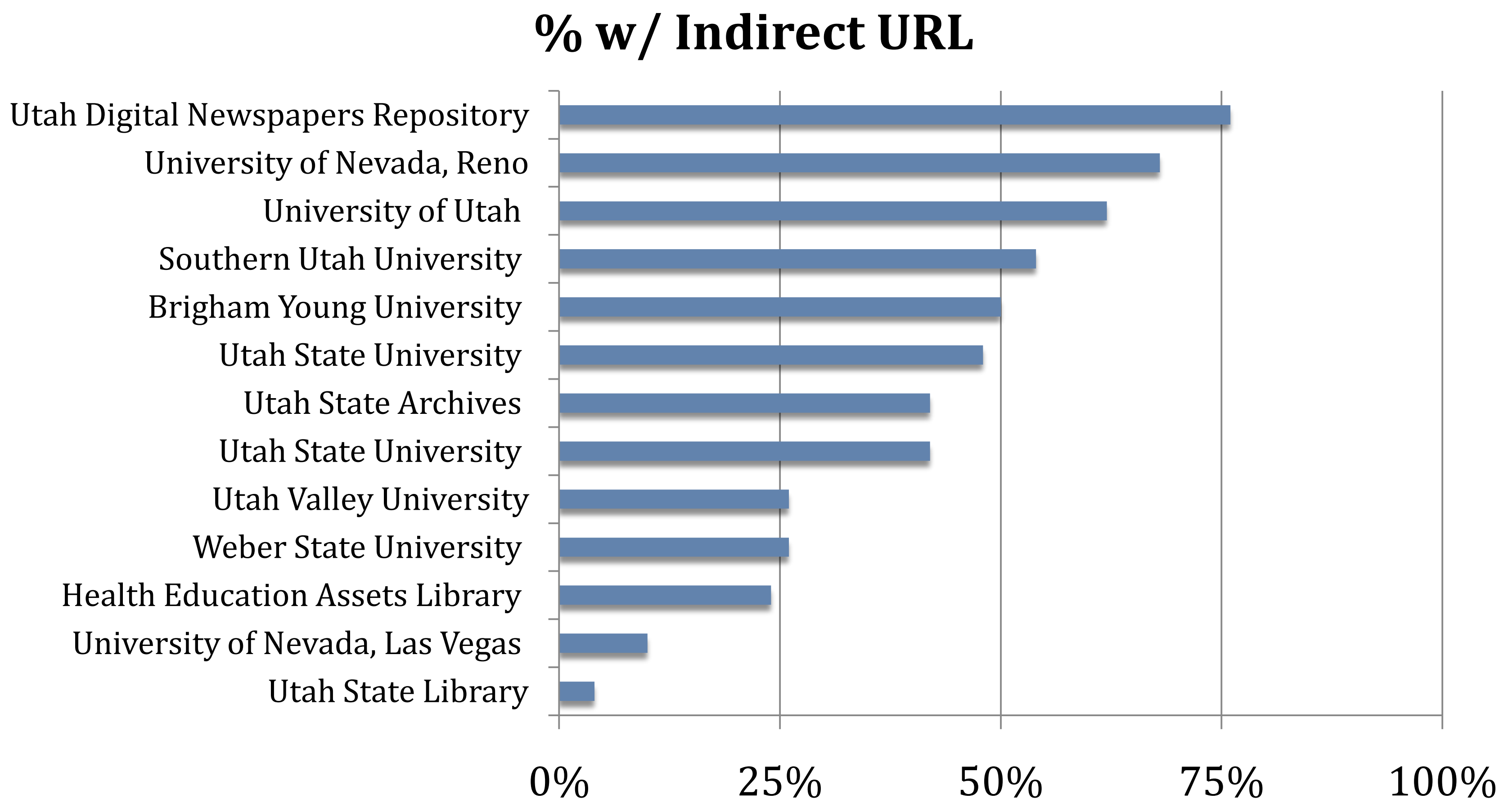
Google Sites led the U.S. expl
16.1 percent and Microsoft site
core searches, followed by AO

Core Search Entity	Explicit Core Search Share (%)		
	Jan-11	Feb-11	Point Change
Total Explicit Core Search	*100.0%*	*100.0%*	*N/A*
Google Sites	65.6%	65.4%	-0.2
Yahoo! Sites	16.1%	16.1%	0.0
Microsoft Sites	13.1%	13.6%	0.5
Ask Network	3.4%	3.2%	-0.2
AOL, Inc.	1.7%	1.7%	0.0

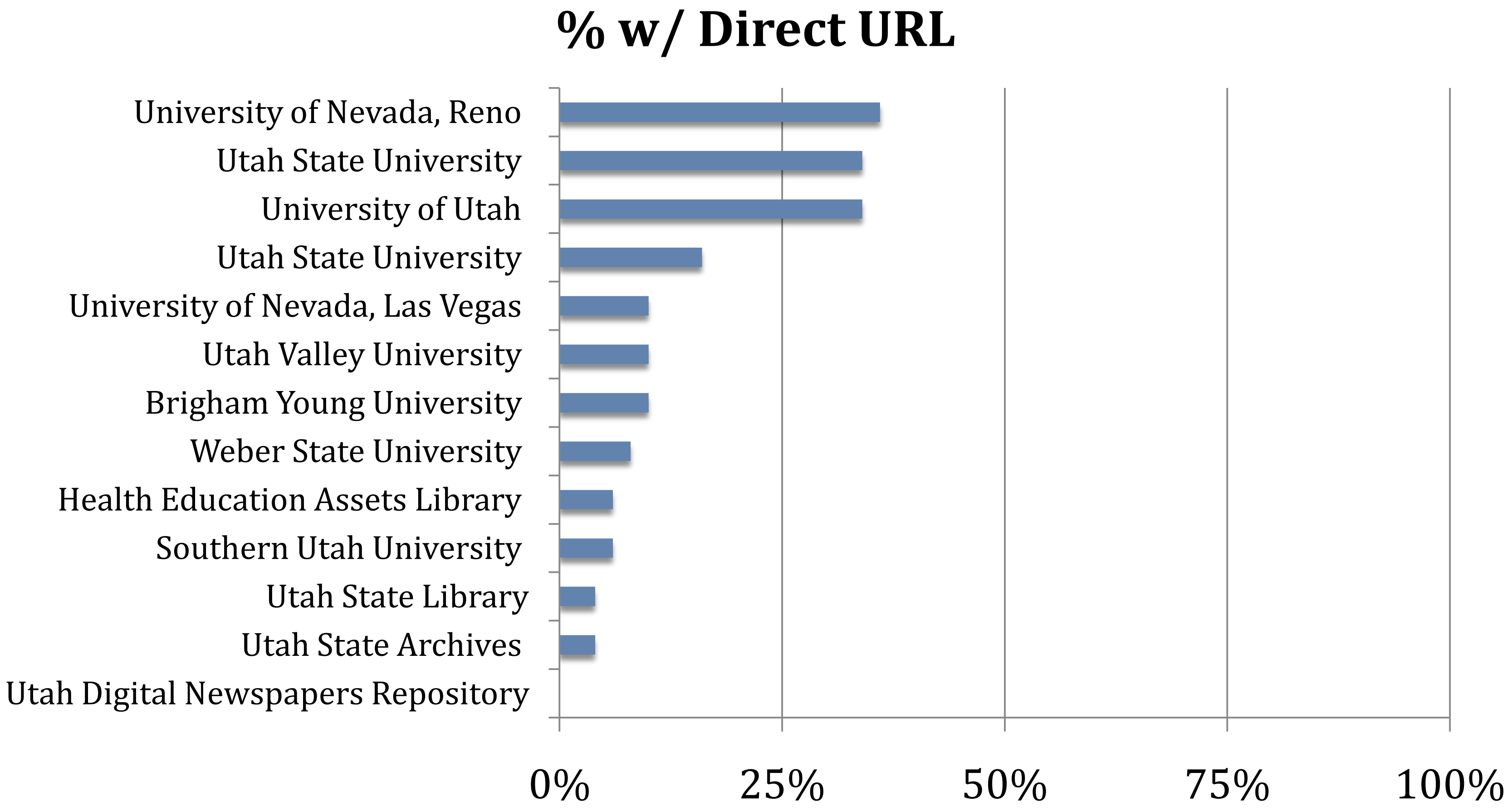
Management Experiences

- 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



MWDL Repositories Survey

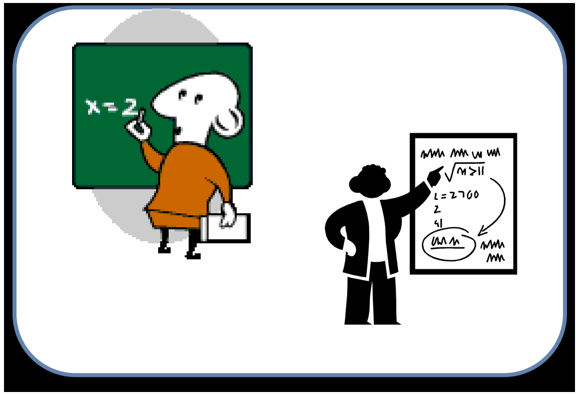


Literature Lessons

- 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



Value

High

- □ Publication Page Views
- □ Publication Downloads
- □ Requests for Information
- □ Publication Citations

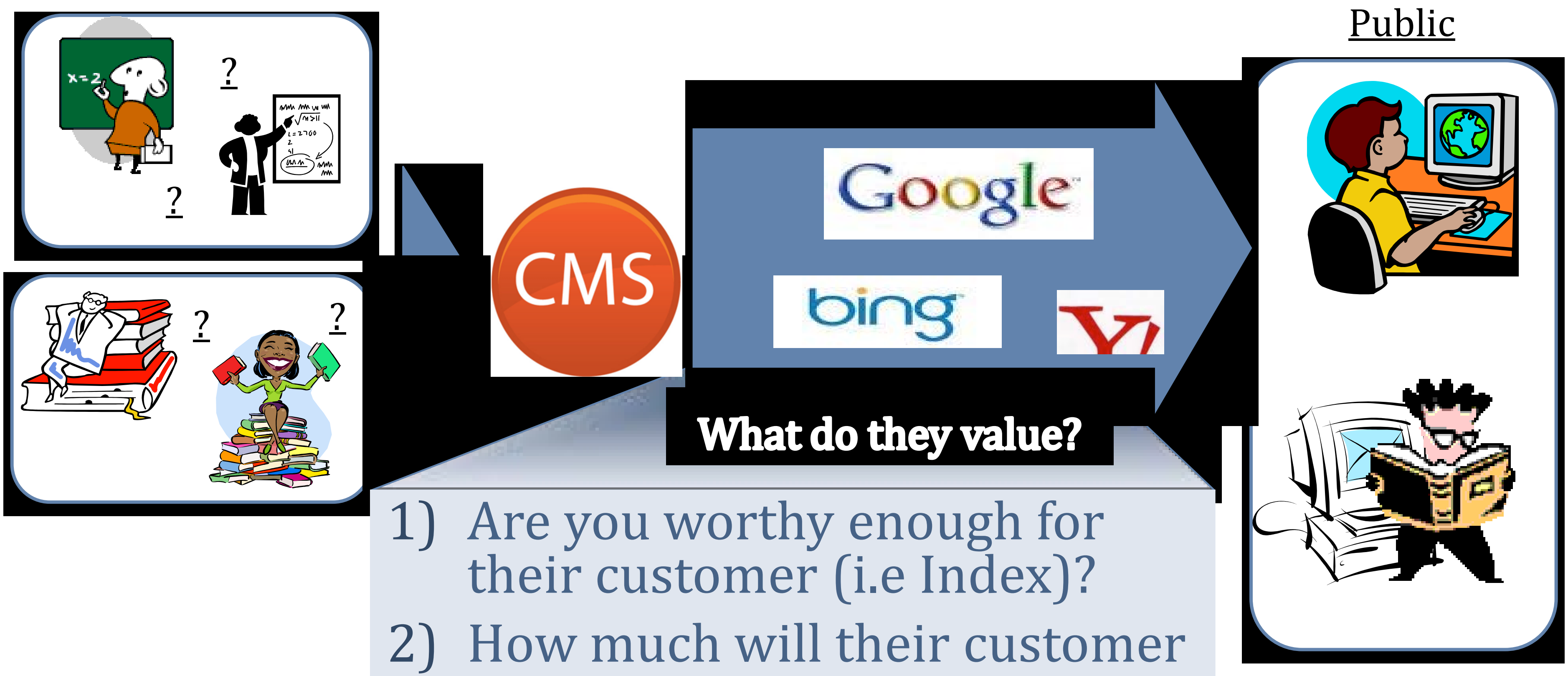
Collection Donors

Value

High

- □ 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?



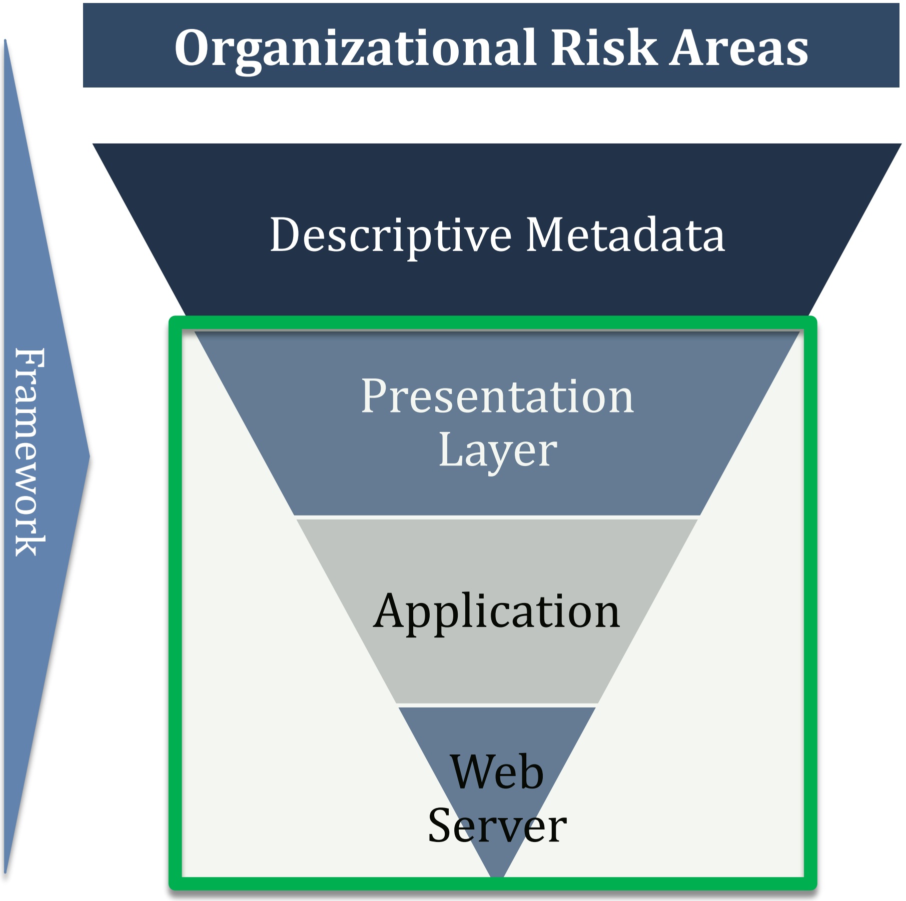
1) Are you worthy enough for their customer (i.e Index)?
2) How much will their customer value the introduction (i.e, Visibility)?

Relate risk to organizational functions

Major Barriers

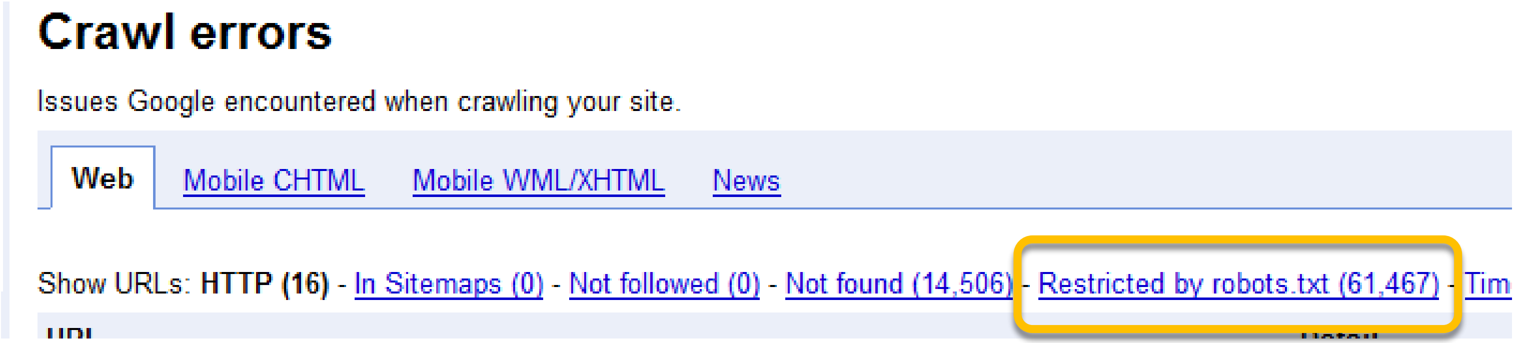
- □ Administrative/Organizational issues
- □ Descriptive metadata uniqueness and structure
- □ Search engines policies and practices
- □ Server configuration and performance

Organizational Risk Areas



Setup Google Webmaster Tools and ask questions.

- □ Reduce Google Crawl Errors

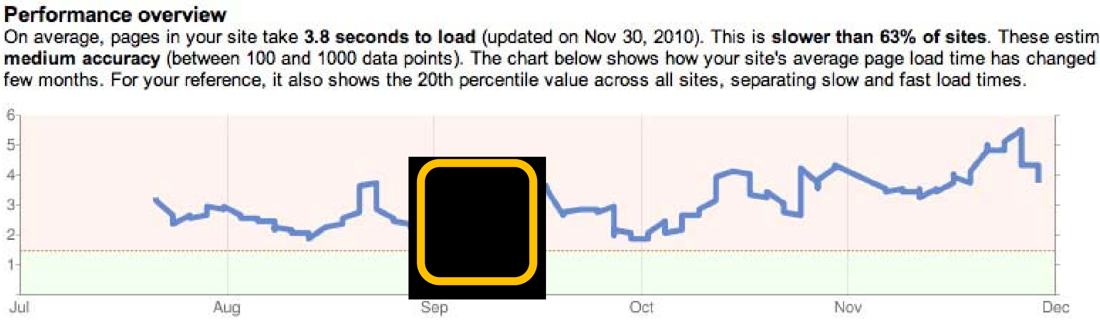


Crawl errors

Issues Google encountered when crawling your site.

Show URLs: **HTTP (16)** - In Sitemaps (0) - Not followed (0) - Not found (14,506) - Restricted by robots.txt (61,467) - Tim

- □ Improve Server Performance



Performance overview

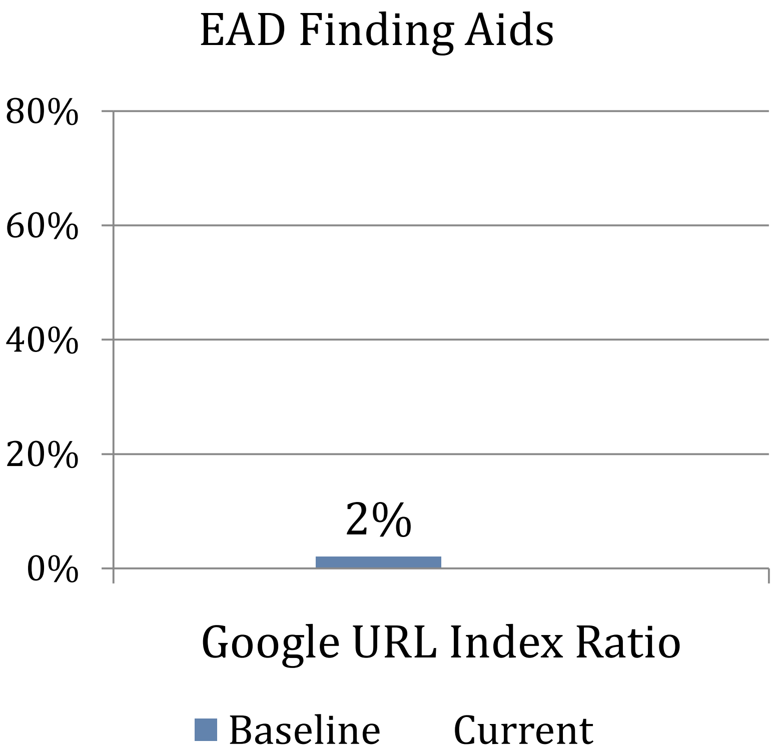
On average, pages in your site take **3.8 seconds to load** (updated on Nov 30, 2010). This is **slower than 63% of sites**. These estimates are of **medium accuracy** (between 100 and 1000 data points). The chart below shows how your site's average page load time has changed over the last few months. For your reference, it also shows the 20th percentile value across all sites, separating slow and fast load times.

Set goals and establish a baseline …

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



75 pages indexed / 3,221 pages submitted as of April 24, 2010

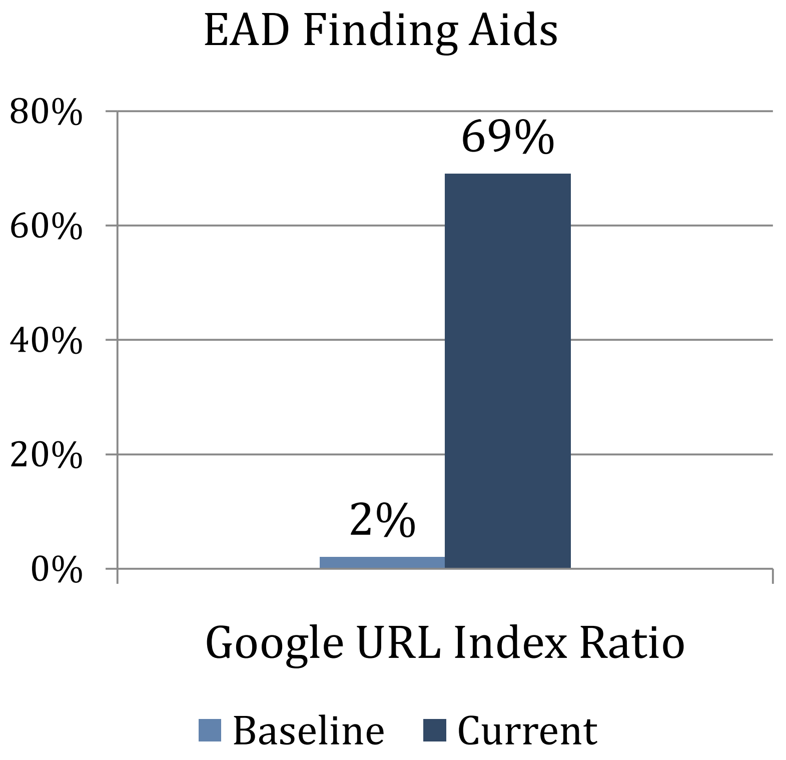
… with objective performance criteria.

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

Pilots

Results

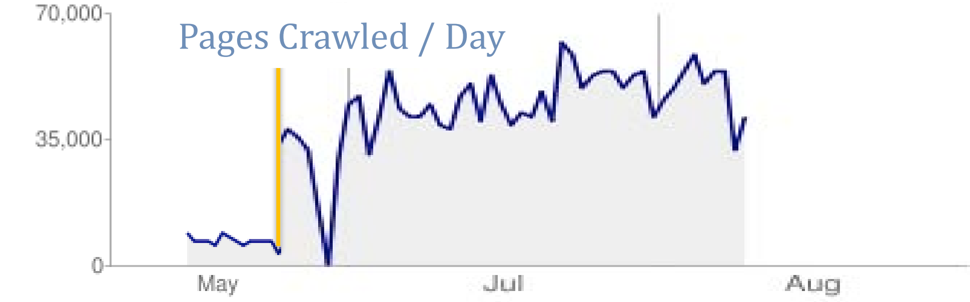


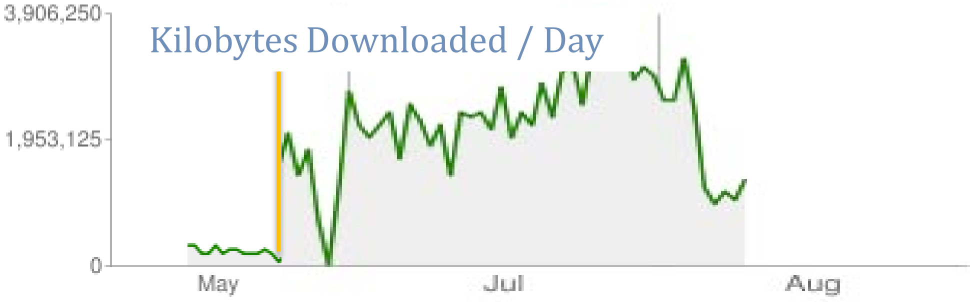
2,239 pages indexed / 3,235 pages submitted as of January 14, 2010

Goal 1: Initial focus was to make it easier for Google to index.

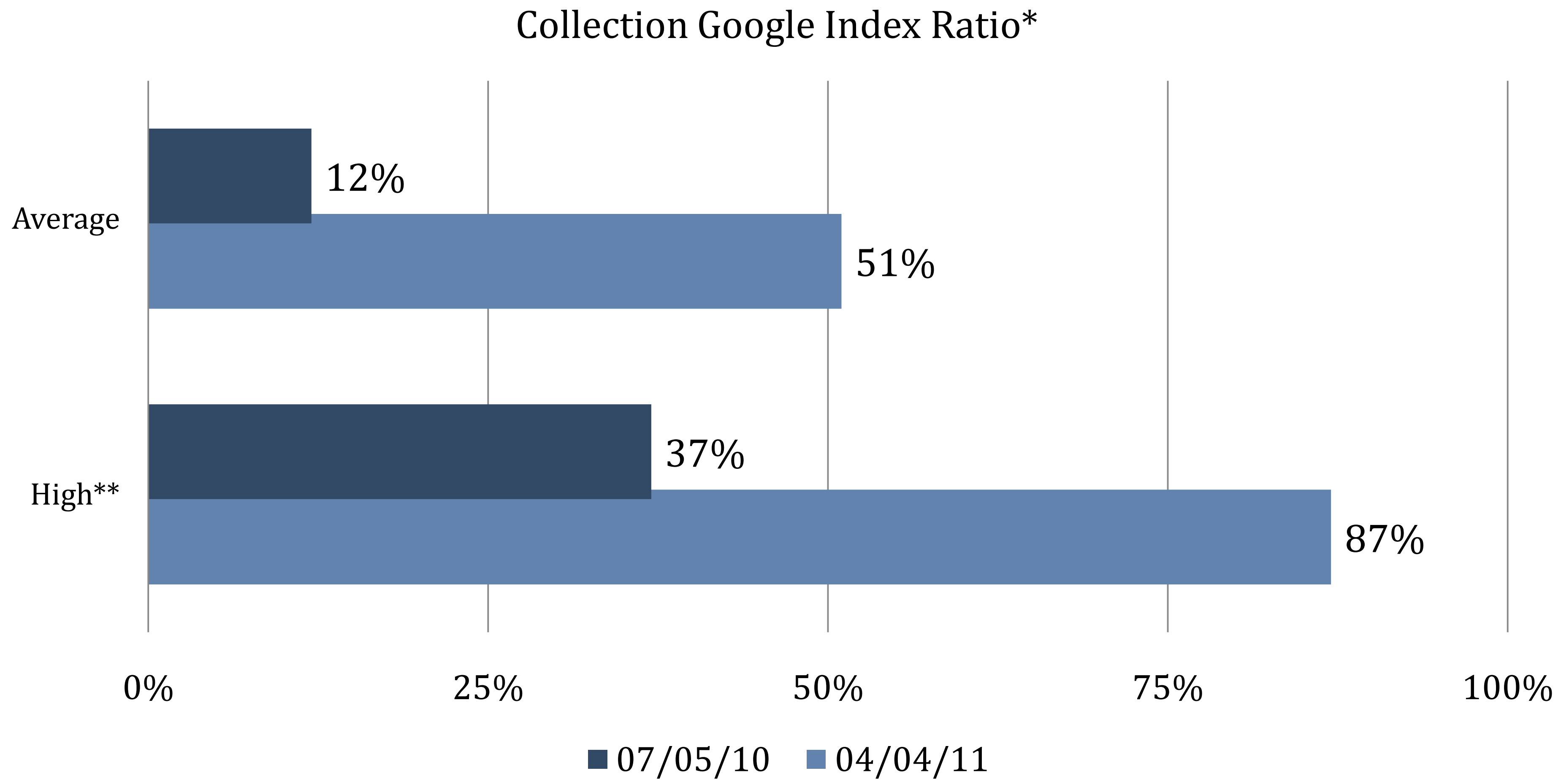
Initial Priorities

- □ Reduce Google Crawl Errors
- □ Developed efficient Google Crawler path
- □ Reconfigure the environment to meet Google's key requirements





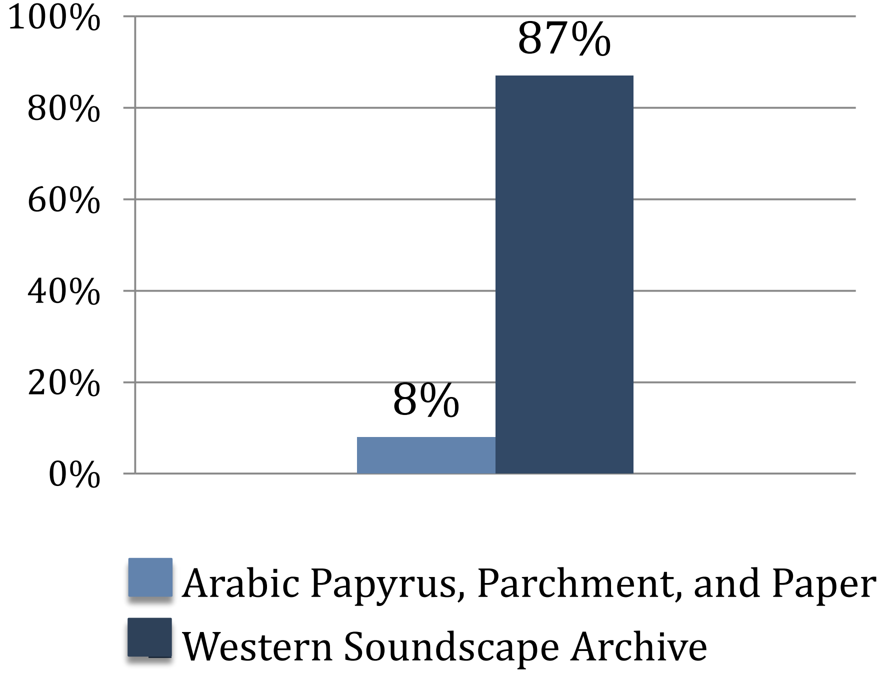
Collection Google Index Ratios have increased across the board.



* 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

Google Index Ratio



Drivers

Key Differences

- □ 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	Marriott Goal	Marriott Activity
Exploit the Digital and Networked Environments	Digitize Collection and share in many venues where users go	• 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
http://scholar.google.com/intl/en/scholar/inclusion.html

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