### **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 <u>Increase Reach</u>
  - Get objects indexed in search engines
- ◆ Priority 2 <u>Increase Visibility</u>
  - Provide robust descriptive content

#### **Literature Lessons**

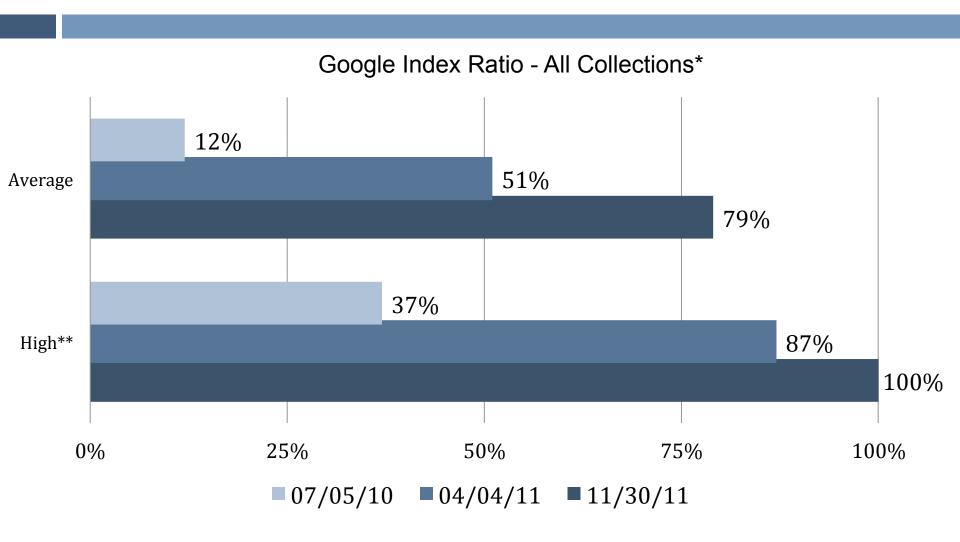
- Most are dated
- Most deal with general websites
- ◆ Few deal with digital collections in db's
- ◆ Some suggest duplicating the content outside the database

### **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

<sup>\*</sup> For full text of "<u>Authors' response Re: Google Scholar discoverability of repository content</u>" 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...



<sup>\*</sup> Google Index Ratio = URLs submitted / URLs Indexed by Google for about 150 collections containing ~170,00 URLs

<sup>\*\*</sup>Highest index ratio achieved for Collections with over 500 URLs submitted to Google

## ...increasing Google referrals by 200% and total visitors by 79%.

content.lib.utah.edu 12 week year-over-year

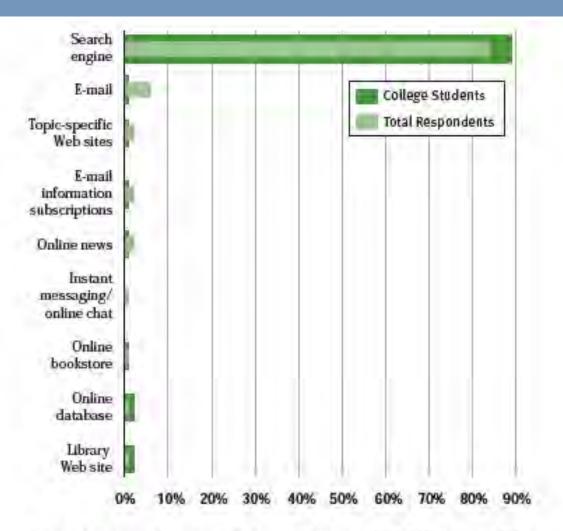
Init	tial Referrers				
Initial Referrer		Custom View: 2/6/11 - 4/30/11 Visits	Custom View: 2/7/10 - 5/1/10 Visits	% Change Visits	
1.	google.com	28,841	9,564	201.56%	
2.	utah.edu	26,440	14,783	78.85%	
3.	Direct Traffic	22,810	16,010	42.47%	
4.	bing.com	3,574	4,376	-18.33%	

	Custom View: 2/6/11 - 4/30/11	Custom View: 2/7/10 - 5/1/10	% Change	
Visitors	89,124	49,774	79.06%	
Visitors Who Visited Once	82,310	44,553	84.75%	
Visitors Who Visited More Than Once	6,814	5,221	30.51%	

### Agenda

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

### **College Students Begin Research - 2005**



Source: Perceptions of Libraries and Information Resources, OCLC, 2005, question 520.

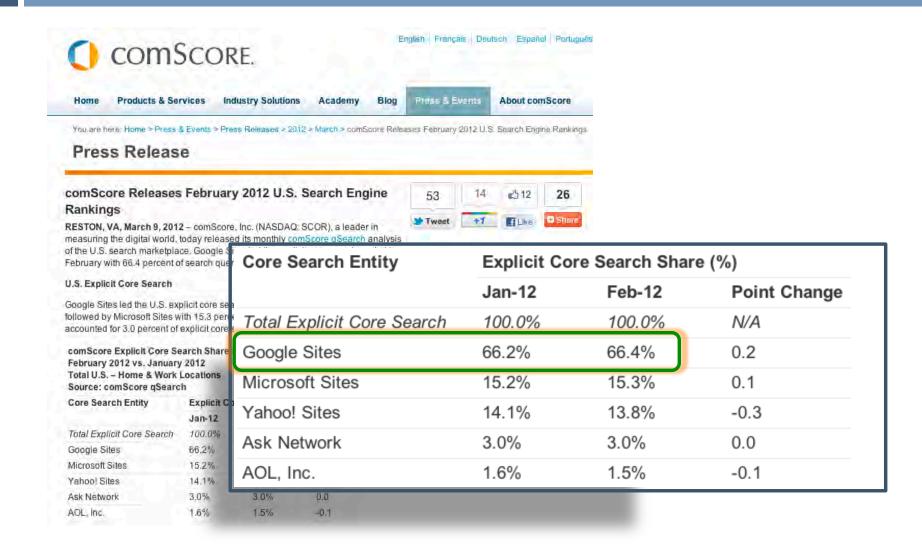
Note: Only electronic resources with usage rates of 1 percent or more are represented on this graph.

### College Students Begin Research - 2010

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%

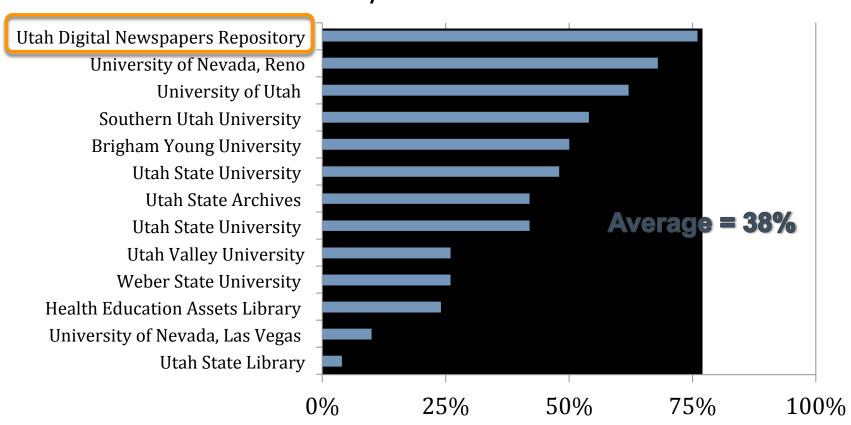
DeRosa, Cathy, et al. "Perceptions of Libraries, 2010: Context and Community: A Report to the OCLC Membership", OCLC, 2010.

### Start with the 800 pound gorilla



### **MWDL** Repositories Survey

#### % w/ Indirect URL



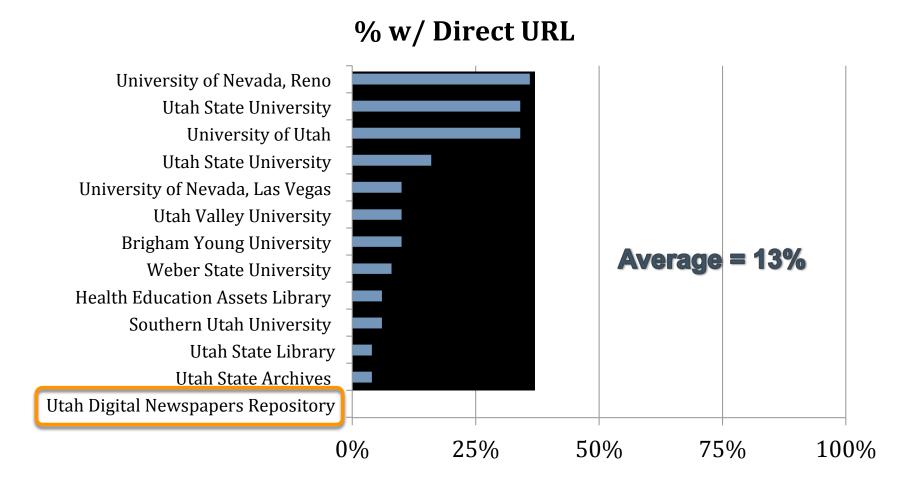
### **Example of indirect URL**



# Users and search engines want links directly to relevant content



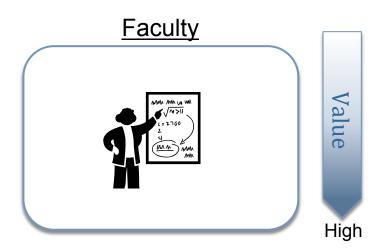
### **MWDL** Repositories Survey



### Agenda

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

#### Know your stakeholders and what they value



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

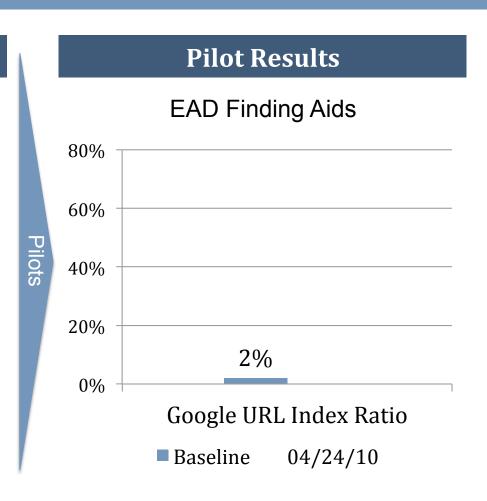


- ◆ Digital Collection Pages Indexed
- Digital Collection Page Views
- Digital Collection Visitors
- Requests for More Info
- Physical Collection Visitors
- Reproductions Ordered

## Set repository goals and establish a baseline ...

#### Goals

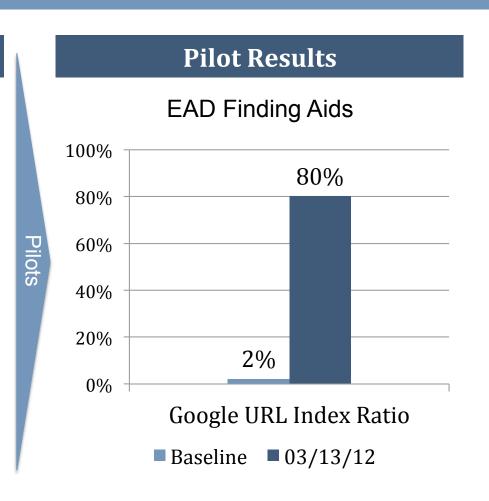
- Increase the number of Digital Collection web pages in the Google search engine.
- Develop internal library staff skills
- Develop a program to maximize a collections visibility and reach



#### ... 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



# Ensure your staff understand the strategic importance of your SEO efforts

#### **Library Strategic Plan**

Exploit the Digital and Networked Environments

#### **SEO Program Activities**

- Optimize Collections to Improve Visibility
  - Descriptions
  - Link Popularity
  - Page Elements
  - Metadata
  - Calls to Action
- Develop Metrics Dashboard to monitor and improve efforts

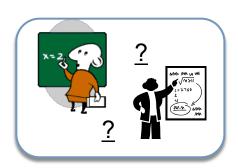
Elevate our position and impact on campus

- 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

### 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

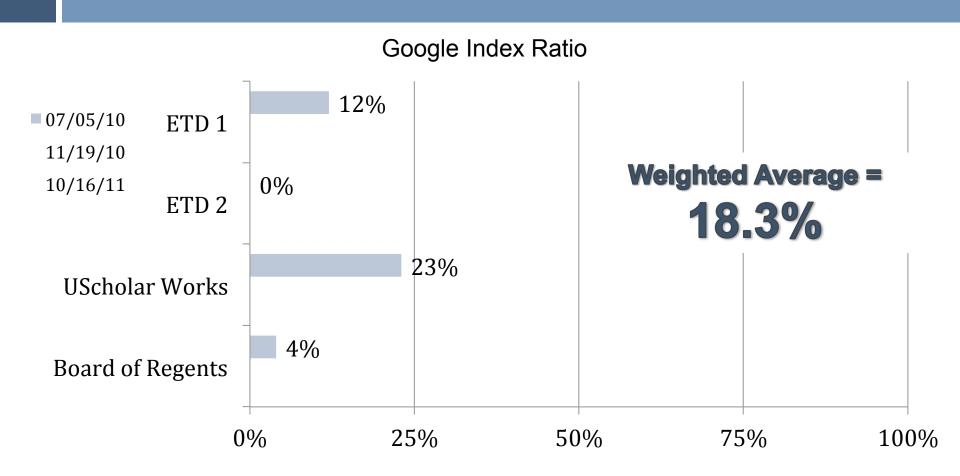
### Agenda

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

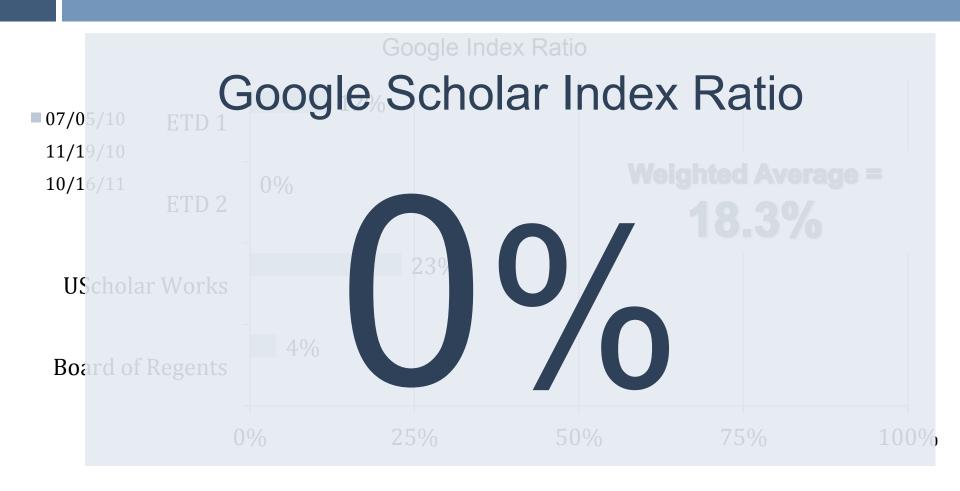
### Why does Google Scholar Matter?

- "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**



### **USpace IR Google Index Ratios baseline**

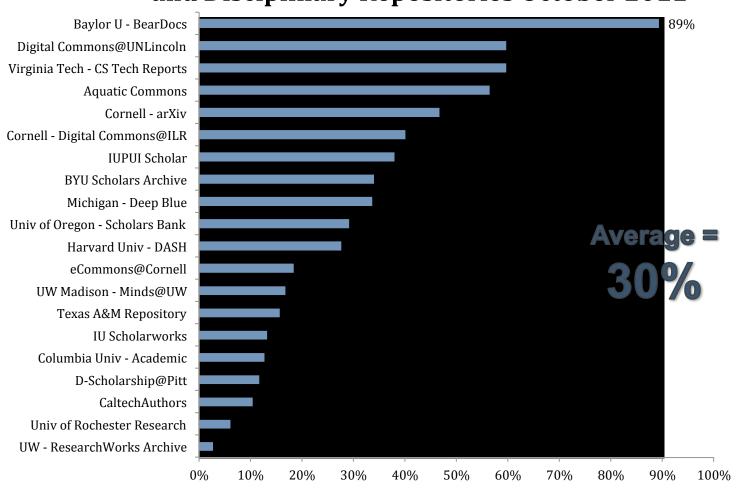


# Google wants the right metadata tags used consistently and accurately

- "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 <u>papers aren't included at all</u>, because documents with the <u>same title are often considered</u> <u>duplicates</u>.
  - Google Scholar Inclusion Guidelines for Webmasters
- "... <u>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."</u>
  - Google Scholar Inclusion Guidelines for Webmasters
- "...the most common cause of indexing problems is <u>incorrect extraction of bibliographic data</u> 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



# Ensure Google Scholar can easily find, crawl, and understand your content

	Cornell	Oregon	Cal Tech	Texas A&M Faculty	UW Aquatic Tech Reports	Columbia	Rochester
Indexing Ratio	98%	88%	88%	48%	46%	45%	38%
Software	Digital Commons	DSpace	ePrints	DSpace	DSpace	Fedora/ Blacklight	IR+
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
Crawling Guidelines							
Browse by Date	No	Yes	Yes	Yes	Yes	No	No
Recently added	No	No	Yes	No	No	No	No
10 clicks from home page	Yes	Yes	Yes	No, only first 200	No only first 200	Yes	No
Robots.txt	Yes	Yes, Not in root	Yes	Yes, Disallows browse by date	Yes, Disallows browse by date	Yes, not configured	Yes
Sitemap Index	Yes	No	No	Yes, not compliant with standards	No	No	No
Indexing Guidelines							
Meta Tag Schema in HTML headers	BePress	DC	ePrints & DC	None	DC & DCTERMS	None	None

# Challenge is presenting bibliographic citations GS can identify, parse and digest

Title University of Utah creator Other Creator Subject.Keyword Subject.LCSH

Description

Publisher Date.Original Type

Format.Extent
Format.Medium
Resource Identifier

Language

Relation

Research Institute
Department

School / College Contributing Institution Publication Type Thanks for nothing: changes in income and labor force participation for never-married mothers since 19 Wolfinger, Nicholas H.

McKeever, Matthew

Motherhood; Single Mothers; Income; Population surveys;

Single mothers

<u>Income</u>

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 dispropriately

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.

University of Utah

2006-07-26

Text

370,155 Bytes application/pdf ir-main,824

ena

Institute of Public and International Affairs Working Papers

McKeever, M. & Wolfinger, N.H. (2006). Thanks for Nothing: Changes in Income and Labor Force Particip

Never-Married Mothers since 1982. Institute of Public & International Affairs (IPIA), 4, 1-43.

(c) Matthew McKeever and Nicholas H. Wolfinger Institute of Public and International Affairs (IPIA)

Family & Consumer Studies

Sociology

College of Social & Behavioral Science

<u>University of Utah</u> <u>working paper</u>

## ... parse each citation into HTML meta tags Google Scholar can read

McKeever, M. & Wolfinger, N.H. (2006). Thanks for Nothing: Changes in Income and Labor Force Participation for Never-Married Mothers since 1982. Institute of Public & International Affairs (IPIA), 4, 1-43.

## ... parse each citation into HTML meta tags Google Scholar can read

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

McKeever, M. & Wolfinger, N.H. (2006). Thanks for Nothing: Changes in Income and Labor Force Participation for Never-Married Mothers since 1982. Institute of Public & International Affairs (IPIA), 4, 1-43.

#### **Dublin Core Tags** DC.creator DC.issued DC.title DC.publisher DC.relation.ispartof DC.citation.volume DC.citation.issue DC.citation.spage DC.citation.epage n/a n/a DC.subject DC.publisher DC.publisher n/a DC.Language DC.publisher DC.identifier

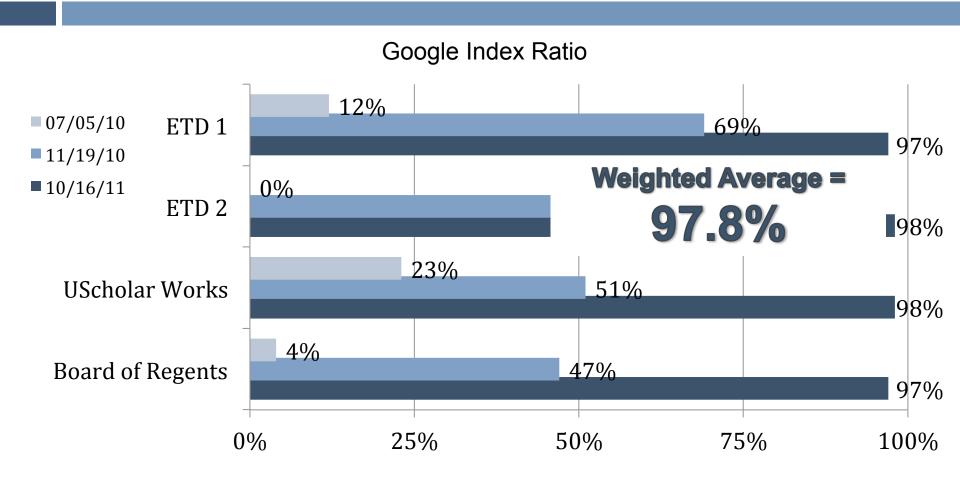
# First step was to begin aligning our Dublin Core fields with Highwire Press and ...

Highwire Press Tags	Dublin Core Tags
citation_author	DC.creator
citation_date	DC.issued
citation_title	DC.title
citation_publisher	DC.publisher
citation_journal_title	DC.relation.ispartof
citation_volume	DC.citation.volume
citation_issue	DC.citation.issue
citation_firstpage	DC.citation.spage
citation_lastpage	DC.citation.epage
citation_issn	n/a
citation_isbn	n/a
citation_keywords	DC.subject
citation_dissertation_institution	DC.publisher
citation_technical_report_institution	DC.publisher
citation_technical_report_number	n/a
citation_language	DC.Language
citation_conference_title	DC.publisher
citation_pdf_url	DC.identifier

### Google Scholar Pilot 1 tested importance of Metadata model

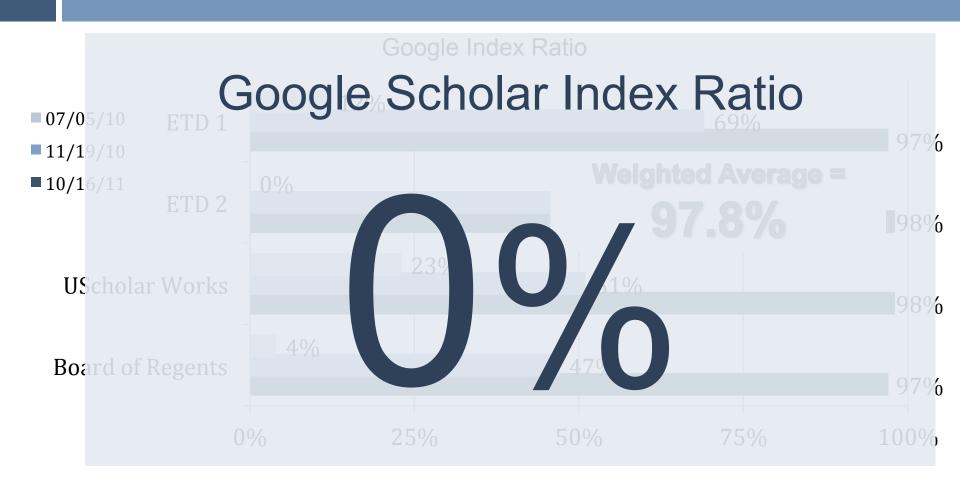
- ◆ 6,482 URLs in Sitemaps submitted via Google Webmaster Tools.
- ◆ Errors generated during Google crawls were analyzed and addressed.
- ◆ Updated & corrected metadata for 20 pilot articles
  - Ensured full-text PDF met GS inclusion guideline requirements.
  - Provided a "landing page" per GS inclusion 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**



<sup>\*</sup>October 16, 2011 Weighted Average Google Index Ratio = 97.82% (10,306/10,536).

### **USpace IR Google Index Ratios increased**



<sup>\*</sup>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 7Google Scholar Index Ratio
 19 Full Text PDFs

 Augmented Content and v.6

 Highwire tree Meta tags
 Browse Bluear
 Recently A Med

College & Department

#### **GS Pilot 3 Increased Sample Size to 56**

- ◆ 56 Papers in GS Pilot 3
  - \* <u>USpace IR</u> 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
 USpaGoogle Scholar Index Ratio
 6 of 7 GS paper types represented
 100% Serve Serve

## A Pre-Print Author Manuscript is not the Journal Article.

Meta Tag	Pre-Print	Journal Article
1 - citation_author	Maloney, Krisellen; Antelman, Kristin; Arlitsch, Kenning; Butler, John	Maloney, Krisellen; Antelman, Kristin; Arlitsch, Kenning; Butler, John
2 - citation date	2009	2010
3 - citation_title	Future leaders' views on organizational	Future leaders' views on organizational culture
	culture	
4 - citation_publisher	N/A	Association of College & Research Libraries
5 - citation journal title	N/A	College and Research Libraries
6 - citation_volume		71
7 - citation issue		4
8 - citation_firstpage	1	322
9 - citation_lastpage	56	347
10 - citation_doi		
11 - citation_issn		
12 - citation_isbn		
13 - citation_keywords	Organizational culture	Organizational culture
16 - citation_technical_report_institution	Uspace Institutional Repository, University of Utah	N/A
17 - citation_technical_report_number		N/A
18 - citation_language	en	en
21 - citation_pdf_url	http://cdm6gs.lib.utah.edu/utils/getfile/collection/uspace/id/10/filename/3.pdf	http://cdm6gs.lib.utah.edu/utils/getfile/collection/ uspace/id/16/filename/17.pdf
22 - citation_abstract_html_url  Not Relevant	http://cdm6gs.lib.utah.edu/cdm/singleitem/ collection/uspace/id/10/rec/1	http://cdm6gs.lib.utah.edu/cdm/singleitem/ collection/uspace/id/16/rec/2

14 - citation\_dissertation\_institution15 - citation\_dissertation\_name19 - citation\_conference\_title20 - 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 OBrien
SEO Research Manager
Patrick.OBrien@utah.edu