Confederation of Open Access Repositories

Driving Traffic to Institutional Repositories

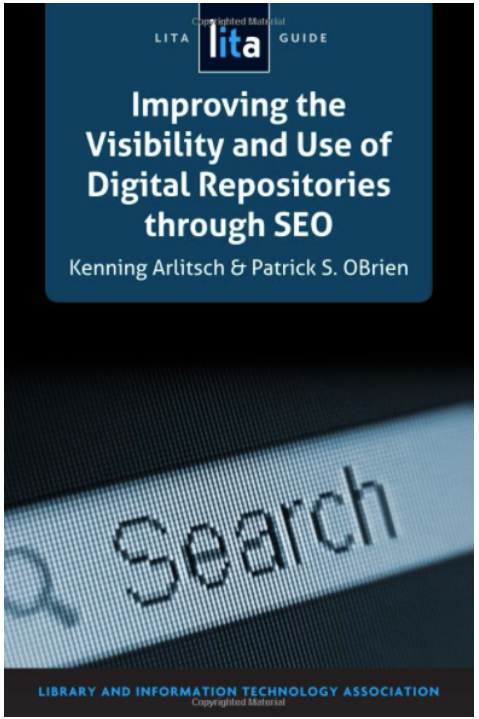
How Search Engine Optimization can increase the number of downloads from IR

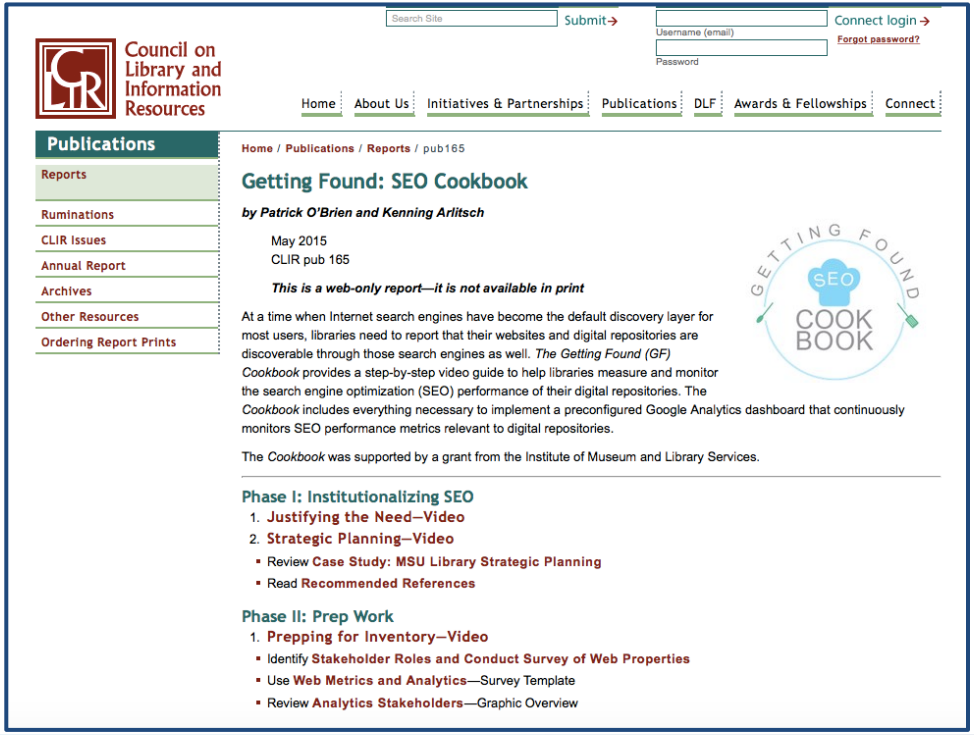
Kenning Arlitsch, PhD, MLIS
Dean
Montana State University Library
@kenning_msu

Presentation Agenda

- Search Engine Optimization
  - Evidence of success
  - IR and Google Scholar
- SEO Deficiencies
  - Organizational
  - Technical
- Analytics and Reporting
  - Google Search Console
  - Google Analytics
  - RAMP (Repository Analytics & Metrics Portal)

Two Resources for the Nuts and Bolts of SEO





Search Site Submit→ Connect login→ Username (email) Password Forgot password?

Council on Library and Information Resources

Home | About Us | Initiatives & Partnerships | Publications | DLF | Awards & Fellowships | Connect

Publications

- Reports
- Ruminations
- CLIR Issues
- Annual Report
- Archives
- Other Resources
- Ordering Report Prints

Home / Publications / 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

High-level overview

SEARCH ENGINE OPTIMIZATION

Repository not optimized for SE? → Low visitation and use

Optimized for SE? → Accessible to disabled users

SEO Research History and Inspiration

- 1999-2012, led digital library team @ University of Utah
  - Digitized more than 1 million newspaper pages
  - >500,000 objects of other formats
  - Major projects:
    - Mountain West Digital Library
    - Utah Digital Newspapers
    - Western Waters Digital Library
    - Western Soundscape Archive
    - USpace institutional repository

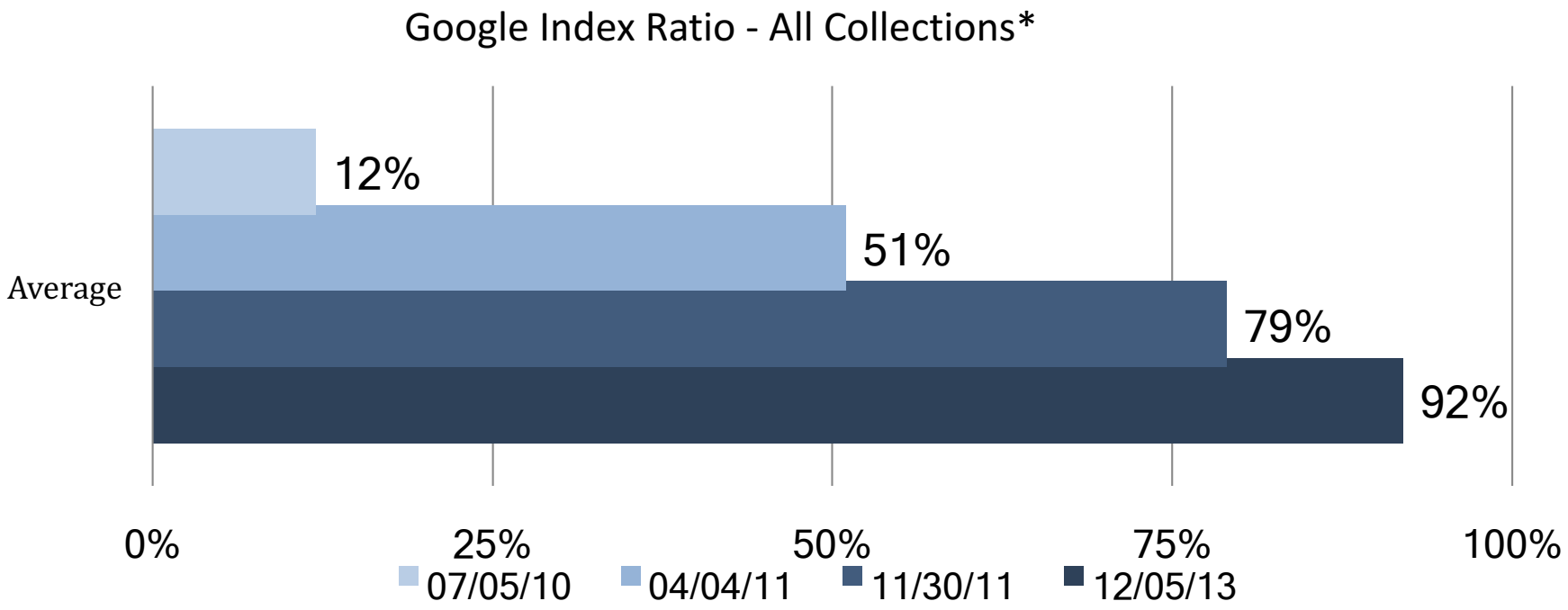
- Were they being used…?

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
- Surveys revealed similar problems in most academic libraries

Patrick OBrien

Basic SEO improved indexing ratio in Google…



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

** ~150 collections containing ~170,00 URLs (07/2010) and ~170 collections containing ~282,000 URLs (12/2013)

…resulting in more referrals and visitors

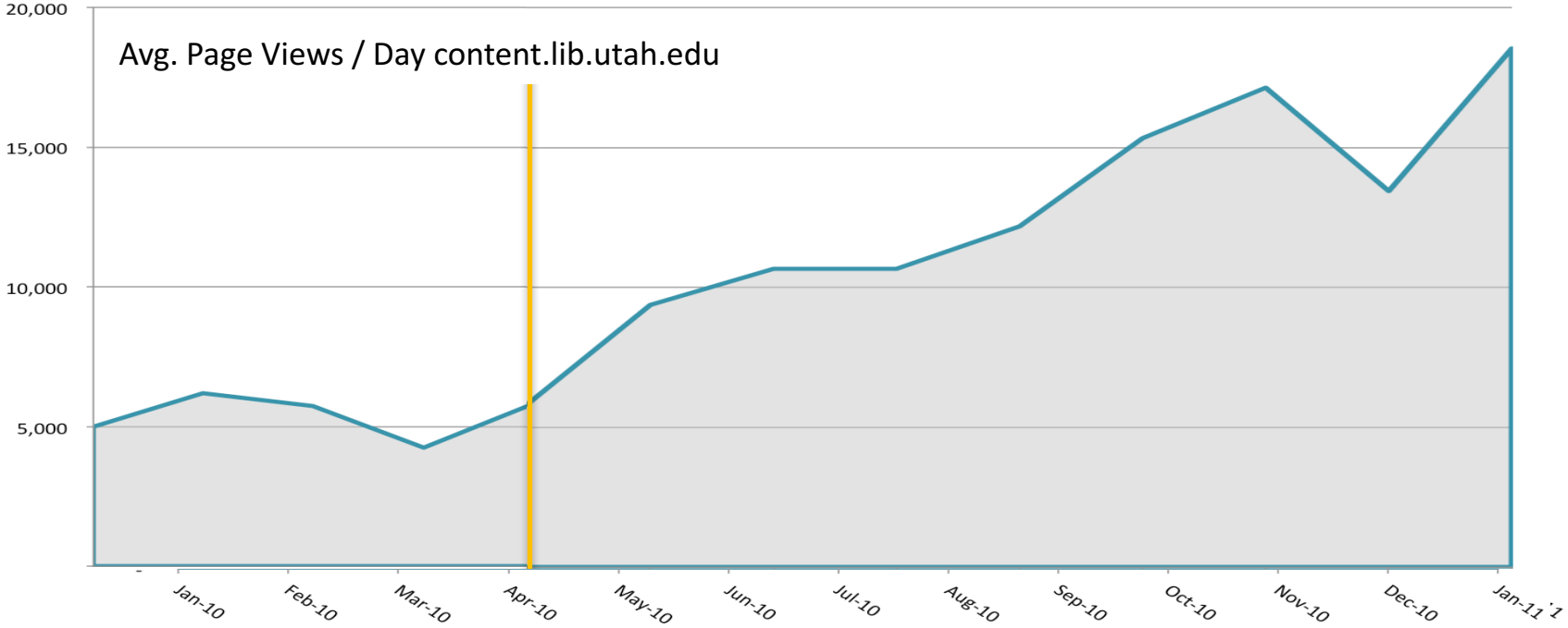
12 week comparison 2010 vs. 2012

Visitor Summary

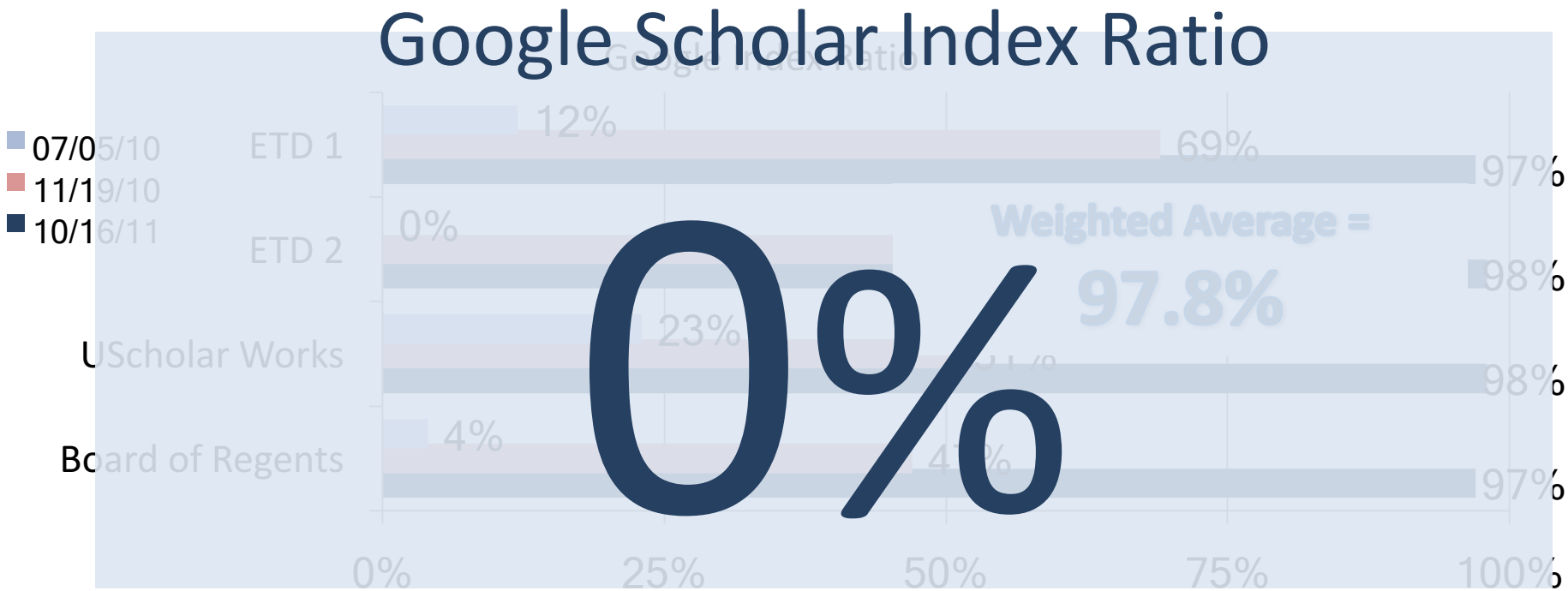
	Custom View: 2/6/12 - 4/29/12	Custom View: 2/1/10 - 4/25/10	% Change
Visitors	115,500	48,962	135.90%
Visitors Who Visited Once	107,749	43,841	145.77%
Visitors Who Visited More Than Once	7,751	5,121	51.36%
Average Visits per Visitor	1.13	1.24	-8.87%

4. google.it	670	38	1,663.16%
5. google.co.in	602	68	785.29%
6. google.fr	475	35	1,257.14%
7. google.es	466	26	1,692.31%
8. google.com.au	463	95	387.37%
9. google.de	441	88	401.14%
10. google.com.br	408	29	1,306.90%
Total	**63,637**	**10,559**	Increase **502.68%**

…and significant increases in the average number of page views per day.



Almost 100% USpace IR content indexed in Google



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

Special SEO problems

INSTITUTIONAL REPOSITORIES AND GOOGLE SCHOLAR

The IR Audience

- People
  - Scholars/researchers/lay public
- Search engines
  - Specifically, Google Scholar (see next slide)
- Whose audience?
  - Search engine users are not your users until SE refers them to you
  - SE will not refer if not confident of a good user experience
    - (see deficiency themes)

What's So Special About Google Scholar?

- May be the most highly-used general academic SE
- Delivers a high-value audience to IR
  - Researchers seeking scholarly publications
- Sub-organization of Google
  - Different index, different harvesting/indexing methods
- 48%-66% of IR traffic referred by GS*
  - If an IR is properly optimized for GS's requirements

*Patrick Obrien, Kenning Arlitsch, Leila Sterman, Jeff Mixter, Jonathan Wheeler & Susan Borda (2016) "Undercounting File Downloads from Institutional Repositories," *Journal of Library Administration*, 56:7, 854-874, DOI: 10.1080/01930826.2016.1216224

Google Scholar's Special Metadata Requirements

- "Use Dublin Core tags (e.g., DC.title) as a last resort - they work poorly for journal papers because Dublin Core doesn't have unambiguous fields for journal title, volume, issue, and page numbers."
- Schema
  - Highwire Press
  - Eprints
  - BEPress
  - PRISM

https://scholar.google.com/intl/en/scholar/inclusion.html

A Quick Note About How SE Crawlers Work

- They do not actually “crawl” through databases
- Follow links, trigger HTML page generation
  - Index metadata that appears in…
    - Displayed HTML page
    - Non-displayed HTML <head> section
- Can also harvest and index PDF files

Structured data GS can identify, parse and digest

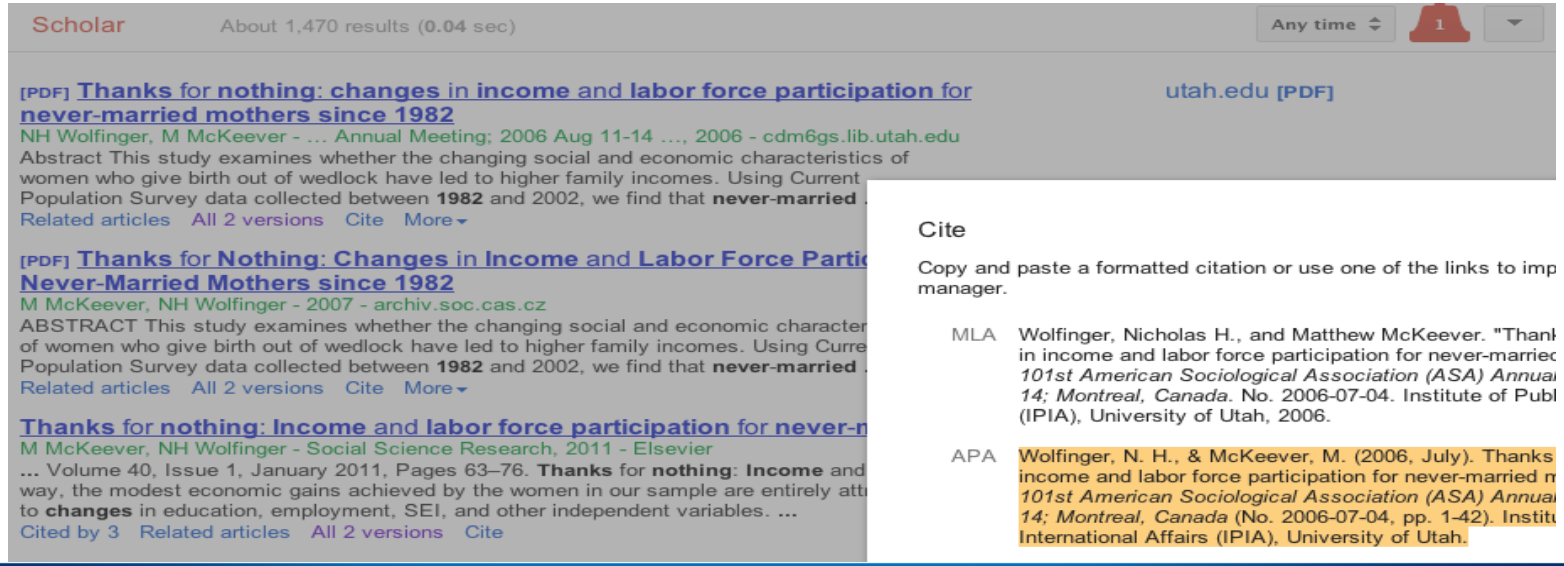
Human Readable Citation

Wolfinger, N. H., & McKeever, M. (2006, July). Thanks for nothing: changes in income and labor force participation for never-married mothers since 1982. In *101st American Sociological Association (ASA) Annual Meeting; 2006 Aug 11-14; Montreal, Canada* (No. 2006-07-04, pp. 1-42). Institute of Public & International Affairs (IPIA), University of Utah.

Citation for Google Scholar

```
<meta name="citation_title" content="Thanks for nothing: changes in income and lab
<meta name="citation_author" content="Wolfinger, Nicholas H." />
<meta name="citation_author" content="McKeever, Matthew" />
<meta name="citation_date" content="2006-07-26" />
<meta name="citation_firstpage" content="1" />
<meta name="citation_lastpage" content="42" />
<meta name="citation_keywords" content="Motherhood; Single Mothers; Income; Popula
<meta name="citation_technical_report_institution" content="Institute of Public &
<meta name="citation_technical_report_number" content="2006-07-04" />
<meta name="citation_language" content="en" />
<meta name="citation_conference_title" content="101st American Sociological Associ
<meta name="citation_pdf_url" content="http://cdm6gs.lib.utah.edu/utils/getfile/co
```

Google Scholar can read and understand!



IR SEO DEFICIENCY THEMES

IR 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.
- Analytics Reporting is Ineffective:
  - 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.

IR SEO Deficiency Themes (Technical)

- Poor experience for search engine customers
  - Slow servers
  - Failing to use secure transfer protocol - https (more, shortly)
  - Incorrect use of redirects
    - e.g. - know when to use 301 vs 302, or 403 vs 404
- Submitted sitemaps to Google to invite crawlers can conflict with server **robots.txt** file
- PDF files exceeding 5Mb

IR SEO Deficiency Themes cont'd (Technical)

- Website Design
  - Excessive use of graphics (SE crawlers are like disabled users)
  - Confusing site hierarchies and paths
    - Too many clicks to the PDF (GS says maximum of 10)
    - CMS/DAM must use canonical links
- Metadata
  - Not knowing what GS wants
    - Highwire Press, PRISM, e-prints, or BePress schema
  - Descriptive metadata not unique
  - Inaccurate and inconsistent (due to re-keying errors)

Secure Hypertext Transfer Protocol (HTTPS)

- Google penalizes sites not using HTTPS
- Encrypts data transferred between server and user
  - Significant privacy feature
- Privacy research: 279 research libraries (ARL, DLF, OCLC-RLP)*
  - 62% had implemented a secure digital certificate
  - Only 20% (of the 62%) used SEO best practice of redirecting non-secure requests to secure fulfillment
  - 15% turned secure requests into non-secure fulfillment

* Young, Scott W.H., Patrick OBrien, Kenning Arlitsch, Karl Benedict. "Measuring the Extent of Third-Party Tracking on Library Websites." Forthcoming publication

Summary - Be a Good Provider to Search Engines

- Responsive network, applications, and servers
- No dead ends
  - Appropriate redirects
- Secure transactions (HTTPS)
- Submit sitemaps
  - Ensure Robots.txt files that don’t conflict with sitemaps
- Metadata
  - Use appropriate schema and appropriate placement
  - Unique descriptions

ANALYTICS AND REPORTING

Analytics Tools to Use in Tandem

- Google Search Console
  - Diagnoses problems encountered by search engine crawlers
  - Provides other valuable SEO diagnostic data
- Google Analytics
  - Measures user visits
  - Be sure to set up as umbrella for all related sites/repositories
  - Fabulous for counting HTML pages, really stinky for non-HTML files

  - Establish a baseline with analytics tools before tweaking SEO!

Google Search Console dashboard

Google

Search Console

http://scholarworks.montana.edu/ | Help

- Dashboard
- Messages (25)
- Search Appearance
- Search Traffic
- Google Index
- Crawl
- Security Issues
- Web Tools

New and important

http://scholarworks.montana.edu/ is now associated with Google Analytics property ScholarWorks — Dec 15, 2016

View all

Current Status

Crawl Errors

Site Errors

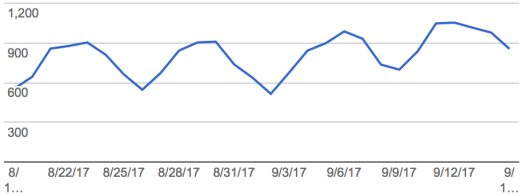
DNS	Server connectivity	Robots.txt fetch
✓	✓	✓

URL Errors

50 Server error
1,961 Not found
0 Other

Search Analytics

22,620
Total Clicks



Sitemaps

All (1)

12,676 URLs submitted
10,410 URLs indexed

4 warnings

Search Console

http://scholarworks.montana.edu/ | Help

- Dashboard
- Messages (25)
- Search Appearance
- Search Traffic
  - Search Analytics
  - Links to Your Site
  - Internal Links
  - Manual Actions
  - International Targeting
  - Mobile Usability
- Google Index
- Crawl
- Security Issues
- Web Tools

Search Analytics

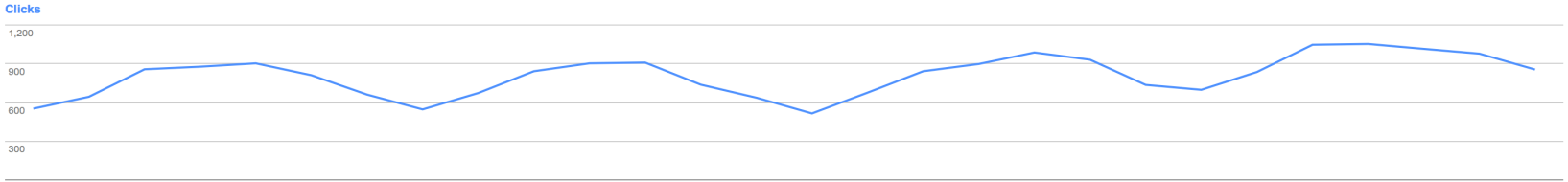
Analyze your performance on Google Search. Filter and compare your results to better understand your user's search patterns. Learn more.

✓ Clicks ☐ Impressions ☐ CTR ☐ Position

Queries	Pages	Countries	Devices	Search Type	Dates
No filter	No filter	No filter	No filter	Web	Last 28 days

Total clicks

22,620



	Queries	Clicks ▼
1	action research in mathematics	108
2	example of descriptive research paper	95
3	action research in mathematics pdf	75
4	what is grammar pdf	59
5	action research sample pdf	58
6	action research sample in mathematics	55
7	definition of grammar pdf	53
8	transformational leadership in education	48

Accurately Measuring File Downloads from IR

REPOSITORY ANALYTICS & METRICS PORTAL (RAMP)

Two Classes of Analytics Tools

- Page tagging analytics services (Google Analytics and others)
  - Great at counting HTML pages, lousy at counting file downloads
  - Used by 85%+ of research libraries
- Log file analytics
  - Captures file downloads very well, but also captures everything else
  - Robot traffic on the Internet
    - Up to 85% of IR downloads is non-human generated*

* Information Power Ltd (2013), "IRUS download data – identifying unusual usage", IRUS Download Report, available at: www.irus.mimas.ac.uk/news/IRUS_download_data_Final_report.pdf (accessed July 1, 2016).

A New Reporting Model

Page Type	Definition	Examples
Citable Content Downloads	Non-HTML scholarly content that may be formally cited in the research process	• Publication (.pdf) • Presentation (.ppt) • Data Sets (.csv)
Item Summary	HTML pages to help user decide to download the full publication	• Title & Abstract • Item Metadata
Ancillary	HTML pages that provide general information or navigation	• Search Results • Browse by Author • Statistics

DISCOVER

Author

- Arlitsch, Kenning (4)
- Mixter, Jeff (4)
- OBrien, Patrick (4)
- Sterman, Leila (3)
- Borda, Susan (2)
- Clark, Jason A. (2)
- Wheeler, Jonathan (2)
- Young, Scott W.H. (2)
- Banner, Katie (1)
- Border, J. Kent. (1)
- ... View More

Date Issued

- 2000 - 2017 (15)

Search

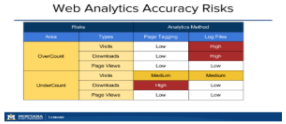
All of ScholarWork

Undercounting

Go

Show Advanced Filters

Now showing items 1-10 of 23



Data set supporting study on Undercounting File Downloads from Institutional Repositories [dataset]

OBrien, Patrick; Arlitsch, Kenning; Sterman, Leila; Mixter, Jeff; Wheeler, Jonathan; Borda, Susan (Montana State University ScholarWorks, 2016-07)

Search term found in abstract:...This dataset supports the study published as "**Undercounting** File Downloads from IR". The following items are included: 1. gaEvent.zip = PDF exports of Google Analytics Events reports for each IR. 2. gaItemSummaryPageViews.zip = PDF exports...

Undercounting File Downloads from Institutional Repositories

OBrien, Patrick; Arlitsch, Kenning; Sterman, Leila; Mixter, Jeff; Wheeler, Jonathan; Borda, Susan (Emerald, 2016-10)

A primary impact metric for institutional repositories (IR) is the number of file downloads, which are commonly measured through third-party web analytics software. Google Analytics, a free service used by most academic ...

(Search term found in fulltext file)

Undercounting File Downloads from Institutional Repositories

A primary impact metric for institutional repositories (IR) is the number of file downloads, which are commonly measured through third-party web analytics software. Google Analytics, a free service used by most academic libraries, relies on HTML page tagging to log visitor activity on Google's servers. However, web aggregators such as Google Scholar link directly to high value content (usually PDF files), bypassing the HTML page and failing to register these direct access events. This paper presents evidence of a study of four institutions demonstrating that the majority of IR activity is not counted by page tagging web analytics software, and proposes a practical solution for significantly improving the reporting relevancy and accuracy of IR performance metrics using Google Analytics.

View/Open

Preprint (833.8Kb)

Date

2016-10

Author

OBrien, Patrick
Arlitsch, Kenning
Sterman, Leila
Mixter, Jeff
Wheeler, Jonathan
Borda, Susan

URI

http://scholarworks.montana.edu/xmlui/handle/1/9943

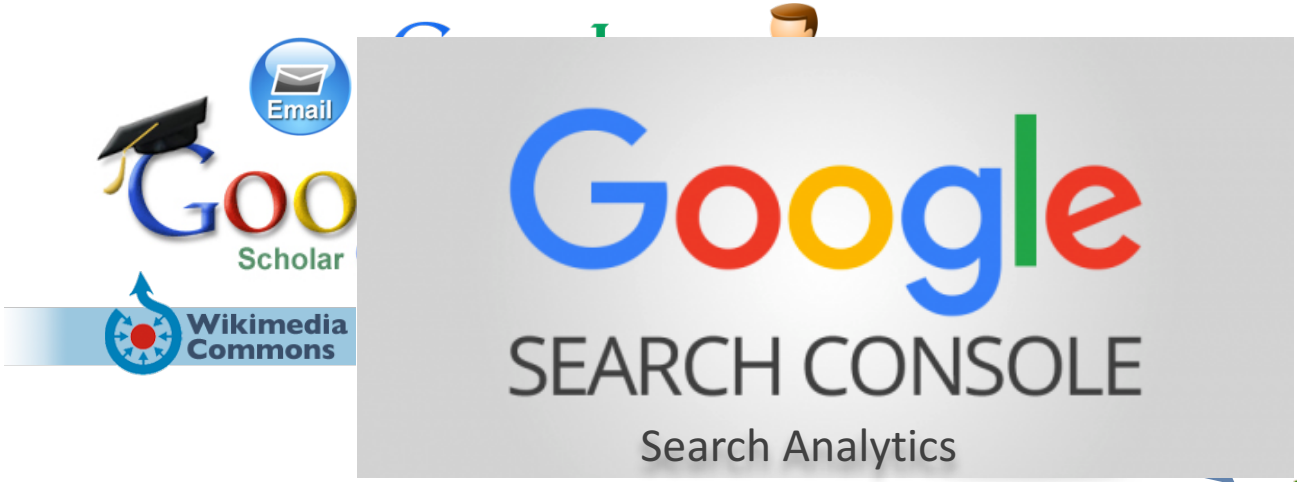
Related Material/Data

http://scholarworks.montana.edu/xmlui/handle/1/9939

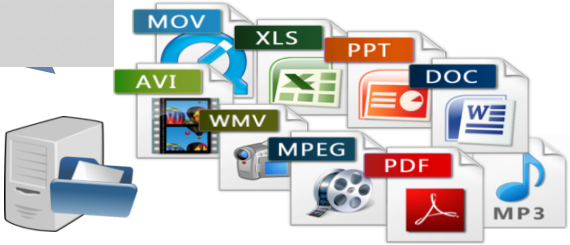
Collections

Scholarly Work - Library

Page tagging <u>does not</u> track non-HTML CCD



Does!



Non-HTML

SO, WE BUILT RAMP, BASED ON GSC

Why Google is Best at Filtering Robots

- Pay-Per-Click advertising model
  - 90% of Google's $75 billion revenue in 2015
- Advertisers pay Google average $.05-$50.00 per user click
  - They need certainty that clicks are humans, not robots
- Only Google has the resources to accurately filter robot traffic

Data set: Jan 5 - May 17, 2016 (n = 134 days)

Study Participant	IR	Platform	URL
Montana State University	ScholarWorks	DSpace	scholarworks.montana.edu
McMaster University	MacSphere	DSpace	macsphere.mcmaster.ca
University of New Mexico	LoboVault	DSpace	repository.unm.edu
University of Utah	USpace	CONTENTdm	uspace.utah.edu

Ancillary Page Views and Item Summary Page vs CCD

IR	Item Sumary PV	Ancillary PV	Total Google Analytics HTML PV	Download Events	Citable Content Downloads
scholarworks.montana.edu	26,735	23,350	50,085	7,129	77,380
macsphere.mcmaster.ca	51,150	71,585	122,735	n/a	133,342
repository.unm.edu	83,491	59,289	142,780	n/a	166,320
content.lib.utah.edu	122,927	47,569	170,496	19,226	159,536

All four IR using Google Analytics

CCD Tracking Improvement

Page Type	Analytics		Search Console
	Pages	Events	Search Analytics
Citable Content Download	-	26,155	562,933
Item Summary	34,103	-	-
Ancillary	201,793	-	-

2,000%

RAMP Landing Page

20 RAMP IR
as of September, 2017

Currently tracking 400,000+ digital items and capturing an average of 20,000 CCD per day that were previously invisible.

Current support for 4 IR Application Stacks

Repository Analytics & Metrics Portal

Search Here

Montana State University

ScholarWorks is an open access institutional repository for the capture of the intellectual work of Montana State University.

University of New Mexico

LoboVault is UNM's Institutional Repository. It hosts scholarly publications from UNM faculty, graduate student theses and dissertations, UNM administrative records, and more.

McMaster University

MacSphere is McMaster University's Institutional Repository (IR). MacSphere aims to bring together all of a University's research under one umbrella, in order to preserve and provide access to that research. The research and scholarly output included in MacSphere has been selected and deposited by the individual university departments and centres on campus.

Maryland MDSOAR

MD-SOAR is a shared digital repository platform for eleven colleges and universities in Maryland. It is jointly governed by all participating libraries, who have agreed to share policies and practices that are necessary and appropriate for the shared platform.

Maryland DRUM

The Digital Repository at the University of Maryland (DRUM) collects, preserves, and provides public access to the scholarly output of the university. Faculty and researchers can upload research products for rapid dissemination, global visibility and impact, and long-term preservation.

OAKTrust digital repository at Texas A&M

The OAKTrust digital repository at Texas A&M is a digital service that collects, preserves, and distributes the scholarly output of the University. The repository facilitates open access scholarly communication while preserving the scholarly legacy of the Texas A&M community.

Digital Collections of Colorado

The Digital Collections of Colorado is a digital service that collects, preserves and distributes digital material provided by a group of Colorado institutions for digital preservation and scholarly communication.

Michigan DeepBlue

Deep Blue is the University of Michigan's institutional repository service. It preserves and provides access to the research and creative work done by our faculty, staff, and students.

About the RAMP Portal

Montana State University, the Association of Research Libraries, the University of New Mexico, and OCLC Research have joined as partners to examine the difficulties that libraries face in producing accurate reports on the use of their digital repositories.

Publications

Patrick OBrien, Kenning Arlitsch, Leila Sterman, Jeff Mixter, Jonathan Wheeler, and Susan Borda. “Undercounting File Downloads from Institutional Repositories,” Journal of Library Administration, vol. 56, no. 7, 2016

Patrick OBrien, Kenning Arlitsch, Leila Sterman, Jeff Mixter, Jonathan Wheeler. “RAMP: Repository Analytics and Metrics Portal: A Prototype Web Service that Accurately Counts Item Downloads from Institutional Repositories,” Library Hi Tech, v35n1, March 2017

Proposal funded by IMLS:

”Measuring Up: Assessing Accuracy of Reported Use and Impact of Digital Repositories” scholarworks.montana.edu/xmlui/handle/1/8924

Conclusion

- SEO drives traffic to IR
  - Particularly when repository is optimized for Google Scholar
- Diagnose, measure and report website visits
  - Google Search Console and Google Analytics
- Accurately count IR file downloads with RAMP
  - Contact us to sign up

Questions?

Kenning Arlitsch, Dean of the Library
kenning.arlitsch@montana.edu
@kenning_msu