Invisible Institutional Repositories: Addressing the Low Indexing Ratios of IRs in Google Scholar

Authors: Kenning Arlitsch & Patrick S. OBrien

This is a preprint of an article that originally appeared in *Library Hi Tech* in 2012.

Kenning Arlitsch, Patrick S. O'Brien, (2012) "Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar", Library Hi Tech, Vol. 30 Iss: 1, pp.60 - 81
http://dx.doi.org/10.1108/07378831211213210

Invisible Institutional Repositories: Addressing the Low Indexing Ratios of IRs in Google Scholar

By Kenning Arlitsch and Patrick S. OBrien

Introduction

Search engine optimization (SEO) research conducted at the University of Utah has revealed that many institutional repositories (IRs) have a low indexing ratio[1] in Google Scholar (GS). IRs were developed to manage and ensure long-term access to academic publications, and GS was created as a search engine for those publications, whether they reside in IRs, at publisher repositories, or other research-oriented sites. This paper addresses the reasons for the low indexing ratio of many IRs, which the authors believe stem mainly from the metadata requirements of GS and which stand in contrast from the practices of many IRs. The authors conducted two surveys of IRs across the country and implemented three pilot projects designed to increase the indexing ratio of the IR at Utah. Transforming the metadata schema in those pilot projects led to a significant improvement in GS indexing ratio of the sample set. Additional reasons for the low indexing ratio of IRs can be tied to the ease with which GS's crawlers can navigate a given repository.

While much has been written about search engine optimization for general websites, very little has been published about SEO specifically for digital repositories and even less for institutional or disciplinary repositories. The subject

[1] Indexing ratio is defined here as the number of unique URLs from a given repository found in a search engine's index divided by the total number of URLs in the repository.

of this paper developed from more general digital repository SEO research the authors have conducted at the University of Utah's J. Willard Marriott Library for the past eighteen months, and whose continued research has recently been funded by a National Leadership Grant from the Institute of Museum and Library Services (IMLS). OCLC is a formal partner on this grant. The authors have dramatically improved the indexing ratio of Utah's digital library (including IR) in Google's main index, and some of that work will be discussed as background.

Digital repositories relevant to this article are defined as databases that store digitized or born-digital objects, making them freely accessible to the public. These repositories typically run on web server technologies, use descriptive metadata, and because their missions are generally directed at open access they all benefit from successful harvesting and indexing by Internet search engines. IRs are that subset of digital repositories that capture and manage the intellectual output of academic institutions or disciplines.

Research Question

The USpace IR at the University of Utah currently experiences an indexing ratio of less than 0.1% in GS, even though the indexing ratio for the same repository improved from approximately 18% to 98% in Google's main index after numerous SEO problems were addressed. The authors hypothesized that the average indexing ratio in GS of IRs across the United States is low, and that altering repository metadata to follow one of the publishing industry schemas recommended by Google Scholar would lead to a substantial improvement in the indexing ratio of USpace content.

Research Method

The authors conducted two surveys to identify the indexing ratio of other IRs, and the methodologies used for both are explained more fully in the S*urvey Methodologies* section. In brief, the authors selected repositories from the OpenDOAR directory (University of Nottingham 2011), gathered total item numbers from each repository, and then systematically sampled GS for evidence that those items had been included in its index. Two surveys (Survey 1 and Survey 2) were conducted, each using a substantially different searching methodology.

To test improvements to the USpace IR the authors gathered indexing ratio statistics and created a feedback loop to measure results of changes they made to repository items, demonstrated in three pilot studies. They submitted sitemaps containing URLs for all items in USpace to GS, and then used Google Webmaster Tools (Google 2011) to observe the activity of crawlers and the resulting indexing ratios. The authors adjusted the metadata for a subset of repository articles, and following a re-harvest they gathered additional statistics (Pilot 1). After this approach failed, the authors conducted discussions with OCLC and GS to confirm a new approach, which led to the development of metadata templates for various academic paper types whose effectiveness was tested during explicit harvests by GS (Pilot 2 and 3).

Google and Google Scholar

Google and Google Scholar are separate indexes, and GS has a different focus from its much larger parent. Dr. Anurag Acharya, GS's founding engineer, has stated that the goal is to offer the "most comprehensive list of research papers available on

the Web," and that GS limits its results to "peer reviewed papers, theses, books, abstracts, and technical reports" (Assisi 2005). More recently, GS has added patents and legal cases to the items it indexes.

GS has its own crawlers (also known as spiders or robots) that visit repositories and publisher sites, among others, to harvest content appropriate for its index. A peculiarity of GS's presentation of academic papers is that it generally provides a link directly to the PDF document. This is expedient for users as it gets them directly to the content, but it also strips any context that may have been provided by the repository's HTML display. In other words, metadata, institutional logos, and other information normally displayed to users are lost unless they are inserted into the PDF itself. The practice can also affect the reporting of visitation statistics through website analytics software that utilize page tagging. For instance, Google Analytics requires a tracking code inserted in the HTML of each page of a given website to gather statistics. Each time that page is displayed in a Web browser it is counted as a visit by Google Analytics, but separating the PDF file from the HTML display means that the visit will not be counted because the required tracking code is not executed when the PDF is called directly. This problem can be overcome by having the webserver execute a PHP script containing the tracking code before serving the requested PDF, but it is unlikely that many repository managers are doing this, or are even aware that their visitation and download statistics may be underreported as a result of GS's item display practice.

Literature Review

Internet search engines dominate general information-seeking behavior of

users, and Google is by far the most popular search engine, consistently grabbing 65% share of the "explicit core search market" (Comscore, Inc. 2011). Bing powers Microsoft and Yahoo! search sites, capturing another 30% of market share. The dominance of search engines is also apparent in the academic sector. A 2005 survey by OCLC demonstrated that 89% of college students began their research with Internet search engines, and that only 2% began at library websites (DeRosa & OCLC. 2005). A repeat of that survey five years later demonstrated that the situation for libraries had only worsened, as 0% of respondents reported visiting library websites at the outset of their research (DeRosa et al. 2010). That same report saw a slight drop in traditional search engine use, but also noted for the first time the use of social media search engines for initial research. Another 2005 survey in the UK found that "students prefer to locate information or resources via a search engine above all options, and Google is the search engine of choice" (Griffiths & Brophy 2005). The information-seeking behaviors of young academic researchers in Sweden displayed an "almost complete dominance of Google as a starting point for searching *scientific* information" (Haglund & Olsson 2008).

Faculty search behavior is similar. A study of active faculty researchers at four major universities reported that "researchers find Google and Google Scholar to be amazingly effective" for their information retrieval needs and accept the results as "good enough in many cases" (Kroll & Forsman 2010). Rieger reports a high degree of use and satisfaction with Internet search engines. She notes that "both faculty and students prefer search engines over other resources to support their academic work" and that "there is a broader awareness of specialized Google tools

such as Google Scholar and Google Book among faculty members and graduate students" (Rieger 2009). In a comparison of GS to Web of Science, Mikki states that "the amount of qualified scholarly content has increased considerably in Google Scholar since it was launched in 2004," and that it has developed into a serious research and citation study tool that should be included in information literacy programs (Mikki 2009).

A review of the literature pertaining to SEO in libraries reveals that much of the published research deals with general websites, e.g. (Cahill & Chalut 2009) and (Rushton et al. 2008). The minimal research dealing with digital repositories sometimes concludes by suggesting that content be replicated outside the database in a static format in order to make it friendlier to search engines, a method that seems arcane and burdensome, but may have been the best option at the time. "Unless links are located on a static web page, crawlers won't find them, and many such links are not followed" (DeRidder 2008). Page rank in search engines is another factor that plays into repository visibility. Malaga has shown that 62% of users click only on results that appear in the first search engine results page (Malaga 2008). The high use of Internet search engines as primary search mechanisms suggests that digital repositories created by libraries are likely to be nearly invisible to users if their contents are not indexed in these search engines.

Search engine and metadata optimization for institutional repositories are also addressed only minimally in the published literature, and the value and use of GS is sometimes questioned. McKay offers that "authors are quite right in perceiving [IRs] as 'islands of information,' … a condition that can be addressed by

search-engine harvesting…" She goes on to say "Google Scholar is not usually the first information source" consulted by academics, though that may have been truer when the article was published in 2007, when GS was relatively new and contained much less content (McKay 2007). Increased use of GS is demonstrated in a more recent University of Mississippi study in which use rose from 4% to 27% of major library databases over a four-year period (Herrera 2010). A 2006 article on optimizing metadata for search engines acknowledges "the problem may not lie with the search engines but with the data providers," and introduces the concept of "data shoogling" to offer more Google-friendly metadata in digital collections (Dawson & Hamilton 2006). It does not, however, specifically address institutional repositories or GS. A survey of 540 librarians at 108 ARL libraries notes complaints of "inadequate use of metadata by Google Scholar" (Drewry 2007), which may support this paper's hypothesis that GS doesn't find metadata supplied by libraries to be appropriately structured or unique.

Beel, Gipp, and Wilde offer related and significant strategies for optimizing academic papers themselves for better inclusion in search engines, and in GS in particular. Their advice includes optimizing graphics for indexing purposes, writing relevant document titles, and selecting appropriate keywords (Beel et al. 2010) . While optimization of the academic papers themselves merits continued exploration and testing, it is beyond the scope of this article.

Background on SEO Research for Digital Repositories at the University of Utah

Digital repositories of every type face a common challenge: having their content found by interested users in a crowded sea of information on the Internet.

Getting found means the repository items must be included in the indexes of major search engines, because that is where the vast majority of users start looking. Unfortunately, many digital repositories show poorly in the results from major search engines. In 2010 the authors conducted a survey of 650 known objects across the thirteen repositories of the Mountain West Digital Library (MWDL), and revealed a disturbing pattern: only 38% of digital objects searched by title were found in Google's index. Worse, this Google Search Engine Results Page (SERP) consisted mostly of links back to a search results screen in the local repository, rather than linking directly to the objects. Only 15% of the hits on the SERP provided users with direct links to the objects. The known-item title searching method employed by the survey probably produced the best results possible at the time; searching by keyword or subject term would likely have presented even fewer items from the repositories of the libraries and archives in the MWDL.

Search engines can be thought of as "users with substantial constraints: they can't read text in images, can't interpret JavaScript or applets, and can't 'view' many other kinds of multimedia content" (Hagans 2005). In order to be indexed by Internet search engines, repository databases and the servers on which they reside must be receptive to crawlers sent out by the search engines. The crawlers follow links to each digital object in the repository; a process greatly facilitated by the submission of sitemaps that function as formal invitations and guides, revealing the preferred URL for each of the repository's objects. The crawlers "harvest" metadata and other information about the objects, sending that information back to the search engine where it is analyzed by algorithms that take many factors into

account in deciding whether to add the metadata to the search engine's index. Crawlers that encounter difficulties in the harvesting phase will throw off errors that can be analyzed and addressed using free Webmaster Tools services offered by both Google and Bing. Errors that are not addressed in a timely manner may discourage crawlers from returning, leading to continuing low search engine indexing ratios, or worse, being dropped from the index altogether. Technical problems encountered by crawlers may include, but are not limited to the following:

1. Conflicts between sitemaps and robots.txt files
2. Slow server response time
3. Dead links or failure to provide appropriate redirects
4. Labyrinths created by repository software, including poorly implemented framesets and JavaScripts, as well as multiple URLs for the same object
5. Poor application of metadata, including re-use of the same metadata terms for multiple objects
6. Metadata schemas deemed unacceptable by the specific search engine

Additional challenges with search engine optimization for digital repositories may be framed as administrative. These include:

1. Aligning the goals of the digital library with institutional goals
2. Informing, training, motivating, and coordinating staff from various departments
3. Establishing an environment of continuous monitoring and addressing crawler errors as they arise

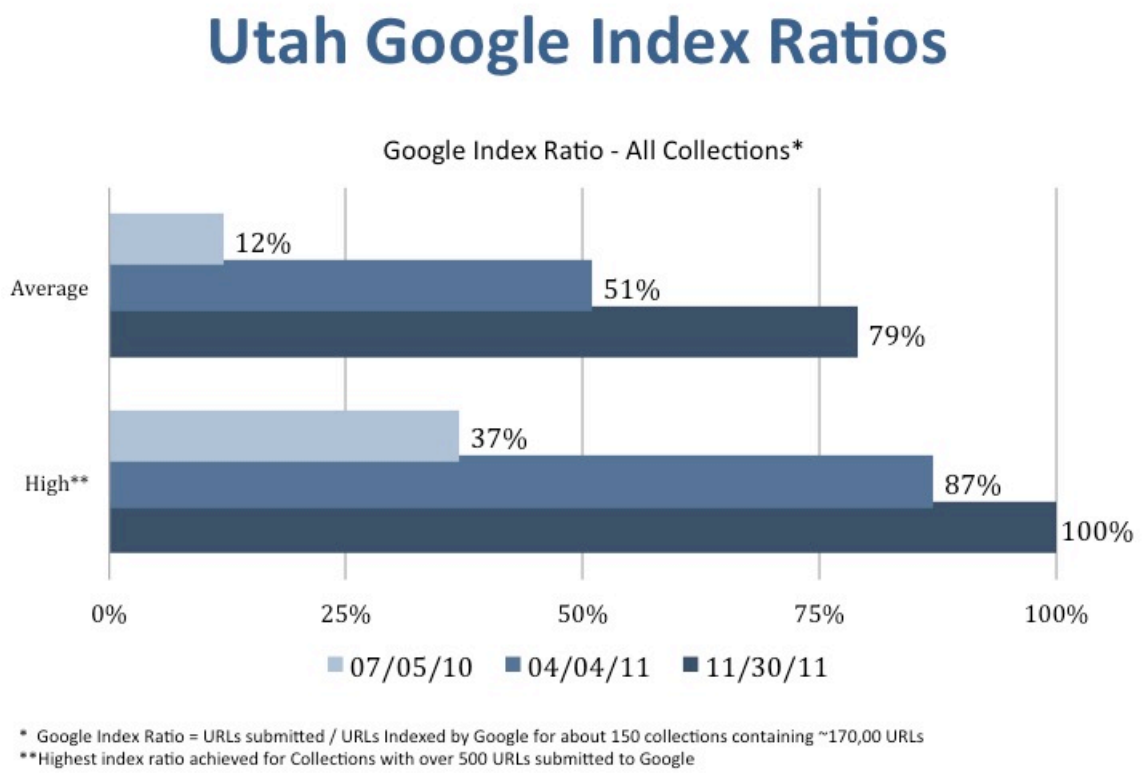
4. Institutionalizing tools to analyze metrics, and using them to inform and convince stakeholders of the impact of the digital library

Results at Utah

Analyzing the SEO problems and applying a variety of solutions has resulted in dramatic improvements to the indexing ratio of Utah's digital repositories in Google. The digital repositories (including the IR) managed by the J. Willard Marriott Library at the University of Utah currently run on CONTENTdm v5.4 (OCLC, Inc. 2011). While version 5.4 includes some features that are considered unfriendly to search engines, such as JavaScripts, framesets (for compound objects), and multiple URLs for digital objects, those ultimately proved not to be barriers to successful SEO (see Chart 1). Version 6, which was released in late 2010, eliminates those barriers altogether, and Utah is planning a migration.

The chart below shows increases in average indexing ratios for all digital collections over a period of fifteen months. It also shows improvements in the highest indexing ratio achieved for collections with more than five hundred URLs.

Chart 1: Google Index Ratio improvement for general digital collections at Utah



* Google Index Ratio = URLs submitted / URLs Indexed by Google for about 150 collections containing ~170,00 URLs
**Highest index ratio achieved for Collections with over 500 URLs submitted to Google

Increased indexing ratios have thus far led to a 200% increase in referrals from Google, and an 80% increase in visits to all digital collections. Indexing ratios of USpace, the University of Utah's IR, have also increased from approximately 18% to 98%, but only in Google, not in Google Scholar (see Chart 3).

Open Access and Institutional Repositories

The open access movement was launched to improve access to publicly funded research, and to help libraries deal with rampant inflation in journal subscription prices. According to Peter Suber the open access movement is dependent on Internet technologies and the consent of the author or copyright

holder (Suber 2004). Institutional repositories were one product of this movement; they capture the intellectual output of the faculty, staff, and students of universities or academic disciplines, and assure perpetual and free access to that output (barring embargo periods or other publisher restrictions). IRs often include electronic theses and dissertations (ETD), and most are managed by academic libraries and some by scholarly societies. Over the past decade IRs have variously enjoyed advances and suffered setbacks, but through the consistent work of many individuals at numerous institutions they are achieving enough mass to become viable sources of research publications. They also hold the promise of contributing significantly to author citation rates. Recent research in the UK suggests that institutional repositories may play a crucial role in measuring research output, and in turn may affect university rankings (Key Perspectives Ltd. & Brown 2009). The *Times Higher Education* publishes an annual ranking of the top world universities, and research citations contribute 32.5% toward each university's score (The Times Higher Education 2010).

Libraries have not developed a mechanism to aggregate and search IRs, and thus GS has become the best de facto search engine available for IR content. But just as institutional repositories are gaining enough mass to make them useful and credible sources of research output, the difficulties associated with SEO threaten to undermine their potential. Faculty and other authors who contribute publications to IRs may lose interest if their publications can't be located (and cited) in academically-oriented search engines like GS.

Surveys of IR Indexing Ratios in Google Scholar

In October and December 2011 the authors conducted two surveys of institutional and disciplinary repositories to arrive at a preliminary determination of how well GS was indexing them. The IRs were identified through the Directory of Open Access Repositories, also known as OpenDOAR (University of Nottingham 2011).

Only institutional or disciplinary repositories housed in the United States were selected for these surveys. They were chosen for their academic content, and to represent an approximate real-world distribution of several repository software types: DSpace, Digital Commons, EPrints, Fedora, IR+, CONTENTdm, DigiTool, and arXiv (see Table 1). While there are a number of other software types in use, many of them are not found in the U.S. Some repositories found in OpenDOAR were ruled out because it was immediately obvious that they included other types of non-IR digital collections, such as photographs. According to OpenDOAR the arXiv repository software is used only by arXiv, but it was included in Survey 1 because of its size and importance to the scientific community. (See Table 1 for a complete listing of the repositories selected for Survey 1)

Survey Methodologies

Search engine indexing is a dynamic environment. Crawlers return to repositories periodically to pick up new additions, sometimes discarding items if they run into errors, and the repositories themselves are (hopefully) continually growing. Therefore these surveys should be understood to be a snapshot from a specific moment in time.

OpenDOAR records list the number of items in most repositories, but those figures are usually outdated. The authors determined the current number of repository items from figures available on the sites themselves, and in one case by contacting the repository manager. DSpace repositories make it is easy to browse by Title to reveal all the items in the repository. In the case of Digital Commons, a dynamic script posts the current total items in the repository. EPrints repositories had a page that listed the number of items by type, the sum of which represented all the items in the repository. Other sites offered similar methods of determining the total number of items contained in the repository.

Survey 1 Methodology

In the first survey searches were conducted to determine the number of items indexed by GS from a given repository by using the "site" operator, i.e. search queries in GS were structured in the following manner: 'site:repositoryURL.' This operator must be used with caution, because in GS it only searches the **primary** versions of academic papers. In other words, a paper that has been formally published in a journal will be considered the primary version. Additional versions of that paper, including those that appear in IRs may be indexed by GS, but are considered "other" versions and will only be revealed by clicking the "versions" link (see Figure 1). Because the "other" versions don't appear on the initial search results page, it is incorrect to assume that the number of results of a search using the site operator shows all the items that GS has indexed from that repository.

Figure 1: Google Scholar Search Result Showing Link to Other Versions of the Paper

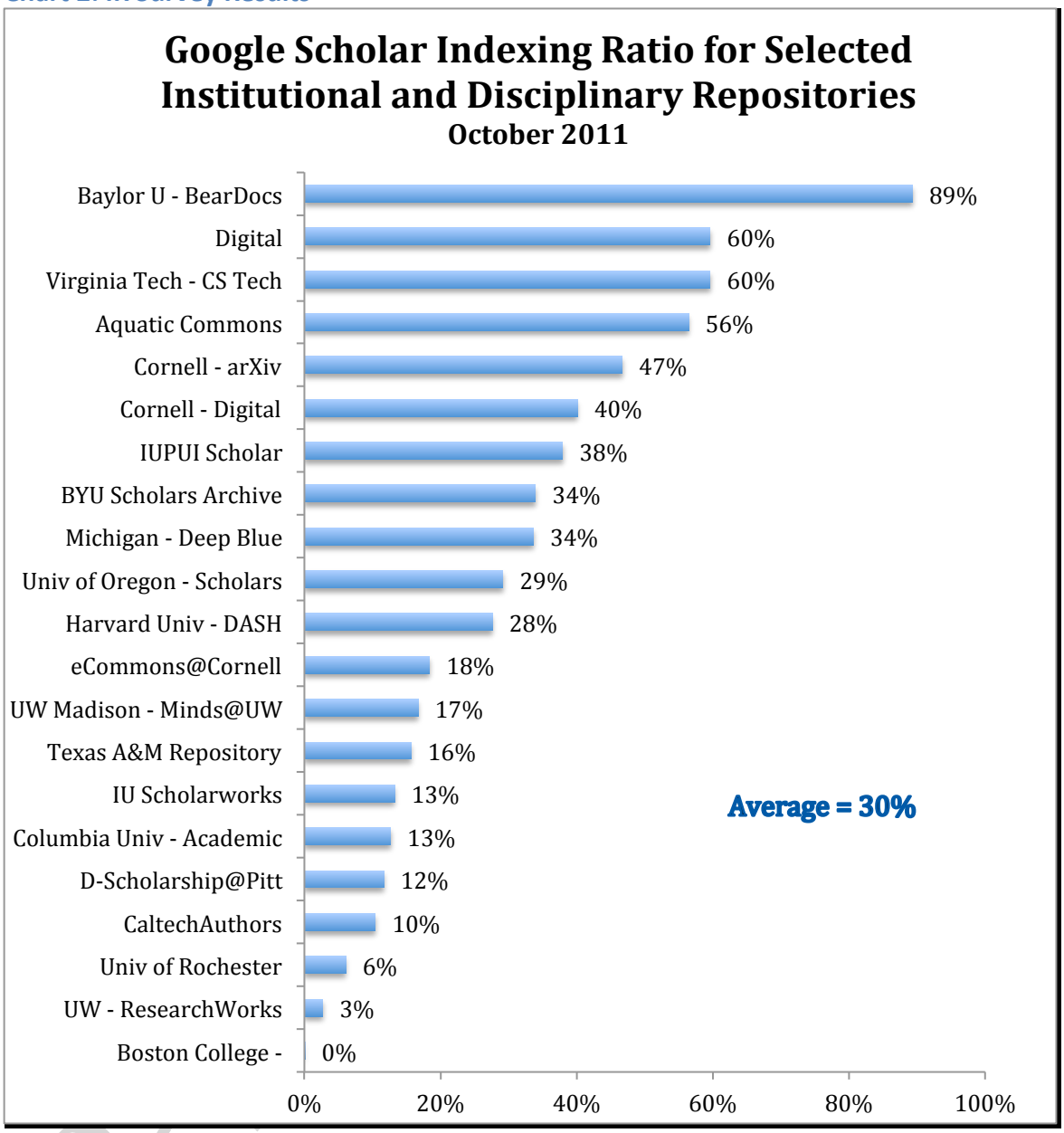
The data from this survey confirm a low average primary publication indexing ratio of only 30% (see Table 1 and Chart 2). Being mindful of casting aspersions, the authors are fully aware that their own IR (USpace) currently shows a near zero percent indexing ratio in GS.

Table 1: Survey 1 of IRs Showing Primary Publication Version Indexing Ratios

Repository Name	Repository Software	Repository URL	Reposit-ory items	Items in Google Scholar	Indexing Ratio
Boston College-eScholarship@BC	DigiTool	dcollections.bc.edu	1,635	1	0%
UW - ResearchWorks Archive	DSpace	digital.lib.washington.edu/dspace	11,285	304	3%
Univ of Rochester Research	IR+	urresearch.rochester.edu	16,184	983	6%
CaltechAuthors	Eprints	authors.library.caltech.edu	22,000	2,290	10%
D-Scholarship@Pitt	Eprints	d-scholarship.pitt.edu	5,888	686	12%
Columbia Univ-Academic Commons	Fedora/ Blacklight	academiccommons.columbia.edu	4,631	586	13%
IU Scholarworks	DSpace	scholarworks.iu.edu/dspace	7,782	1,030	13%
Texas A&M Repository	DSpace	repository.tamu.edu	46,324	7,250	16%
UW Madison-Minds@UW	DSpace	minds.wisconsin.edu	15,078	2,520	17%
eCommons@Cornell	DSpace	ecommons.library.cornell.edu	18,544	3,410	18%
Harvard Univ - DASH	DSpace	dash.harvard.edu	6,193	1,710	28%
Univ of Oregon-Scholars Bank	DSpace	scholarsbank.uoregon.edu/xmlui	9,740	2,840	29%

Michigan - Deep Blue	DSpace	deepblue.lib.umich.edu	66,038	22,200	34%
BYU Scholars Archive	CONTENTdm	scholarsarchive.lib.byu.edu	7,421	2,520	34%
IUPUI Scholar	DSpace	scholarworks.iupui.edu	2,109	800	38%
Cornell-Digital Commons@ILR	Digital Commons	digitalcommons.ilr.cornell.edu	14,669	5,880	40%
Cornell-arXiv	arXiv	arxiv.org	706,906	330,000	47%
Aquatic Commons	Eprints	aquaticcommons.org	5,722	3,230	56%
Virginia Tech - CS Tech Reports	Eprints	eprints.cs.vt.edu	983	586	60%
Digital Commons@UNLincoln	Digital Commons	digitalcommons.unl.edu	50,657	30,200	60%
Baylor U - BearDocs	DSpace	beardocs.baylor.edu	928	829	89%

Chart 2: IR Survey Results



Survey 2 Methodology

In the second survey the authors used a similar approach to the one they had employed in 2010 to survey the repositories of the Mountain West Digital Library, i.e. they searched in GS for known repository items by their titles. This method is, of course, slower and more laborious, but it is also more accurate, allowing articles to

be counted whether they appear as the primary link in the initial list of results or are hidden behind the "versions" link.

The authors created a data set for seven repositories from Survey 1 by using crawler software to harvest titles from each repository. This method mimicked the process used by Internet search engine crawlers, and collected 500 to 1,400 article titles from each repository and saved them into Excel spreadsheets. In some cases, scholarly papers in the IRs were easy to identify and entire collections could be crawled. In other cases, it was difficult to isolate the publicly available scholarly papers for an automated crawler because the repositories don't follow the GS recommendations. These difficulties in crawling the IR resulted in less than optimal sampling of titles within the IR collection; in fact the sample may have been biased to favor a higher indexing ratio. Using a sampling methodology developed for verifying database backups (LaRock 2010) those titles were then randomized, and forty titles from each set were searched by copying article titles from the spreadsheets and pasting them into the GS search box. The authors used Zotero to create metadata records and snapshots for each search result, whether the article was found or not. "Versions" links were followed whenever found and the resulting screen was also captured as a snapshot attached to the same metadata record in Zotero.

Of the seven repositories that were sampled, three showed very high indexing ratios (88% - 98%), while the other four showed ratios below 50% (see Table 2). A discussion about the likely reasons for these differences follows in the section titled "Drawing Conclusions from the Surveys."

Table 2: Survey 2 Indexing Ratios for Seven Institutional Repositories

	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							
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							
Meta Tag Schema in HTML headers	BePress	DC	ePrints & DC	None	DC & DCTERMS	None	None
Title	Yes	Yes	Yes	No	Yes	No	No
Author	Yes	Yes	Yes	No	Yes	No	No
Pub Date	Yes	Yes	Yes	No	DCTERMS	No	No
Publisher	Yes	Yes	Yes	No	No	No	No
Journal	No	No	Yes	No	No	No	No
Volume	No	No	Yes	No	No	No	No
Issue	No	No	Yes	No	No	No	No
First Page	No	No	Yes	No	No	No	No
Last page	No	No	Yes	No	No	No	No
Absolute URL to PDF	Yes	Yes	Yes	No	No	No	No
Institution	n/a	n/a	n/a	na	n/a	No	n/a
Dissertation Name	n/a	n/a	n/a	na	n/a	No	n/a

Survey Conclusions

The first survey had limitations in terms of calculating a complete index ratio for each IR. However, since use of the site operator in GS reveals only the primary versions of the articles, the average indexing ratio of 30% indicates that most IRs do not contain very many primary articles. This raises some interesting questions about the purpose of IRs. Specifically, is there really much point in IRs capturing pre-prints of published literature? How much value is really derived from having those pre-prints in the IR, given the amount of labor required to put them there, particularly if the primary publisher is open access as well? On the other hand, IRs that largely contain grey literature that is not published elsewhere will likely see a much higher indexing ratio with GS precisely because those are the primary articles.

Data from the second survey are much more interesting. Because the authors used crawler software to harvest article titles, they encountered many of the same problems that Internet search engine crawlers face when trying to harvest institutional repositories. The guidelines shown in Table 2 were drawn from stated requirements and recommendations from GS's Webmaster Inclusion Guidelines website (Google Scholar 2010). In general, IRs that followed these guidelines had a much higher indexing ratio (88%-98%) than sites that did not (38%-48%). For the purposes of this paper, the most validating differences were found in the expression of publisher metadata schemas (Bepress, Highwire Press, PRISM, or Eprints) in the meta tags within the header tags of the HTML display pages (see Figure 2). Those repositories that did not make their metadata available in one of the recommended publisher schemas within the HTML meta tags generally fared much more poorly

than those that did. Further, the repositories that offered an absolute URL to the PDF file for their documents also had far higher indexing ratios than those that did not. Finally, improving crawler efficiency by providing chronological listings of papers, recently added papers, and a limited number of clicks to publically available scholarly papers also seemed to affect indexing ratio.

Figure 2: Example of HTML Meta Tags Using of Bepress Schema

```
<meta name="bepress_citation_author" content="Webber, Douglas A.">
<meta name="bepress_citation_author_institution" content="Cornell University">
<meta name="bepress_citation_author" content="Ehrenberg, Ronald G.">
<meta name="bepress_citation_author_institution" content="Cornell University">
<meta name="bepress_citation_title" content="Do Expenditures Other Than Instructional Expenditures Affect
<meta name="bepress_citation_date" content="2010">
<meta name="bepress_citation_pdf_url" content="http://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?
```

GS makes specific recommendations for IR software on its Inclusion Guidelines for Webmasters site (see reference below), but the surveys in this paper demonstrate that software makes little or no difference; the problem cuts across institutions, repository focus, and repository software. Instead, indexing ratio success has much more to do with how carefully a repository follows the guidelines described, above.

> **If you're a university repository,** we recommend that you use the latest version of Eprints (eprints.org), Digital Commons (digitalcommons.bepress.com), or DSpace (dspace.org) software to host your papers. If you use a less common hosting product or service, or an older version of these, please read the rest of this document and make sure that your website meets our technical guidelines (Google Scholar 2010)

Why Google Scholar Has Difficulty with Institutional Repositories

Librarians are great believers in standards, and while building digital repositories they have dutifully followed them for scanning, metadata creation, harvesting, and web services, among others. Search engines, however, are not required to honor standards. For example, in August 2008 Google announced that it was "Retiring support for OAI-PMH[2] in Sitemaps" (Mueller 2008), causing consternation across the library community. Two years later, GS made the following announcement on its Webmaster Inclusion Guidelines site: "Use Dublin Core tags (e.g., DC.title) as a last resort - they work poorly for journal papers…" (Google Scholar 2010).

Although Dublin Core is recognized to be a standard of the lowest common denominator, libraries have used it widely for most digital repositories, including IRs. The Dublin Core schema works "poorly for journal papers" because it does not include adequate fields for citation data and because it is interpreted inconsistently. Citation information such as journal name, volume and issue number, and page numbers span of the article is usually entered into a single field, such as DC.Relation or DC.Source in simple Dublin Core, and there is no specified format or consistency. This makes it difficult for a search engine like GS to accurately parse and index the data into their individual bibliographic components. The Dublin Core Metadata Initiative website (DCMI 2005) does include guidelines for encoding bibliographic citation information using a qualification of the DC.Identifier field (called "bibliographicCitation") but this is still only a single field. It is also unlikely that

[2] (Open Archives Initiative Protocol for Metadata Harvesting, a common standard for sharing metadata in the library community)

many repositories have updated to reflect the relatively recent development of DC Qualifiers. Dublin Core also doesn't facilitate various academic paper types: there is no specific field to distinguish a pre-print from a journal article, a book chapter from a book, a working paper from a conference proceeding, or a dissertation.

Instead of Dublin Core GS recommends using one of the following schemas: Highwire Press, Eprints, Bepress, and PRISM. These schemas are more adept at structuring citation data appropriately. Highwire Press, a division of Stanford University, developed its schema for journal articles and GS extended the tags to cover additional academic paper types, such as working papers, dissertations, manuscripts, conference papers, books and book chapters. The authors used the extended Highwire Press tags in their pilot projects to test the hypothesis that transforming metadata would lead to an increase in indexing ratio in GS for an IR.

Pilot 1

Due to the Marriott Library USpace's non-existent showing in GS, the authors began to strategize methods to modify USpace IR metadata to fit the recommendations. GS explains how Highwire Press tags could map to Dublin Core fields (Google Scholar 2010). Thus the first step was to begin aligning existing Dublin Core fields with those mappings (see Table 3).

Table 3: Map used in first GS pilot

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

The indexing ratio for USpace at the University of Utah prior to the pilot (July 5, 2010) was poor, at best, and can be summarized as follows:

1) Index ratio for the three primary USpace IR collections containing 6,482 papers
    a) Ranged between 4% and 23% within Google
    b) Average overall Google Index Ratio was 18.33% (1,188/6,482)
    c) Index ratio for IR collections within GS was less than 0.1%

The following steps were taken to address the poor indexing ratio:

1) Sitemaps representing three IR collections were submitted through Google Webmaster Tools
    a) A total of 6,482 URLs were submitted
        i) Each collection contained between 500 and 4,200 academic papers
2) Errors generated during Google crawls were analyzed using Webmaster Tools and improvements were made
    a) Improved server performance

b) Implemented unique title and description tags containing the paper's name and abstract, respectively

c) Implemented "rel=canonical" tags, indicating the preferred URL of each digital object (there were often multiple URLs pointing to each paper).

To address the metadata requirements per the Google Scholar inclusion guidelines the authors did the following:

1) Mapped Dublin Core to Google-supported Highwire Press tags

a) Extended Dublin Core fields according to GS recommendations

i) Journal Volume (DC.volume)

ii) Journal Issue (DC.issue)

iii) Starting Page Number (DC.citation.spage)

iv) Ending Page Number (DC.citation.epage)

Twenty papers were selected for a pilot

1. Verified metadata was accurate and mapped correctly to the HTML "meta name=" fields on display templates as understood from GS inclusion guidelines (see Table 3 and Figure 3)
2. Ensured each of the 20 papers had a full-text PDF that met GS inclusion guideline requirements
3. Embedded the metadata schema directly into five of the PDF files of the papers
4. 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. This

landing page contained links to both a paper's HTML page and its full-text PDF

Figure 3: Converting Bibliographic Data

Below is an example of converting the bibliographic data being stored in a single field. The DC.Relation field data, were parsed and presented in the HTML header for Google Scholar.

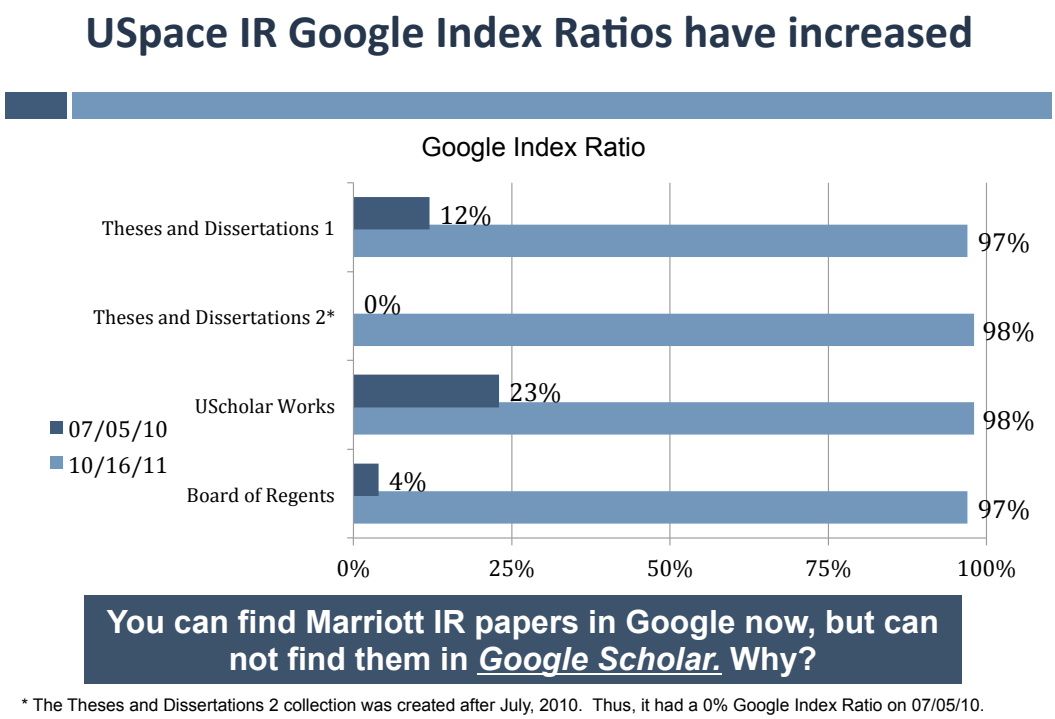
McKeever, M. & Wolfinger, N.H. (2006). Thanks for Nothing: Changes in Income and Labor Force Participation for Never-Married Mothers since 1982. Institute of Public & International Affairs (IPIA), 4, 1-43.

```
1 <meta name="DC.title" content="Thanks for nothing: changes in income and labor force
2 <meta name="DC.creator" content="Wolfinger, Nicholas H.">
3 <meta name="DC.issued" content="2006/07/26">
4 <meta name="DC.citation.spage" content="1" />
5 <meta name="DC.citation.epage" content="43" />
6 <meta name="DC.Publisher" content="University of Utah">
7 <meta name="DC.Contributors" content="Institute of Public and International Affairs (
8 <meta name="DC.identifier" content="http://content.lib.utah.edu/cgi-bin/showfile.exe?
```

The experiment delivered a significant increase in the Google index ratio for the IR collections (see Chart 3), and as of October 16, 2011 the Google Index ratio for the IR collections was 97.82% (10,306/10,536). However there was no effect on the IR's GS index ratio. In fact, not one of twenty USpace papers that had been isolated and optimized was included in the GS index.[3]

[3] USpace added a second Theses and Dissertations collection after the first GS pilot was started in July, 2010.

Chart 3: Increase in USpace Indexing Ratios in Google



* The Theses and Dissertations 2 collection was created after July, 2010. Thus, it had a 0% Google Index Ratio on 07/05/10.

Pilot 2

During the summer of 2011 the authors consulted with OCLC and Google Scholar with the aim of developing and testing a second pilot project. Nineteen papers from USpace were selected for the second pilot

1. Six of seven GS paper types were represented and the full text PDF document was included for each paper. The book paper type was out of scope for this pilot. (See Appendix for examples of each paper type)
    a. Dissertation & Thesis
    b. Conference Article
    c. Working Paper

d. Manuscript & Pre-Print

e. Journal Article

f. Book Chapter

2. CONTENTdm v6.0 display templates were augmented

   a. Embedded Highwire Press meta tags in the HTML page header of display templates using an automated script (see Figure 4)

   b. Created a Browse By Year page that provided links to papers in chronological order of publishing date

   c. Created a Recently Added page that listed papers added to the IR within the last 30 days

Figure 4: Highwire Press tags embedded in HTML headers

```
<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
```

The second pilot was a moderate success, with 62% of papers indexed on the first harvest. However, due to unexpected campus network and power outages that took down the test server for an extended period, the pilot was cut short and the results were dropped from GS's index.

Pilot 3

For the third and final pilot project, the authors uploaded fifty-six papers with full-text PDF files, and transformed the Dublin Core metadata to Highwire Press tags as described earlier. The same six paper types were represented as before. This time more than 90% appeared in the GS index after four weeks. Continuing conversations with GS and OCLC will help address lingering issues, but the authors consider this success to be a significant breakthrough.

Transforming Metadata

The thought of manually transforming metadata for an IR might induce nausea in repository managers. Fortunately, the IMLS NLG grant recently awarded to the University of Utah intends, as one of its deliverables, to help address this problem. OCLC is a partner in the grant and will develop formal crosswalks between Dublin Core and one or more of the publishing industry schemas recommended by GS. Automated transformation and linked data mechanisms will also be developed to minimize the work required to express citation data more effectively for indexing. The products of that grant will be published in a toolkit in 2014.

Conclusion

Transforming metadata to GS-preferred metadata schemas is very likely to raise indexing ratio of IRs. The second and third pilot projects described in this paper were successful, demonstrating that transforming from Dublin Core metadata tags to more precise bibliographic Highwire Press tags increased the sample data

set GS indexing ratio from 0% to 62% in the second pilot, and then to more than 90% in the third. The authors are cautiously optimistic that continuing discussions with GS and OCLC will eliminate most remaining indexing problems. Transforming metadata to EPrints, PRISM, and Bepress schemas is also likely to have a positive effect, though this assertion will require additional testing.

The low indexing ratio of IRs in GS cuts across institutions and repository software. Despite GS's endorsement of three software packages, the surveys conducted for this paper demonstrates that software is not a deciding factor for indexing ratio in GS. Each of the three recommended software packages showed good indexing ratios for some repositories and poor ratios for others. Rather, the major deciding factors seem to lie in: 1) whether the IR has provided crawlers an efficient method to access its scholarly papers; and 2) whether acceptable metadata schemas are provided that offer precise bibliographic information within the HTML page header tags.

While transforming metadata seems to be an effective route to getting indexed, individual IRs may have additional SEO-related problems that must be addressed as well. Slow or misconfigured servers, failure to submit viable sitemaps, crawler errors that remain unresolved, failure to provide appropriate server response codes, lack of communication across the organization, and a host of other potential problems must be considered for effective SEO that will raise repositories' visibility in all search engine indexes. Advanced methods for optimizing PDF files may also help to assure inclusion in the GS index. More research and testing is needed, but it is fair to say that a crawler-friendly repository will fare much better in

GS than one that poses difficulties to crawlers. Upgrading to current repository software packages may help in this endeavor as product development teams become aware of and address SEO issues.

The growing use of GS by researchers underscores the need to address the problem of low IR indexing ratio. As the economic recession has tightened university budgets, more emphasis is being placed on assessment and measurement of outputs. IRs have the potential to raise author citation rates, and in turn to affect university rankings, but this potential may be seriously hampered if IR content is redundant or invisible to researchers who use GS.

Assisi, F.C., 2005. Anurag Acharya Helped Google's Scholarly Leap. *INDOlink - Science & Technology*. Available at: http://www.indolink.com/SciTech/fr010305-075445.php [Accessed October 13, 2011].

Beel, J., Gipp, B. & Wilde, E., 2010. Academic Search Engine Optimization. *Journal of Scholarly Publishing*, 41(2), pp.176-190.

Cahill, K. & Chalut, R., 2009. Optimal Results: What Libraries Need to Know About Google and Search Engine Optimization. *The Reference Librarian*, 50(3), pp.234-247.

Comscore, Inc., 2011. comScore Releases September 2011 U.S. Search Engine Rankings. *comScore Releases September 2011 U.S. Search Engine Rankings*. Available at: http://www.comscore.com/Press_Events/Press_Releases/2011/10/comScore_Releases_September_2011_U.S._Search_Engine_Rankings [Accessed October 22, 2011].

Dawson, A. & Hamilton, V., 2006. Optimising metadata to make high-value content more accessible to Google users. *Journal of Documentation*, 62, pp.307-327.

DCMI, 2005. Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata. *Dublin Core Metadata Initiative*. Available at: http://dublincore.org/documents/dc-citation-guidelines/ [Accessed October 26, 2011].

DeRidder, J.L., 2008. Googlizing a Digital Library. *The Code4Lib Journal*, (2). Available at: http://journal.code4lib.org/articles/43 [Accessed October 5, 2011].

DeRosa, C. & OCLC., 2005. *Perceptions of libraries and information resources : a report to the OCLC membership*, Dublin Ohio: OCLC Online Computer Library Center.

DeRosa, C. et al., 2010. *Perceptions of Libraries, 2010: Context and Community*, OCLC, Inc. Available at: http://www.oclc.org/reports/2010perceptions.htm [Accessed October 4, 2011].

Drewry, J.M., 2007. *Google Scholar, Windows Live Academic Search and beyond: A study of new tools and changing habits in ARL libraries*. University of North Carolina at Chapel Hill. Available at: http://etd.ils.unc.edu/dspace/handle/1901/429 [Accessed October 21, 2011].

Google, 2011. Google Webmaster Central. Available at: http://www.google.com/webmasters/ [Accessed October 29, 2011].

Google Scholar, 2010. Inclusion Guidelines for Webmasters. Available at: http://scholar.google.com/intl/en/scholar/inclusion.html [Accessed October 4, 2011].

Griffiths, J.R. & Brophy, P., 2005. Student searching behavior and the Web: use of academic resources and Google. *Library Trends*, Spring, pp.539-554.

Hagans, A., 2005. High Accessibility Is Effective Search Engine Optimization. *A List Apart*. Available at: http://www.alistapart.com/articles/accessibilityseo [Accessed October 4, 2011].

Haglund, L. & Olsson, P., 2008. The Impact on University Libraries of Changes in Information Behavior Among Academic Researchers: A Multiple Case Study. *The Journal of Academic Librarianship*, 34(1), pp.52-59.

Herrera, G., 2010. Google Scholar Users & User Behaviors: An Exploratory Study. *College & Research Libraries*. Available at: http://crl.acrl.org/content/early/2010/07/23/crl-125rl.abstract [Accessed October 4, 2011].

Key Perspectives Ltd. & Brown, S., 2009. *A comparative review of research assessment regimes in five countries and the role of libraries in the research assessment process: a pilot study commissioned by OCLC Research*, Dublin, Ohio: OCLC Research.

Kroll, S. & Forsman, R., 2010. *A slice of research life information support for research in the United States*, Dublin, Ohio: OCLC Research.

LaRock, T., 2010. Statistical Sampling for Verifying Database Backups. *simple-talk*. Available at: http://www.simple-talk.com/sql/database-administration/statistical-sampling-for-verifying-database-backups/ [Accessed December 10, 2011].

Malaga, R.A., 2008. Worst practices in search engine optimization. *Communications of the ACM*, 51(12), p.147.

McKay, D., 2007. Institutional Repositories and Their "Other" Users: Usability Beyond Authors. *Ariadne*, (52). Available at: http://www.ariadne.ac.uk/issue52/mckay/ [Accessed October 15, 2011].

Mikki, S., 2009. Google Scholar compared to Web of Science: A literature review. *Nordic Journal of Information Literacy in Higher Education*, 1(1), pp.41-51.

Mueller, J., 2008. Retiring support for OAI-PMH in Sitemaps. *Official Google Webmaster Central Blog:* Available at: http://googlewebmastercentral.blogspot.com/2008/04/retiring-support-for-oai-pmh-in.html [Accessed October 19, 2011].

OCLC, Inc., 2011. CONTENTdm Digital Collection Management Software. *CONTENTdm [OCLC - Digital Collection Services]*. Available at: http://www.oclc.org/contentdm/default.htm [Accessed October 27, 2011].

Rieger, O.Y., 2009. Search engine use behavior of students and faculty: user perceptions and implications for future research. *First Monday*, 14(12). Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2716/2385 [Accessed October 21, 2011].

Rushton, E.E., Kelehan, M.D. & Strong, M.A., 2008. Searching for a New Way to Reach Patrons: A Search Engine Optimization Pilot Project at Binghamton University Libraries. *Journal of Web Librarianship*, 2(4), pp.525-547.

Suber, P., 2004. Very Brief Introduction to Open Access. Available at: http://www.earlham.edu/~peters/fos/brief.htm [Accessed October 15, 2011].

The Times Higher Education, 2010. The Times Higher Education World University Rankings 2010-2011. Available at: http://www.timeshighereducation.co.uk/world-university-rankings/ [Accessed October 4, 2011].

University of Nottingham, 2011. OpenDOAR - Home Page - Directory of Open Access Repositories. Available at: http://opendoar.org/ [Accessed October 12, 2011].

Appendix: Highwire Press metadata mappings for 7 paper types

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 culture	Future leaders' views on organizational 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	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

Meta Tag	PhD	Masters
1 - citation_author	Rague, Brian William	Wu, Shangduan
2 - citation_date	2010/08	2010/07
3 - citation_title	A CS1 pedagogical approach to parallel thinking	Electronic structure and transport property of disordered graphene
8 - citation_firstpage	1	1
9 - citation_lastpage	234	84
13 - citation_keywords	Computer; CS1; Education; Parallel; Programming;	Disorder; Electronic structure; Graphene; Transport property; Electronic structure;
14 - citation_dissertation_institution	University of Utah, College of Engineering	University of Utah, College of Science
15 - citation_dissertation_name	PhD	MS
18 - citation_language	en	en
21 - citation_pdf_url	http://cdm6gs.lib.utah.edu/utils/getfile/collection/uspace/id/5/filename/19.pdf	http://cdm6gs.lib.utah.edu/utils/getfile/collection/uspace/id/0/filename/4.pdf
22 - citation_abstract_html_url	http://cdm6gs.lib.utah.edu/cdm/singleitem/collection/uspace/id/5/rec/1	http://cdm6gs.lib.utah.edu/cdm/singleitem/collection/uspace/id/0/rec/1

Not Relevant
4 - citation_publisher
5 - citation_journal_title
6 - citation_volume
7 - citation_issue
10 - citation_doi
11 - citation_issn
12 - citation_isbn
16 - citation_technical_report_institution
17 - citation_technical_report_number
19 - citation_conference_title
20 - citation_inbook_title

Meta Tag	Book Chapter	Book
1 - citation_author	Riloff, Ellen M.	Ram, Ashwin
2 - citation_date	1999	1999
3 - citation_title	Information extraction as a stepping stone toward story understanding	Understanding Language: Understanding Computational Models of Reading
4 - citation_publisher	MIT Press	MIT Press
8 - citation_firstpage	435	1
9 - citation_lastpage	460	519
12 - citation_isbn	0-262-18192-4	0-262-18192-4
13 - citation_keywords	Information extraction; Story understanding;	Information extraction; Story understanding;
18 - citation_language	en	en
20 - citation_inbook_title	Understanding Language: Understanding Computational Models of Reading	N/A
21 - citation_pdf_url	http://cdm6gs.lib.utah.edu/utils/getfile/collection/uspace/id/9/filename/5.pdf	
22 - citation_abstract_html_url	http://cdm6gs.lib.utah.edu/cdm/singleitem/collection/uspace/id/9/rec/1	

Not Relevant

5 - citation_journal_title
6 - citation_volume
7 - citation_issue
10 - citation_doi
11 - citation_issn
14 - citation_dissertation_institution
15 - citation_dissertation_name
16 - citation_technical_report_institution
17 - citation_technical_report_number
19 - citation_conference_title

Meta Tag	Working Paper
1 - citation_author	Wolfinger, Nicholas H.; McKeever, Matthew
2 - citation_date	2006-07-26
3 - citation_title	Thanks for nothing: changes in income and labor force participation for never-married mothers since 1982
6 - citation_volume	
7 - citation_issue	
8 - citation_firstpage	1
9 - citation_lastpage	43
10 - citation_doi	
13 - citation_keywords	Motherhood; Single Mothers; Income; Population surveys;
16 - citation_technical_report_institution	Institute of Public & International Affairs (IPIA), University of Utah
17 - citation_technical_report_number	2006-07-04
18 - citation_language	en
19 - citation_conference_title	101st American Sociological Association (ASA) Annual Meeting; 2006 Aug 11-14; Montreal, Canada
21 - citation_pdf_url	http://cdm6gs.lib.utah.edu/utils/getfile/collection/uspace/id/7/filename/21.pdf
22 - citation_abstract_html_url	http://cdm6gs.lib.utah.edu/cdm/singleitem/collection/uspace/id/7/rec/1

Not Relevant
4 - citation_publisher
5 - citation_journal_title
11 - citation_issn
12 - citation_isbn
14 - citation_dissertation_institution
15 - citation_dissertation_name
20 - citation_inbook_title