ibres volume 28, issue 1, pages 13-24 (2018)

Relationships Between Metadata Application and Downloads in an Institutional Repository of an American Law School

Hollie C. White
Libraries, Archives, Records, and Information Science (LARIS),
School of Media, Creative Arts, and Social Inquiry
Curtin University, Australia
hollie.white@curtin.edu.au

Sean Chen
Duke Law School,
Duke University, United States
schen@law.duke.edu

Guangya Liu
Duke Law School,
Duke University, United States
guangya.liu@law.duke.edu

ABSTRACT

Background. The Duke Law Scholarship Repository is a successful digital repository of an American law school, with over 1 million downloads per year. A series of studies were conducted to understand the relationship between metadata work and downloads.
Objective. The paper reports an analysis of the relationships between certain metadata elements and repository downloads.
Methods. Quantitative statistical methods, specifically correlation, t-test and multiple regression analysis, were used.
Results. Statistically significant relationships were found between download frequency and factors relating to abstract, co-authors, page count and discipline. Negative statistically significant relationships were found between download frequency and free text keywords, as well as controlled vocabulary subject terms.
Contributions. This study is an example of how in-use repository system administrators can demonstrate the impact of metadata work for institutional scholarly outreach. Also, this study adds another dimension to the keyword and searching/download literature that has been building since the 1970s.

INTRODUCTION

Digital repositories are one environment where information professionals create metadata resulting in the direct use of information objects. According to Zastrow (2013, p. 16), "one of the biggest challenges for digital professionals is finding the balance between providing

enough description to make an item findable but not so much that a job gets bogged down in time-consuming quantities of metadata". Like cataloging, metadata creation is time and resource consuming. Evaluating the effectiveness and value of metadata work is challenging for information professionals because, "current institutional repository software provides few tools for metadata librarians to understand and analyze their collections" (Nichols et. al, 2009, p. 230). In many cases, Google Analytics Solutions, download reports, and search logs are tools that provide insight into how repository collections are used.

For digital repositories, information retrieval shows use through download and visit numbers. These measures are called usage-based metrics and provide one way of evaluating how effective the metadata for a resource is in reaching a library repository's audience. Downloads and visits are outcome-based assessment—more concrete ways for libraries and their services to demonstrate value or success (Price & May 2011). According to Neylon and Wu (2009),

> *A simple way of measuring interest in a specific paper might be via usage and download statistics; for example, how many times a paper has been viewed or downloaded, how many unique users have shown an interest, or how long they lingered. This method can certainly provide a rapid means of assessing interest in a paper by comparing the trend in downloads and page views against the average.* (Neylon & Wu 2009, p. 3)

In order to understand more about how effective metadata application and workflows are in an already existing system, a series of small studies were conducted to examine how the presence of certain metadata relates to downloads.

This paper begins by reviewing the literature addressing repositories and metadata plus search engine optimization. Next, the study methodology begins with an introduction to the Duke Law Scholarship Repository, and continues with an explanation of how usage data are collected and statistical analysis used to determine the relationship between downloads and certain types of metadata. Analysis results related to the correlations, t-tests and multiple regression are presented. In addition, interpretations of what is suggested about the relationship between metadata workflow and repository metadata enhancement are presented.

LITERATURE REVIEW

Many literatures were consulted when conducting this research, but only two will be introduced in this paper due to limited space: studies connecting repositories and metadata, and the search engine optimization literature. Other relevant literatures are integrated as necessary in the Results and Analysis section.

Repositories and Metadata

In the early 2000s, the development of digital repositories managed by libraries created an opportunity for information professionals to move beyond its more passive role of information provider to be information distributor and publisher (Lynch 2003). Repositories are now used to publish open access journals, faculty-authored papers, student works, and a variety of other types of material to a worldwide online audience. Information describing each work, or object, is recorded in the repository system in order to facilitate information dissemination. Data describing basic digital item characteristics (like author or title) as well as subject metadata can support a number of activities including management, search, retrieval and evaluation.

Metadata is often referred to as "data about data" or "the sum total of what one can say about any information object at any level of aggregation" (Gilliland 2008). Using metadata often means describing an information object and using tools like controlled vocabularies or ontologies to elaborate on an item's "aboutness". The effectiveness of metadata for information retrieval within repository and other environments, specifically subject terms found in controlled vocabularies, has been long debated in the information science community. Metadata professionals claim that "metadata is an essential building block to facilitate effective resource discovery, access, and sharing across ever-growing distributed digital collections" (Park & Tosaka 2010, p. 104). In contrast, canonical research in information retrieval by Cleverdon (1970, 1984), Fidel (1992) and Rowley (1994) has shown that the use of controlled vocabularies and other indexing systems does not improve precision and recall results during information retrieval. Indexing researchers, such a Svenonius (1986), Hooper (1965), Leonard (1977), Reich and Biever (1991), and Sievert and Andrews (1991) have shown that, in situations outside of lab-based controlled research (as done by information retrieval researchers), controlled vocabularies can improve indexing consistency. Despite the mixed findings from decades of research, metadata standards and controlled vocabularies are persistently used in library-based systems such as repositories and online collections to this day. Yet, the value of metadata work is often questioned since it can be financial costly as well as time consuming.

Research specifically related to repository metadata has been limited (Barton, Currier & Hey, 2003; Park & Tosaka, 2010). Much of the literature on repositories and metadata address starting a repository (e.g., Wang, 2011; Rodgers & Sugarman, 2013) and how a certain domain uses metadata that could benefit repositories (e.g., White, 2013 & 2014). Park has been a key researcher in the area of repository metadata, studying both quality and creation (Park 2009, Park & Tosaka 2010). Park and Tosaka's (2010) research on metadata creation focused on the survey work of ALA librarians in 2008. In her literature review on metadata quality (Park, 2009), she pointed to three areas—accuracy, completeness, and consistency—as being the most agreed upon measures of good metadata. In contrast to Park's research, the study presented in this paper does not argue quality, but metadata effectiveness in terms of access to repository information. Though accessibility is an area of interest to metadata researchers like Bruce and Hillman (2004), there is in general a lack of research on how metadata affects access to information. This gap is filled by a handful of repository-based search engine optimization studies that looks more systematically at relationships between access and metadata.

Search Engine Optimization

Search engine optimization (SEO) is another area of focus for researchers interested in the convergence of metadata, repositories and access to information. According to Arlitsch and O'Brien (2013), SEO is "the practice of assuring that websites are indexed and effectively presented in Internet search results". No one (except those at Google) knows the complete workings of Google's ranking factors. In reality, when non-Google entities discuss SEO it must include caveats. For example, one research firm states that "the analysis and evaluation of Ranking Factors using this data not only has interpretation value, but in fact, represents a profound interpretation (and thus not a mere conjecture) on the basis of facts, namely the evaluation and categorization of website features that achieve high ranking in the search results" (Tober, Hennig, & Furch, 2013). While Google does share some information on the process, and outlines a set of best practices (Google, n.d.; Google Scholar, n.d.), only a

handful of research has been done by researchers and repository vendors that directly relates to repository work. Repository vendors often give general guidelines about SEO to clients, but declines to share specific information with clients in fear of giving unfair advantage to a single institution.

One set of researchers who have published multiple pieces on the convergence of search engine optimization and repositories are Arlitsch and O'Brien. They characterized the plight of repositories very simply: "[d]igital repositories of every type face a common challenge: having their content found by interested users in a crowded sea of information on the Internet" (Arlitsch & O'Brien, 2012, p. 64). Arlitsch and O'Brien's (2012) study used a content analysis method to determine how well repositories were indexed by Google Scholar. They found that all repositories, irrespective of system platform (ContentDM, Digital Commons, DSpace, EPrints, or Fedora), had low indexing ratios, with most performing under 60%, thus making them essentially invisible to Google Scholar. Institutional repositories are powerful tools with "the potential to raise author citation rates, and in turn to affect university rankings, but this potential may be hampered if IR content is redundant or invisible to researchers who use GS [Google Scholar]" (Arlitsch & O'Brien 2012). Their 2013 book goes on to outline the ways they increased their own repository's visibility in both Google and Google Scholar.

METHODOLOGY

This study investigated a method of using downloads to understand the effectiveness of metadata. The study methodology is based on usage metrics as discussed by Neylon and Wu (2009). According to O'Brien et al. (2016, p. 856), "[u]ltimately, citations may be the most valued measure of reuse and worth, and it is reasonable to expect publications to be downloaded and read before being cited. Using file download counts a metric for scholarly value is therefore crucial for IR assessment but it is a surprisingly difficult metric to measure accurately […]". The purpose of this study is to use statistical methods to evaluate which metadata features impact article downloads. Specific methods used in this study to collect and analyze Duke Law Repository usage statistics are discussed in detail below.

Study Environment

The Duke Law Scholarship Repository was established in 2005 using the ePrints platform focusing primarily on current Duke Law faculty publications in law reviews and journals. In 2009, Duke Law switched platforms to Digital Commons by migrating faculty publications into that system, which is optimized for both Google and Google Scholar indexing. In 2011, all nine of Duke Law's student-run journals were migrated from the Duke Law website into separate collections within Digital Commons. The Duke Law Scholarship repository reached 10,000 items around May 2013.

Two main collections compose the majority of the Duke Law School Scholarship Repository, faculty scholarship and student-run law journals. Duke Law uses what Chapman, Reynolds and Shreeves (2009) call a "mixed metadata environment", meaning that metadata is ingested into the repository through multiple workflows. Unlike other mixed metadata environments, Duke Law's ingest processes are carefully monitored and controlled by various Duke Law School's Information Services full-time staff members. Each collection has a separate workflow that has developed due to the nature of the scholarship being placed in the repository.

Faculty Scholarship

The faculty scholarship workflow focuses on adding individual pieces to the repository. An item is submitted for inclusion, researched for copyright information, and the PDF is prepared for ingest. Basic information about metadata is sent via email to experienced cataloging and acquisitions support staff for manual entering into the repository. Support staff enters descriptive and subject metadata into the Digital Commons backend using pre-designated metadata fields and local style guides. Until August 2014, subject metadata was assigned based on cataloger's judgment using free-text keywords. After 2014 the same staff member started using Library of Congress Subject Headings to assign keywords and stopped assigning free-text keywords in order to streamline the metadata workflow.

Journals

The journals collection workflow focuses on adding entire journal issues to the repository as the issues become available from the student editorial teams. Descriptive and subject metadata are entered in batch spreadsheet loads by website support staff. Subject metadata is pre-assigned to all articles based on the overall journal topic. The Duke Law Scholarship Repository and its collections has been discussed in more detail at various presentations and webinars (White 2012).

Data Collection

Data were collected two times. The first collection occurred in April 2013, and a second set of data was collected in April 2016. For 2013, a total of 9,692 repository entries were extracted from the Duke Law Scholarship Repository and placed into a Microsoft Excel spreadsheet. The metadata were then coded according to eight factors: Abstract presence, Abstract word count, Page count of PDF, Number of co-authors, Keyword presence, Number of keywords, Presence of subject categories/discipline, and Number of disciplines. These eight factors were then statistically analyzed in relation to repository-based PDF download counts.

The second data collection in April 2016 focused on keyword-based factors. Learning from the first experiment, keywords were mapped to Library of Congress Subject Headings (LCSH) and then these controlled terms were applied to all the existing metadata records from April 2013 to August 2014. Repository staff added controlled terms to all post-August 2014 new entries as articles were added. The data set collected in April 2016 was a comparison set looking at both pre- and post-August 2014 data. Instead of looking at number of articles, a sample of downloads were chosen as the number of observations with n=219,384.

Data Analysis

Two rounds of statistical analyses were run using the Stata 12 statistical software package. For the first data set analysed in 2013 (later checked and re-analyzed in 2015), a series of Pearson product-moment correlation coefficients were computed to assess the relationship between downloads and the eight factors. Multiple regression analysis was also used to test if each of the eight factors significantly predicted repository-based downloads over and above the effect of the other factors. These factors were analysed in three sections, by using data from (a) the entire repository, (b) the faculty scholarship collection, and (c) the Duke Law student-run journal collection.

For the second data set collected in 2016, pre-post t-tests with keyword type as the independent variable and number of downloads as the dependent variable, were run on the

219,384 download-based sample. The analysis looked specifically at keyword-based factors in relation to the entire collection. Multiple regression analysis was also run. Since there was a change in the assignment of metadata keywords/controlled-terms in August 2016, the date August 2014 was used as a pre-post comparison point. T-tests and regression analysis were run on the new data set from August 2013 to August 2014, in comparison to post-keyword change data from August 2014 to August 2015.

The research team changed the way the data set was analyzed in order to get a better picture of how the sample records changed over time. As mentioned later in the study limitations section, since the repository database was part of an "in-use" system, comparison of the same type of data pre- and post-change is important to understanding how the type of keyword/controlled-term impacts downloads.

ANALYSIS AND RESULTS

The following section reports analysis results of both the 2013 and the 2016 data sets. Overall results for 2013 are reported first, and then individual metadata factor results are discussed in more detail. The Keyword result section discusses results for the 2013 and the 2016 data sets.

Overall Results for 2013

Table 1 shows:

- correlation (Pearson r) between each factor and download counts, for the entire repository data set of 9,692 entries;
- t-statistics for individual factors from the multiple regression results, using the entire repository data set of 9692 entries ($F(8, 9683)=177$; $R^2= 0.13$);
- t-statistics from the multiple regression using the faculty scholarship collection subset of 2,405 entries ($F(8, 2396)=32$; $R^2= 0.10$);
- t-statistics from the multiple regression using the journals collection subset of 7,282 entries ($F(8, 7273)=148$; $R^2= 0.14$).

The correlation results (Pearson r values in column 2) indicate that downloads were significantly correlated with seven of the eight factors ($p<.01$), including presence of abstract, abstract word count, page count, number of co-authors, presence of discipline, and number of disciplines. In addition, presence of keywords was found to have a significant relationship with download counts.

The regression results for the entire repository data set indicate that three factors—*number of co-authors, presence of keywords*, and *number of keywords*—had negative significant relationships with downloads. *Number of disciplines* was found to be non-significant at the entire repository level.

The regression results for the faculty scholarship collection subset indicate a statistically significant relationship with downloads for five of the eight factors. *Presence of keywords* and *number of keywords* were found to have negative significant relationships with downloads. *Number of co-authors* and *number of disciplines* were found to be non-significant for the faculty scholarship collection.

Regression results for the journal collection indicate a statistically significant relationship with downloads for five out of the eight factors. *Number of co-authors* was the only negative significant factor for this collection. *Presence of abstract* and *presence of disciplines* were found to be not significant.

Table 1. Correlation with download counts, and t-statistic from multiple regression analysis for the 2013 Data Set

Factor	All Repository *Pearson r* (n=9,692)	All Repository *t-statistic* (n=9,692)	Faculty Scholarship Collection *t*-statistic (n=2405)	Journals Collection *t*-statistic (n=7,291)
Presence of abstract	0.2664***	9.89***	4.62***	-0.62
Abstract word count	0.2822***	8.85***	3.63***	9.68***
Page count	0.1455***	9.62***	6.39***	6.59***
Number of co-authors	0.0166	-10.97***	-0.35	-5.03***
Presence of keywords	0.319**	-2.69**	-2.48*	5.21***
Number of keywords	-0.1357***	-4.10***	-3.32**	1.97*
Presence of discipline	0.2153***	8.76***	3.93***	1.12
Number of disciplines	0.1924***	-0.54	-1.71	13.85***

*** $p<.001$, ** $p<.01$, *$p<.05$

Abstract-Related Results

Understanding more about abstracts and how effective they are in the repository for providing access to articles was the original impetus for the research. Both the presence and word count of abstract were statistically significant in the correlation analysis. From the regression analysis results, all three collections showed a positive statistically significant relationship between word count and downloads. This indicates that longer abstracts lead to more downloads.

The presence of an abstract was significant for the faculty scholarship collection, but not for the journals collection. This indicates that for the journals collection, presence of abstract doesn't add any extra weight to word count of abstract. A non-zero word count already implies the presence of an abstract.

Discipline-Related Results

Disciplines function differently than *keywords*. Both can enhance the "aboutness" of a piece using a word or phrase. *Disciplines* as a metadata field has a categorization function. According to Jacob (2004 p. 518), "[c]ategorization is the process of dividing the world into groups of entities whose members are in some way similar to each other". Instead of relating the unique characteristics of a piece, as with *keywords*, *disciplines* focuses on those characteristics that bring multiple articles together. *Disciplines* further promotes this linking by connecting all Digital Commons repositories together based on subject via the Digital Commons network (http://network.bepress.com/), which was first released in October 2012.

Results for disciplines are quite contradictory. As indicated in Table 1, the correlation results indicate positive statistically significant relationships for both discipline presence and number of disciplines in relation to downloads. Regression results are mixed, with presence statistically significant for the entire repository ($t=8.76$, $p<.001$) and faculty scholarship ($t=-3.93$, $p<.001$), but not journals ($t=1.12$, $p=.264$). Questions can be raised about why disciplines appear to be effective in promoting downloads in some cases and not others.

Page Count and Co-Author-Related Results

Page count was the one factor that was statistically significant in all three regressions. Since all papers in the Duke Law Scholarship Repository are made into optical character recognition pdfs before being uploaded, this finding only confirms what is already known about Google page rankings (Webmaster FAQ, n.d.). The longer a paper, the more words are available for a full text search. More words increase the likelihood of a paper being picked up by a search and subsequently downloaded (Tober, Henning, & Furch, 2013). The fact that these pdfs are linked from a higher education (.edu) website only increased the likelihood of downloads as well.

Co-authors was another factor that was analyzed in relationship to downloads. The results for this factor were mixed, based on the collection analyzed. Many bibliometric studies have examined the relationship between co-authorship and h-index rankings (Schreiber 2009). The researchers were also interested in how co-authorship numbers relate to downloads. In the faculty scholarship collection, co-authorship was found to be not statistically significant. For the journals collection, co-authors were negatively statistically significant ($t=-5.03$, $p<001$). More research is needed to understand when and how the number of co-authors contribute to downloads.

Keyword-Related Results

The most compelling results from the initial 2013 data set were related to keywords, and it was those results that led to repository metadata revisions and a second study.

In this study, keywords as a factor in contributing to downloads had varying results. As mentioned previously, keywords were studied in terms of presence and by number (when present). Initial correlation results found both keyword presence and number to be non-significant. Regression results indicate both significant and not significant results depending on the repository collection being evaluated.

Results suggest that in most cases, including keywords, no matter the number, either has a negative impact on downloads or no impact at all. The one exception to this is with the journal's collection. Keyword assignment within the journals was limited to the *Duke Law and Technology Review* and the *Duke Journal of Constitutional Law & Public Policy* for the main purpose of assisting with website organization. Applications of terms was constrained by functional requirements which used keyword terms to provide collocation functionality for users of each journal website. This suggests a few possibilities:

- Option 1: Free text keywords are not adequately describing the resources in the faculty scholarship collection. A more controlled method (like that used for the journals) should be applied to all of repository collections in order to increase downloads.
- Option 2: Free text keywords are applied well in the faculty scholarship collection and their presence allows users to know what they do and do not want to download and read. Therefore, the keywords present let user know they do not want to download items that are not relevant to them.

- Option 3: All keywords are a waste of resources and keyword application should cease all together.

Arlitsch and O'Brien (2013) warned against using controlled vocabularies, using many of the arguments also presented by folksonomy enthusiasts. Instead, they suggested working with Google Keyword Tool in order to examine how users find a given site. Yet, this is in conflict to an earlier statement about the power of linked data vocabularies. They recommended making available only those resources where you can give robust descriptions. This points to the need to use vocabularies (particularly those used with RDF and linked data) in order to give the appropriate contexts to chosen keywords.

With this in mind it was determined that a second study needed to be conducted in order to narrow down the options previously presented. Option 1 was chosen to investigate. From initial data analysis in 2013 until August 2014, free text keywords were mapped to Library of Congress Subject Headings by a graduate student in the Master's program in information studies. Mappings were supervised and verified by Duke Law's Digital Resources Librarian prior to implementation into the repository system. From 2014 to 2016, data were collected about downloads using the controlled vocabulary terms and not free-text keywords. In April 2016, new pre/post t-tests and regression analysis were run using the August 2014 date as the division marker for the keyword application change. As mentioned in the Methods section, a new dataset was pulled to compare the differences based on the same data set, but the new data set focused on download level data as opposed to article level data. Since the Duke Law Scholarship Repository experiences over 1 million downloads a year, a sample of records was used with 219,384 observations used in the dataset.

The follow-up study's pre/post t-test results ($t=-35$, $p<.001$) and regression results ($F(7, 219{,}376)=569$; $R^2= 0.02$) resulted in negative significant relationships between presence of controlled vocabulary-based keywords and downloads. These results confirmed that both free-text keyword (i.e. tagging/folksonomy) and controlled vocabulary terms both resulted in a negative significant relationship with downloads. Therefore, it was concluded that the presence of any type of keyword, within the Duke Law Scholarship Repository context, was not linked to encouraging downloads.

Metrics are a way to show the value of library work to funding agencies and administrators (White 2017). Once dismissed, downloads are an increasingly important measure of user engagement with repository content (O'Brien et al. 2016). Keywords results from both of these Duke Repository-based studies indicate that subject term application is not essential for getting users engaged with repository content. At the local level, these results start a new round of questions. Some of those questions include, if the Duke Law Repository were to start over, would it include any type of keyword? Since the repository has over a decade of existence, should keywords be discontinued? Or should all keywords be removed entirely? After much questioning and discussion, the Duke Law Repository Staff has decided to keep applying controlled keywords because the group believe in the value of subject term application beyond downloads. Using controlled terms help streamline Duke Law staff workflow. Core to this decision is a belief that metadata is best applied in anticipation of the systems to come, as opposed to those currently in existence. Also, a reliance on strong descriptive and subject metadata is a best practice approach for system migration.

The results from these Duke-based studies are surprising, especially compared to results from researchers like Gross, Taylor, and Joudrey (2015) who looked at data from both 2005 and 2015 showing that controlled vocabulary use persisted in contributing to downloads. These Duke Repository keyword results add one more study to a long-standing debate in the

information science literature as to the value of keywords as a tool to enhance information retrieval results (Cleverdon, 1970; Svenonius, 1986; Fidel, 1992; Rowley, 1994; Yang, 2016).

STUDY LIMITATIONS

As with any study, there are weaknesses to the approaches discussed in this paper. True quantitative analysis works best in controlled laboratory environments, while these studies were conducted in a live, working repository. Statistical methods may point to relationships, but not necessarily show true causation. For libraries, it has been confirmed, "statistical measures alone are not enough to prove that libraries are a valuable asset" (Price & Fleming-May, 2011, p. 196). The same is true for repositories. Usage-based measures, such as downloads and visits, show one part of the much larger picture of repository success. The statistical methods used in this study are just one way to examine metadata and its value within the Duke Law Scholarship Repository. Other methods, both traditional (like citation analysis and h-index) as well as Altmetrics could give more insight into repository success within certain contexts (Konkiel & Scherer, 2013). In addition, the case study environment (of the Duke Law Scholarship Repository) is limited and not easily generalizable.

CONCLUSION

Results from this study suggest that certain metadata practices contribute to more downloads depending on collection. Each collection has its own metadata strengths in terms of downloads. The results also suggest that each collection is found by users due to different interests. Download counts to metadata relationships suggest that the "author" of a piece is the most important download factor for items in the Faculty Scholarship collection, while journal collections papers are found more by keyword/topical interest.

Abstract presence is most effective in the faculty scholarship collection which relies heavily on author-created metadata. Authors while not depositing the metadata themselves are creating the metadata that is used to populate various metadata fields, especially the abstract field. The journals collection is populated with information provided by journal editors and boards. At this time, not all of the Duke Law student-run journals require abstracts from their authors. Suggestions will be made to the journals about requesting abstracts.

Some subject metadata-related results, specifically keywords and disciplines application, indicate that the more systematic and controlled approach as used in the journal collections is more successful in contributing to downloads. Follow-up studies on keywords indicate that their presence does not contribute to increased downloads, which calls into question whether adding any keyword is necessary in the repository environment. This study hopes to show other repositories how statistical analyses, like correlations and regression analysis, can be used to show the relationship between metadata and downloads. Currently, the study results cannot be generalized beyond the duke law scholarship repository. Yet, similar methods employed in other repository environments could show how metadata application workflows have an impact on access and outreach of scholarship.

REFERENCES

Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google. *Library Hi-Tech, 30*(1), 60-81.

Arlitsch, K., & O'Brien, P. S. (2013). *Improving the visibility and use of digital repositories through SEO: A LITA guide*. Chicago: ALA TechSource.

Barton, J., Currier, S., & Hey, J. M. N. (2003). Building quality assurance into metadata creation: An analysis based on the learning objects and e-prints communities of practice. In *Dublin Core Conference 2003 (DC-2003): Support Communities of Discourse and Practice—Metadata Research and Applications, 28 Sep-02 Oct 2003, Seattle, USA*. Dublin Core Metadata Initiative. Retrieved from http://dcpapers.dublincore.org/pubs/issue/view/26

Bruce, T. R., & Hillmann, D.I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. I. Hillman & E. L. Westbrooks (Eds.), *Metadata in practice* (pp. 238-256). Chicago: American Library Association.

Chapman, J.W., Reynolds, D., & Shreeves, S. (2009). Repository metadata: Approaches and challenges. *Cataloging & Classification Quarterly, 47*(3-4), 309-325. doi:10.1080/01639370920735020

Cleverdon, C. W. (1970*). The effect of variations in relevance assessments in comparative experimental tests of index languages*. Cranfield, UK: Cranfield Institute of Technology.

Cleverdon, C. W. (1984). Optimizing convenient online access to bibliographic databases. *Information Services and Use, 4*(1-2), 37-47.

Fidel, R. (1992). Who needs a controlled vocabulary? *Special Libraries*, *83*(1), 1-9.

Gilliland, A. J. (2008). Setting the stage. In Murtha Baca (Ed.), *Introduction to metadata: Pathways to digital information* [online edition, version 3.0]. Retrieved from http://www.getty.edu/research/publications/electronic_publications/intrometadata/

Google. (n.d). Inclusion guidelines for webmasters. Retrieved from https://scholar.google.com/intl/en/scholar/inclusion.html

Google Scholar. (n.d.). Webmaster FAQ. Retrieved from https://support.google.com/webmasters/answer/1050724

Gross, T., Taylor, A.G., & Joudrey, D.N. (2015). Still a lot to lose: The role of controlled vocabulary in keyword searching. *Cataloging & Classification Quarterly, 53*(1), 1-39. http://dx.doi.org/10.1080/01639374.2014.917447

Hooper, R.S. (1965). *Indexer consistency tests—Origin, measurements, results and utilization*. Bethesda, MD: IBM.

Jacob, E. K. (2004). Classification and categorization: A difference that makes a difference. *Library Trends, 52*(3), 515-540.

Konkiel, S., & Scherer, D. (2013). New opportunities for repositories in the age of altmetrics. *Bulletin of the Association for Information Science and Technology, 39*(4), 22-26.

Leonard, L. E. (1977). Inter-indexing consistency studies, 1954-1975: A review of the literature and summary of study results. *Occasional Papers* (University of Illinois at Urbana-Champaign. Graduate School of Library Science), no. 131.

Lynch, C. A. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *ARL Bimonthly Report, 220*, 1-7.

Neylon, C., & Wu, S. (2009). Article-level metrics and the evolution of scientific impact. *PLoS Biology, 7*(11). https://doi.org/10.1371/journal.pbio.1000242

Nichols, D.M., Paynter, G.M., Chan, C., Bainbridge, D., McKay, D., Twidale, M. B., & Blandford, A. (2009). Experiences in deploying metadata analysis tools for institutional repositories. *Cataloging and Classification Quarterly, 47*(3-4), 229-248. doi:10.1080/01639370902737281

O'Brien, P., Arlistch, K., Sterman, L., Mixter, J., Wheeler, J., & Borda, S. (2016). Undercounting file downloads from institutional repositories. *Journal of Library Administration, 56*(7), 854-874.

Park, J. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly, 47*(3-4), 213-228. doi:10.1080/01639370902737240

Park, J., & Tosaka, Y. (2010). Metadata creation practices in digital repositories and collections: Schemata, selection criteria, and interoperability. *Information Technology and Libraries, 29*(3), 104-116.

Price, A. N., & Fleming-May, R. (2011). Downloads or outcomes? Measuring and communicating the contributions of library resources to faculty and student success. *The Serials Librarian, 61*(2), 196-199.

Reich, P., & Biever, E. J. (1991). Indexing consistency: The input/output function of thesauri. *College & Research Libraries, 52*(4), 336-42.

Rodgers, J. R., & Sugarman, T. (2013). Library technical services: Key ingredients in the recipe for a successful institutional repository. *The Serials Librarian, 65*(1), 80-86.

Rowley, J. (1994). The controlled versus natural indexing language debate revisited: A perspective on information retrieval practice and research. *Journal of Information Science, 20*(2), 108–119.

Schreiber, M. (2009). A case study of the modified Hirsch index accounting for multiple coauthors. *Journal of American Society for Information Science & Technology, 60*(6), 1274-1282.

Sievert, M. C., & Andrews, M. J. (1991). Indexing consistency in information science abstracts. *Journal of the American Society for Information Science, 42*(1), 1-6.

Svenonius, E., (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science, 37*(5), 331–340.

Tober, M., Hennig, L., & Furch, D. (2013). SEO ranking factors—rank correlation 2013—Google USA. Retrieved from http://www.searchmetrics.com/media/documents/ranking-faktoren/download-ranking-factor-study-2013.pdf

Wang, F. (2011). Building an open source institutional repository at a small law school library: Is it realistic or unattainable? *Information Technology and Libraries, 30*(2), 81-84.

White, H. (2012). Duke presents: Institutional repositories and law reviews [Bepress Webinar, November 1, 2012]. Retrieved from https://www.bepress.com/webinar/duke-presents-institutional-repositories-law-reviews/

White, H. (2013). Examining scientific vocabulary: Mapping controlled vocabularies with free text keywords. *Cataloging & Classification Quarterly, 51*(6), 655-674.

White, H. (2014). Descriptive metadata for scientific data repositories: A comparison of information scientist and scientist organizing behaviors. *Journal of Library Metadata, 14*(1), 24-51.

White, H. (2017). Building a repository metrics program to enhance the value of library services. *AALL Spectrum, 21*(6), 26-29.

Yang, L. (2016). Metadata effectiveness in Internet discovery: An analysis of digital collection metadata elements and internet search engine keywords. *College & Research Libraries, 77*(1), 7-19.

Zastrow, J. (2013). Digital changes everything: The intersection of libraries and archives. *Information Today, 33*(9), 16-18.