2015

Availability and accessibility in an open access institutional repository: A case study

Jongwook Lee, Gary Burnett, Jung Hoon Baeg, Micah Vandegrift, and Richard Jack Morris

Availability and accessibility in an open access institutional repository: a case study

Abstract

Introduction. This study explores the extent to which an institutional repository (IR) makes papers available and accessible on the open web by using 170 journal articles housed in DigiNole Commons, the IR at Florida State University.
Method. To analyze the IR's impact on availability and accessibility, we conducted independent known-item title searches on both Google and Google Scholar (GS) to search for faculty publications housed in DigiNole Commons.
Analysis. The extent to which the IR makes articles available and accessible was measured quantitatively, and the findings that cannot be summarized with numbers were analyzed qualitatively.
Results. Google and GS searches provided links to DigiNole metadata for a total of 145 (85.3%) of 170 items, and to full texts for 96 (96%) of 100 items. With one exception, access to either metadata or full text required no more than three clicks.
Conclusions. Overall, the results confirm the contribution of the IR in making papers available and accessible. The results also reveal some impediments to the success of OA: including impediments linked to contractual arrangements between authors and publishers, impediments linked to policies, practices, and technologies governing the IR itself, and the low level of faculty participation in the IR.

Introduction

Open access (OA) refers to a variety of approaches for making the products of scholarly research freely available for others to access and, in some cases, reuse. Some authors argue OA increases the accessibility of research papers by reducing or eliminating restrictions on access caused by the licensing requirements and copyright agreements common in traditional subscription journals (Chan, 2004; Harnad et al., 2008). Arguments have been forwarded for a move to OA on financial grounds, as rapidly increasing prices for journal subscriptions may result in restrictions to readership if libraries reduce the number of their subscriptions, (see, for example, Parks, 2002). Further, many researchers have claimed that OA provides a "citation advantage" over traditional publication models (Antelman, 2004; Eysenbach, 2006; Harnad & Brody, 2004; Norris, Oppenheim, & Rowland, 2008), because it typically makes research papers available via the Web, increasing their "citeability" (Gargouri et al., 2010; Lawrence, 2001).

Generally speaking, there are two primary OA models currently in use:

1. The "gold OA" model in which papers are published in open access journals, often (though not always) with associated article charges paid by authors, and
2. The "green OA" model, in which authors themselves archive their work in repositories, personal websites, or elsewhere (Bailey, 2010).

A third type, sometimes called the "hybrid journal," allows authors to pay fees to make individual articles originally published in traditional subscription journals freely available (Miguel, Chinchilla-Rodriguez, & Moya-Anegon, 2011). Of the models, green OA has typically been considered to be the most effective method as a result of an increasing number of repositories and journals that allow author self-archiving (Harnad et al., 2008). For instance, Miguel et al. (2011) found that, of journals indexed in SCOPUS, 32% followed the green model by contractually allowing authors to self-archive, whereas only 9% could be considered to be true gold OA journals.

Authors can self-archive their papers by uploading them to their own personal websites or by depositing them in discipline-specific, government-sponsored, or institutional repositories (IRs). A tradition of exchanging preprints among researchers in several science fields has given rise to numerous disciplinary repositories (such as arXiv and PubMed Central), which have often been developed and maintained by the communities of researchers who use them (Björk, 2004). IRs, by contrast, can be defined as "digital collections capturing and preserving the intellectual output of a single or multi-university community" (Crow, 2002, p. 1). Typically, such IRs are housed by university libraries (Björk, 2004), which, by adapting traditional collection development practices, can systematically manage and maintain IR materials for the long term (Björk, 2004); as a result, the green approach of author self-archiving in IRs has been widely recommended (Gargouri et al., 2010). This practice may have several advantages; it may, for instance, help to:

1. facilitate scholarly communication,
2. increase the visibility of institutions,
3. mitigate the monopoly power of publishers, and
4. support the teaching, research, and administrative missions of universities (Crow, 2002; Markland, 2006; McCord, 2003).

Claims about the value of IRs, like those about OA in general, are rooted in an assumption that they inherently enhance the accessibility – and, as a result, help to increase citation counts – of the materials they house. Consequently, much effort has been expended to

increase faculty participation, which is typically low, in their institutions' IRs (Chan, 2004; Davis & Connolly, 2007; Oguz & Assefa, 2014; Swan & Brown, 2005). However, few researchers have tested this basic assumption.

This case study addresses this dearth of research, exploring the extent to which OA makes articles accessible on the web by examination of a particular institutional repository. The authors proposed the following research questions:

1. How effective is an IR in making articles accessible?
2. What, if any, are the potential impediments to the effectiveness of an IR in furthering OA goals?

Many researchers have associated the citation advantage of OA with its function of improving the accessibility of papers on the web (Antelman, 2004; Eysenbach, 2006; Harnad & Brody, 2004; Lawrence, 2001). However, such studies typically do not provide empirical support for such a claim. In this study, the accessibility of materials housed in an IR is tested using Google and Google Scholar (GS) searches. Since the mere availability of a paper (i.e., the mere fact that it is present within an IR) does not necessarily guarantee that it is easily accessible, the study differentiates the concepts of availability and accessibility as dimensions of physical access: we examine availability as the ability of search engines to retrieve clear links to an individual paper within the first two pages of results, and, further, measure accessibility as the number of clicks required for a user to navigate from those results to the full text of the paper itself. The study provides empirical findings for scholarly communication researchers and librarians who are interested in the promoting the success of OA and IRs. The result of this study can be used as a source for encouraging the OA movement and for enhancing the performance of IR software.

Literature Review

OA citation advantage. Numerous previous studies have attempted to explain the causes of the OA citation advantage through three postulates (Craig, Plume, McVeigh, Pringle, & Amin, 2007; Davis & Fromerth, 2007; Kurtz et al. 2005; Koler-Povh, Južnič, & Turk, 2014; Xia & Nakanishi, 2012). The first (the "OA postulate") suggests that OA increases citation count by directly improving the accessibility of papers. On the other hand, the second and third postulates ("early access" and "selection bias") explicitly reject the assumptions inherent in the OA postulate. The early access postulate proposes that papers are more likely to be cited because OA papers are often made public in early pre-print versions, and are thus accessible for a longer time than non-OA papers. Similarly, the selection bias postulate argues that authors tend to favor their highest quality – and, thus, most likely to be cited – work when choosing materials to make available in IRs.

Lawrence (2001) examined the correlation between paper availability on the web and citation count, analyzing 119,924 conference papers in computer science and related disciplines, finding that papers on the web are more likely to be cited. Harnad and Brody (2004) compared the citation counts of articles published in a non-OA journal that had been placed by authors into IRs with those of articles from the same journal that had not been so placed. They found that, in the fields of computer science, astronomy, and physics, the citation rates of OA articles were 2.5-2.8 times higher than those of non-OA articles.

Antelman (2004) tested a hypothesis that citation counts of OA articles were higher than those of non-OA articles, choosing ten journals each from four disciplines whose practitioners are known to be heavy users of pre-prints (mathematics, electrical and electronic engineering,

political science, and philosophy). Using Google to distinguish freely available full-text OA articles from non-OA articles, the study found that citation counts of OA articles were 51% to 91% higher than those of non-OA articles. Eysenbach (2006), in a longitudinal study examining the impact of OA as well as article and author characteristics on citation rates found OA to be a significant independent predictor. In addition, OA articles tend to be cited earlier and more often than non-OA articles, even when there is no significant difference in the quality of articles.

Some researchers have considered all three postulates. For example, Kurtz et al. (2005) measured the effects of OA, early access, and self-selection bias on citations in seven astronomy journals. Comparing citation changes for older and newer articles to test the OA and early access postulates, and testing selection bias by using the Monte Carlo simulation to analyze the "probability that a particular number of non-arXiv submitted papers [would] be [among] the top 100 or 200 most cited papers," (p. 1398), they found that, while early access and selection bias strongly influence citations, the effect of OA itself is unobservable, perhaps because the astronomy research community typically has easy access to core journals. Davis and Fromerth (2007) analyzed 2,765 articles published in four mathematics journals, finding that articles deposited in arXiv tend to have more citations than non-deposited articles. However, while they reported a positive impact on citations due to selection bias, they detected an OA effect only among highly cited articles and no impact from early access.

Moed (2007) investigated the citation impact of articles deposited in arXiv with that of articles not deposited, using citation time windows to measure the effects of early access, and analyzing the proportion of prominent authors in arXiv to test selection bias. Although the study found an increase in citation counts for papers available in arXiv, this was due to early access and selection bias rather than to the impact of OA *per se*; as Moed (2007) put it, arXiv increases citation counts not because it makes papers freely accessible, but because it makes them "available earlier" (p. 2054). Davis, Lewenstein, Simon, Booth, and Connolly (2008) carried out a randomized controlled experiment to measure the OA effect on downloads and citations, finding that OA articles are downloaded more often than non-OA articles, with a strong impact from article characteristics such as article type and length, etc. While their study suggests that OA augments readership through increased downloads, there is no evidence of a true OA citation advantage.

Status of IRs. Early work by Crow (2002) suggested that IRs could be seen as contributing factors in "a new disaggregated model" (p. 6) of scholarly publishing, one that may help to weaken the monopolistic power of the traditional academic journal system over scholarly communication. Through developing and maintaining "institutionally defined," "scholarly," "cumulative and perpetual," and "open and interoperable" IRs (p. 16), he argues that institutions can increase their visibility and prestige by centralizing the intellectual work of their members, thus enabling researchers to find relevant materials more easily. Shearer (2002) identified potential factors that need to be considered for IRs to be successful, including "input activity," "disciplines," "advocacy activities," "archiving policies," "copyright policies," "content type," "staff support," "quality control policies," "software," and "use" (pp. 98-99). Shearer assumed that the input activity – that is, submission of papers by researchers – would be one of the most important factors, and wanted to see the relationship between it and other factors. This 2002 study, however, did not provide the results based on the analysis of data.

Markland (2006) examined the effectiveness of Google in retrieving papers deposited in IRs, choosing one item each from 26 UK institutional repositories, checking their availability and investigating the ease of finding them through five search strategies ("a search at the

repository interface," "a Google search using a keyword or phrase from the title," "a Google search using the complete title," "a Google Scholar search using a keyword or phrase from the title," and "a Google Scholar search using the complete title" (p. 224)). The study showed that 3 of the items could not be retrieved via the repository interface. For results of searches from Google and Google Scholar using keyword phrases from titles, 17 of 26 items in repositories were retrieved from Google, and 8 of 26 from Google Scholar. When using a complete title search, 25 of 26 were retrieved via Google and 17 of 26 via Google Scholar, suggesting that a simple title search via Google was the most effective means of retrieving repository items.

Some researchers have reported low awareness and usage of IRs. Swan and Brown (2005) examined the perceptions of OA and self-archiving in a survey of 1,296 researchers. While 49% of respondents had self-archived their papers in repositories or websites, the remainder had not. Of those who had not yet self-archived, 71% were unaware of OA and self-archiving. In an evaluative study of IRs, Davis and Connolly (2007) collected data from Cornell's DSpace in order to calculate descriptive statistics and interviewed eleven faculty members for a deeper understanding of their attitudes and behaviors. DSpace had 2,646 items as of October, 2006 categorized into 196 collections, of which almost 30% contained no materials. Further, of 519 unique contributors, nearly 50% uploaded only a single item, reinforcing the interview finding that faculty members lacked both knowledge and motivation to use IRs. In a study of attitudes and behaviors, Watson (2007) interviewed 21 researchers from Cranfield University. Interviewees considered it important to share their work, but most were not aware of the potential of IRs as a way to do so; even among those who knew of the existence of IRs, many were not using them. Xia (2010) found researchers to be increasingly aware of OA but only at a very basic level, with insufficient understanding to enable them to participate in OA initiatives, suggesting that increased awareness alone may not be sufficient to increase faculty use of IRs.

More recently, Nicholas, Rowlands, Watkinson, Brown, and Jamali (2012) investigated the perceptions of scientific researchers on digital repositories. They analyzed 1,685 survey responses obtained from faculty members and students who had been registered in the Institute of Physics Publishing. They found that 1,079 (63.7%) of survey respondents had deposited their research outcomes in some kind of repository, and that 44.1% had specifically used IRs. Oguz and Assefa (2014) surveyed faculty members at a medium-sized university to investigate their perceptions and attitudes toward IRs and found positive perceptions among 52.9% and negative perceptions among 47.1%. In general, although there are some variations across disciplines and institutions (Cullen, & Chawner, 2011; Oguz & Assefa, 2014), there appears to be a growing rate of author participation in IRs, but there is still plenty of room for further growth (Björk, Laako, Welling, & Paetau, 2014).

Dimensions of Physical Access: Availability and Accessibility

In a traditional brick-and-mortar library, the mere presence of an item in the collection does not guarantee full accessibility; for instance, a book may be available, but may be shelved on a top shelf, with the result that there are important barriers to the accessibility of that item for users in wheelchairs; in online settings – such as IRs – there may be comparable impediments limiting the accessibility of items that are present (and, thus, available) in a collection. Therefore, in this study we treat *availability* as a necessary, but not sufficient, element of *accessibility* because the mere presence of papers within an IR does not guarantee their accessibility (Hargittai & Hinnant, 2006); this allows us to identify possible impediments to accessing documents housed in IRs. Like Fidel and Green (2004), we consider *availability* to be a dimension of

accessibility, arguing that *availability,* while a necessary precondition for users to gain access to and "use a source at a particular time" (p. 577), does not ensure that users will easily be able to put their hands on that source. Ugah (2008), similarly, defines availability as the presence and readiness for use of materials in libraries or virtually; a source is unavailable if it lacks either physical presence or readiness for use. The accessibility of materials, following such a definition, depends upon prior availability, simply because unavailable sources are also inaccessible.

The term *accessibility*, however, has been treated much more broadly, with focus on the physical, intellectual, and social aspects (Culnan, 1985; Burnett, Jaeger, & Thompson, 2008; Fidel & Green, 2004). Physical access is often defined in terms of users' ability to reach – or put their hands on – and make use of available materials; in other words, the concept of availability is often seen as a necessary dimension or precondition of physical access. Beyond physical accessibility, intellectual access focuses on human cognitive factors important in seeking and understanding information and social access has to do with a variety of social factors – including social norms, political and legal matters, etc. – related to information access. As Burnett, Jaeger and Thompson (2008) suggest, physical access is the prerequisite for other types of access, just as availability is a precondition of physical accessibility; in this study, however, we focus only on the first form of accessibility, physical accessibility, in order to investigate possible impediments that may complicate or block users' ability to "put their hands on" materials housed – and, thus, available – in IRs. Further impediments related to either intellectual or social access may be relevant for the use of materials in IRs; in the current study, however, we do not investigate such other possible impediments.

Many OA studies have used the terms *availability* and *accessibility* interchangeably or have reported a positive relationship among them, suggesting that increased availability can help to improve accessibility. The *Budapest Open Access Initiative* (2002) defines OA as "free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of … articles…." Similarly, Bullinger, et al. (2003), in the *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, describe OA as "granting to all users a free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly…." That is, OA may increase accessibility for users through making sources available.

However, given the diverse facets of access – and even physical access – mere availability may not always enhance accessibility. Lawrence (2001), for instance, argues that other variables such as search tools, search engines, and indexing can alter the *physical* accessibility of a document even if it is, strictly speaking, available. Culnan (1985) also notes that individuals' prior knowledge and the context of use can affect accessibility. This implies that OA, even if it does improve availability, may not improve every aspect of accessibility; for one thing, while the relationship between availability and true physical accessibility may be enhanced by OA, OA itself may not directly related to increasing intellectual and social accessibility. While the current study acknowledges the importance of both intellectual and social access, it focuses on the prerequisite concepts of availability and physical accessibility to explore physical barriers (or, strictly speaking, their virtual analogs) in accessing IRs – that is, the potential of IRs to make materials both available and easily accessible. As noted above, we investigate availability as the ability of search engines to retrieve clear links to an individual paper within the first two pages of results – that is, availability refers to the simple presence of an item in a set of search results, an indication that the item exists. Further, we examine accessibility as the number of clicks required for a user to navigate from those results to the full

text of the paper itself – thus, accessibility, in this study, refers to the amount of labor required of a user to actually obtain the item after having determined that it is available.

Method

In this study we explored the extent to which OA supports physical accessibility by conducting a case study of the institutional repository at Florida State University (FSU). FSU launched its IR, *DigiNole Commons* (http://diginole.lib.fsu.edu/), in mid-2011 in order to provide a common, openly accessible repository for scholarly and creative works of the university's faculty. To date, this IR has been hosted through Digital Commons, IR software provided by Berkeley Electronic Press (bepress) and managed by the FSU Libraries. As of the end of 2012, DigiNole Commons hosted a total of 5,020 items: 4,600 electronic theses and dissertations, 146 honors undergraduate theses, and 214 works by faculty. The dataset used for this study is a sub-set of the latter, and includes 170 faculty publications found in the IR that have also been published in peer-reviewed journals.

To analyze the IR's impact on physical accessibility, we conducted independent known-item title searches on both Google and Google Scholar (GS) to search for the faculty publications housed in DigiNole Commons. Numerous prior studies have provided a rationale for using Google and GS in collecting research materials. Jacso (2005) reported that GS is a powerful tool for searching scholarly information because its crawlers run "databases of the largest and most well-known scholarly publishers and university presses; their digital hosts/facilitators; societies and other scholarly organizations and government agencies, and preprint/reprint servers" (p. 209). Markland (2006) searched Google and GS to assess their efficiency in finding papers deposited in IRs, based on the assumption that users tend to use Google because of its simplicity and ease of use. Björk, Roos, and Lauri (2009) used title and author name searches on Google to estimate the proportion of green OA papers, treating any papers not found through this approach as unavailable, since most users assume papers are not available if they do not turn up in the first few results of a Google search. Norris, Oppenheim, and Rowland (2008) used both Google and GS to check the OA status of papers, and compared search results from Google and GS with two centralized IR access points, OAIster and OpenDOAR. While OAIster and OpenDOAR retrieved 14% of 2,280 OA papers, Google and GS were able to retrieve the other 86%, although there was little overlap in search results between Google and GS. Xia, Myer, and Wilhoite (2011) conducted title searches (sometimes supplemented with author names) on Google, GS, and Yahoo to look into the availability of papers, finding a similar difference in search results between Google and GS.

Data Collection. Open access IRs do not exist in isolation, but are just one possible place where the outcomes of faculty research can be found. Papers may be deposited by their authors or publishers in multiple repositories, for example, or may exist in multiple versions in varying file formats (e.g. HTML, PDF, etc.) or even with differing contents, since works made available as preprints are often subject to revision between the preprint stage and publication. To account for this phenomenon of "multi-locations" and "multi-versions" (p. 22), Xia et al. (2011) operationalize OA availability as "the number of web search engines that can return a link to the free full text of an article".

In the current study, however, we were concerned with the ability of a single IR to enhance physical access, and thus focuses on only one (out of many possible) version of a publication: that made available through DigiNole Commons at FSU via Google and GS

searches. Because of this more limited focus, availability is treated more narrowly in this study, as described above; we consider a paper to be available if a link to it is found on the first two pages of search results, and we then measure access by counting the number of clicks required to move from the search results to the full text of the item itself. If getting to a full text requires first moving to the second page of search results and then following a link from that page to the full text within DigiNole Commons, access to that article is considered to require two clicks; if finding the full text requires moving to the second page, following a link to DigiNole Commons metadata, and then following a link to the full text, access requires three clicks.

Google Scholar, possibly since it is more narrowly focused on scholarship, adds a particular complication for our measurement of access: rather than presenting unique links to specific items, it collapses multiple links into a single link covering multiple versions, both initially hiding specific search results and requiring an additional click in order to gain access to IR materials (see Figure 1, below). Thus, if an item housed in DigiNole Commons is one of several items retrieved in a GS search, an extra click is required to gain access to that item over what is required in a Google search.

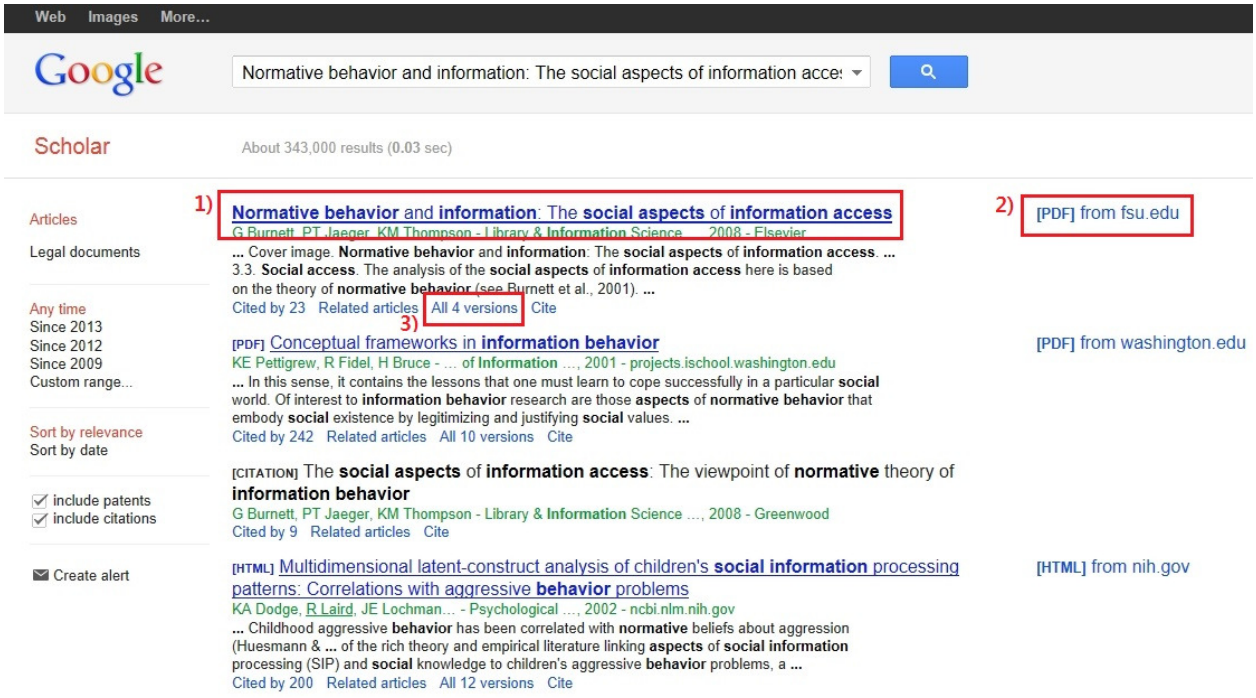

Figure 1: Google Scholar Links

Researchers first checked for the presence of metadata and full texts for 170 faculty publications in DigiNole Commons itself. In contrast to the complete availability of metadata, only 100 (58.82%) full texts out of 170 were available within DigiNole Commons. In the other 70 cases, links to 18 items external to DigiNole Commons were provided (7 to openly accessible sites such as author or departmental websites, and 11 to non-OA subscription-based sites such as JSTOR, publisher sites, etc.); an additional three items had links that appeared to no longer be current, and two items were under publisher embargo. For 47 items, neither full-text copies nor links to full texts were available at DigiNole Commons. Accordingly, while this study analyzes

metadata availability via Google and GS for 170 items, the analysis of full text availability and accessibility is limited to those 100 items for which full texts exist in DigiNole Commons. The researchers first conducted full title searches on Google and GS in March, 2013; although some scholars (e.g. Vaughan, 2004; Vaughan & Shaw, 2005) have suggested that the consistency of Google search results over time is typically fairly high, identical title searches were repeated in May, 2013. Fairly high consistency (ranging from 84.7% to 99.4%) was found in both availability and accessibility between the two search results, with most discrepancies found in GS searches, most commonly as a result of changes in GS links to multiple copies of an item, as discussed above, or small changes in search result rankings (e.g., items moving from the first page of results to the second or vice versa). In what follows, we report the findings based on the March 2013 searches, and notes the possible impact of search inconsistencies on availability and accessibility in the discussion.

Results

Availability. In Google searches, links to DigiNole metadata were found in the first two pages of search results for 78 (45.9%) out of 170 items; 74 (74.0%) out of 100 full texts housed in DigiNole Commons were available either directly (from a Google link to the item itself) or indirectly (through a Google link to DigiNole metadata). Searches in Google Scholar, by comparison, turned up links to DigiNole metadata in 127 (74.7%) cases out of 170, and to full texts in 78 (78.0%) out of 100 cases. A chi-square test for comparing two proportions shows a statistically significant difference (χ^2 =28.306, df=1, $p < .001$) in metadata availability between Google (45.9%) and GS (74.7%) at the .05 alpha level. However, the difference in full-text availability is not statistically significant.

Table 1: Availability of DigiNole metadata and full-text

	Metadata		Full-text	
	Google	Google Scholar	Google	Google Scholar
Available links	78 (45.9%)	127 (74.7%)	74 (74.0%)	78 (78.0%)
Unique links	18 (10.6%)	67 (39.4%)	18 (18.0%)	22 (22.0%)
Shared links	60 (35.3%)		56 (56.0%)	
Total combined links	145 (85.3%)		96 (96.0%)	
Total DigiNole Items	170 (100%)		100 (100%)	

As summarized in Table 1, GS searches uncovered 67 unique links to DigiNole metadata that did not turn up in Google searches, compared to 18 unique links that were found via Google

but not GS. For full texts, GS retrieved 22 unique items not found by Google, while Google retrieved 18 unique items. Considered together, GS and Google searches provided links to DigiNole metadata for a total of 145 (85.3%) of 170 items, and to full texts for 96 (96.0%) of 100 items.

Accessibility. The previous section describes the degree to which items housed in DigiNole Commons are made available – that is, the degree to which links appear in search results – in Google and GS. This section presents findings related to the accessibility of those items in terms of the number of clicks required to navigate to them from the search results. The examination of the search results for the 170 DigiNole Commons items revealed five different scenarios, illustrated in Figure 2.

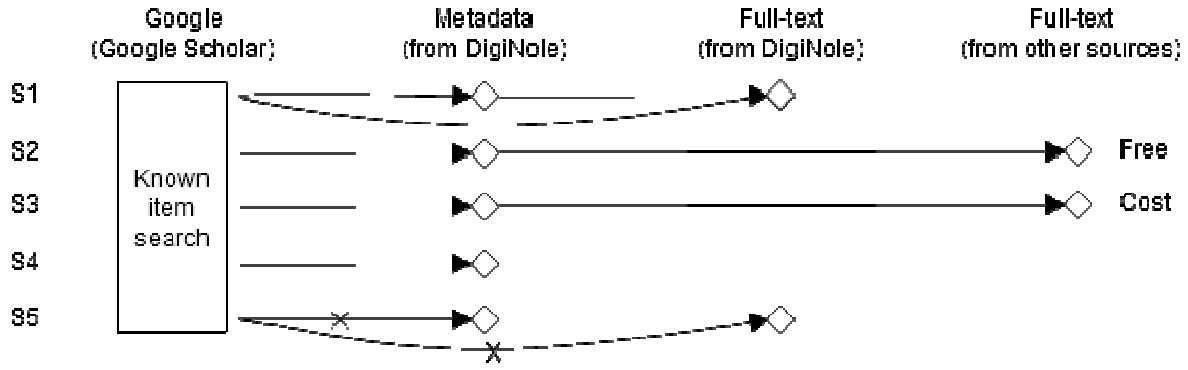

Figure 2: Five accessibility scenarios

Of these, only the first two – the first leading directly from a search to an item in the IR, and the second to a copy of an item not housed in the IR but still freely available at an author's website or some other location -- can be considered to be true OA scenarios. Each of the three other scenarios, because they do not result in access to free copies of materials, do not mesh with the fundamental objectives of OA initiatives: scenario three does, ultimately, lead a user to the item sought, but only externally to the IR itself, and with a cost via a subscription-based vendor; for DigiNole Commons and the FSU community, this means that only on-campus users – and off-campus users officially logged in – can still freely access materials, but others cannot; thus, access is limited. The final two scenarios – one retrieving metadata while failing to provide access to a full text from DigiNole Commons, and the other, which neither turns up DigiNole metadata nor leads a searcher to the full text of an item – cannot be considered to support OA.

Figure 3 shows the distribution of DigiNole Commons' 170 items into these five scenarios. Given the number of available full text items shown in Table 1, it would be reasonable to expect scenario one to include 96 items. However, curiously, two additional DigiNole Commons full texts, neither of which shows up in an examination of the IR itself, were retrieved by a GS search (for instance, DigiNole Commons provides an external link to one of these, *Evaluation of dynamically downscaled reanalysis precipitation data for hydrological application in the southeast United States*; however, an internal DigiNole version turns up among GS search results). In addition, four out of the 100 items with full texts available in DigiNole Commons are not retrieved in either Google or GS searches.

Six items, without full texts available in DigiNole Commons but with links to freely available external copies, fall into scenario 2. Ten items, with links to publisher sites or other subscription-based vendors, fall into scenario three. Thirty nine fall into scenario four, because their DigiNole Commons metadata pages provide no access at all to full text copies. Finally, scenario five includes 17 items, none of which can be retrieved from DigiNole Commons through Google or GS searches.

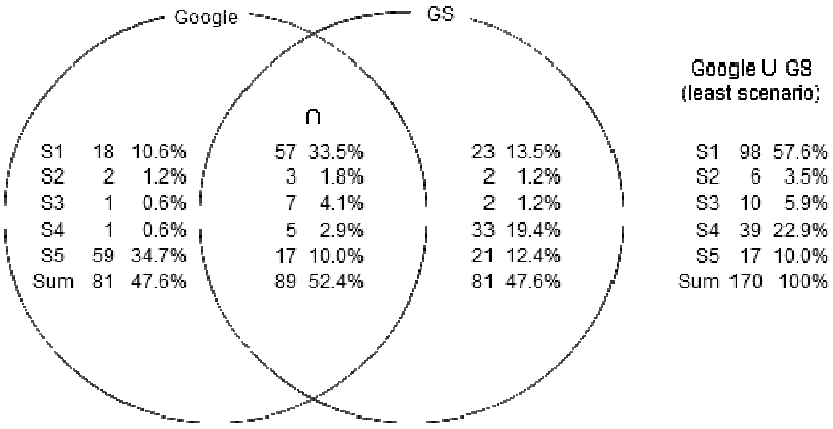

Figure 3: Comparison of scenario types between Google and Google Scholar

As noted earlier, this study measures accessibility as the number of mouse clicks required to reach either DigiNole metadata or full text copies from a set of Google or GS search results. Results differed a bit between the two, with GS requiring more clicks, largely because of the way GS groups similar items into a single link. From Google, an average of 1.12 clicks were required to access 78 DigiNole metadata records, and an average of 1.33 clicks were required to access 75 DigiNole full text items. All available metadata could be obtained with either one or two clicks from Google; the same is true for full texts, with the exception of one item requiring three clicks. In GS, access to 127 metadata records required an average of 1.69 clicks; access to 80 full texts required an average of 1.78 clicks. Again, with one exception, access to either metadata or full text required no more than three clicks; the one exception required four clicks to access the full text. The Mann-Whitney U Test, for comparing two sample means that do not fall into a normal distribution, was applied to compare accessibility across the two search engines; Google and GS are significantly different in both metadata ($p < .001$) and full-text accessibility ($p < .006$) at the .05 alpha level.

Other Issues. The analysis above provides an overview of the availability and accessibility of metadata and full texts housed in FSU's DigiNole Commons. However, several other issues arose during the analysis of both the materials themselves and the Google and GS search results that cannot be summarized statistically. These anomalies themselves have implications for the ability of the IR to provide access to the materials in its collection, and are briefly discussed here.

Title or authorship issues. IRs are often used to house articles in versions other than the final published version, such as pre-prints; a small handful of items housed in DigiNole Commons fall into this category, most of which display minor differences from their published counterparts, and a couple of which are radically different. For example, metadata for an article by Hart, Taylor, and Schatschneider (2013) is present in the IR under the title "There is a world

outside of experimental designs: Using twins to explore causation". This article (unavailable in full text in the IR because of an embargo) was published online in 2012 and in print (in the journal *Assessment for Effective Intervention*) under the similar – but not identical – title "There is a world outside of experimental designs: Using twins to *investigate* causation". Such a discrepancy, however, does not necessarily impede an article's accessibility, and, in this case, it does not: a title search in GS turns up both the DigiNole and the published versions.

Other changes, however, are neither so minor nor always so innocuous in terms of their impact on accessibility. The most extreme example of this can be seen in an article by Coutts (2009) published in the *Journal of Urban Planning and Development* under the title "Multiple case studies of the influence of land-use type on the distribution of uses along urban river greenways," but appearing in the IR under the title "Locational influence of land use type on the distribution of uses along urban river greenways". In another case, an article not only shows different titles between the DigiNole pre-print and the published versions, but the pre-print metadata gives the name of only one of seven authors listed in the publication. Clearly, such extreme variations can have important implications for OA, since they may serve to make the OA copies of the work inaccessible from Google or GS, as they did in the first of these two cases.

The algorithms governing how GS derives author names may also have an impact, although the three instances in which this was seen in the current study did not influence either availability or accessibility, given that the study used known-item title searches. In these three instances, GS mistakes the letters "MD" after an author's name as his initials, rather than as an indication that he is a medical doctor: thus, in GS, Jose E. Rodriguez, MD appears as "MD Rodriguez," with "E Jose" appearing as a separate author.

Item visibility and other anomalies. On initial examination, one case appeared to be a true anomaly: an article by Falk, Lepore, and Noe (2013) housed in the IR and titled "The cerebral cortex of Albert Einstein: A description and preliminary analysis of unpublished photographs," fell into scenario five (that is, it could not be retrieved) in a Google search, but was easily retrieved by GS. Initially, researchers found the inability of Google to uncover this item to be curious, since the article, upon publication, received considerable attention in both scholarly circles and the mass media; further, the article was published in a non-OA journal, with full open access to the published version and the right to archive the final published version in the IR immediately upon publication secured through the payment of an article processing charge (with assistance from one of the current article's authors). Upon reflection, however, it became clear that, while scenario five typically is a sign that the goals of open access are not being met – if items cannot be easily retrieved, that is, access cannot be considered to be truly open – in this case the most likely explanation is the opposite: access via Google to the DigiNole Commons copy does not constitute a failure of OA, but, rather, reflects the success of OA in a much broader sense. Because of the level of attention accorded the article, and its open accessibility on the journal's website, the article's presence in DigiNole Commons is not as important as it may be for lesser-known works; Google's algorithms treat the DigiNole Commons copy as simply one more among many available copies, resulting in lower relevance ranking.

One other item displays a similar pattern: "Development of a new academic digital library: A study of usage data of a core medical electronic journal collection" (Shearer, Klatt, & Nagy, 2009) turns up in a GS search, but not in a Google search. In this case, the likely reason is similar, though not identical: an examination of Google search results shows that the article is easily available in numerous other OA repositories, which, in turn, appears to have negatively

impacted the relevance ranking for the DigiNole Commons copy while enhancing the overall availability of the work through other outlets.

One final anomaly must be noted, although, strictly speaking, it falls outside of the parameters of this study: the search engine built into DigiNole Commons itself fails to retrieve certain items that are clearly present in the IR. In each case, this appears to be due to the fact that the articles' titles include non-alphanumeric characters such as parentheses, question marks, and asterisks. In each case, Google and GS both successfully retrieve the articles, even though the IR's search engine cannot do so. While this failure has no impact on the findings of this study, since it is internal to the IR rather than related to the availability and accessibility of the items via Google and GS searches, it does constitute an impediment for the success of DigiNole Commons' goal of providing open access to the materials in its collection.

Discussion

This case study confirms that IRs, at least overall, can contribute to making papers available and accessible on the open web. Nevertheless, it also uncovers some potential impediments to the success of IRs. As pointed out in the previous section, some situations either do not satisfy the goals of IRs at all or satisfy those goals only in part. In some cases, for instance, access to full texts requires the payment of fees to vendors, whether through subscription agreements between vendors and libraries hosting IRs, or through access fees paid by individual searchers. For many – in this case, for users conducting their searches from computers on the FSU campus or users logged in as authorized off-campus users – such costs may be hidden because they are covered by the institution with which they are affiliated; however, the very fact that such fees exist violates the spirit of open access for all but a defined set of authorized users.

Google or GS searches that retrieve metadata from DigiNole Commons but fail to retrieve full texts may occur for several reasons, including:

1. Contractual embargos, in which authors must withhold their work from open access for a contractually-determined period of time;
2. Undetected file upload errors;
3. Institutional policy or procedural issues; and
4. Erroneous or outdated links.

As the development of DigiNole Commons is ongoing, there are cases where metadata was entered to demonstrate the IR's functions and value to a department or faculty member, with the goal of submitting the full text of the articles at a later date following subsequent individual outreach. Also, there are cases where specific departments begin to utilize DigiNole Commons solely as a database for the research produced in the department. The institutional goal for the repository is that all metadata records will lead to full-text accessible items. In three cases, links to external full text copies turned out to be invalid, and three other articles were, at the time of the study, currently under embargo.

Instances in which neither metadata nor full texts were retrieved by Google or GS searches despite their presence in the IR, can most likely be attributed to issues related to the algorithms used by search engine crawlers either in searching or in determining relevance rankings. Moreover, although it was, as noted above, not a serious concern – and is, in any event, largely beyond the ability of IR administrators to mitigate – inconsistencies in "hits" across multiple searches is a potential impediment for full OA implementation, whether because of the implementation of search and relevance algorithms in Google and GS or because of the ways in

which links are updated over time. In addition, as noted above, title changes between pre-prints and final published versions can cause retrieval problems if the published versions of titles are used in "known item" searches (the same may be true in instances when there are changes in authorship between versions of a work, as noted above); if the title changes are minor, open retrieval may still be nearly seamless for users searching via Google or GS, but more significant title changes may make retrieval nearly impossible when the OA pre-print title is not used (Björk, Roos, & Lauri, 2009).

Some potential impediments are clearly beyond the means of libraries managing OA IRs to address; little, for instance, can be done about the ways in which search engine algorithms or relevance rankings cause existing items to go missing from Google or GS search results. However, there may be ways to mitigate some of the other potential impediments, edging IRs closer to full implementation of OA goals. Since some items become inaccessible because of differences between pre-print and published versions, metadata records – following Dublin Core or other metadata standards – can make linkages and connections between these multiple versions explicit, and, thus, searchable. Also, although it not possible to alter third party search algorithms, libraries can increase IR paper availability by representing and organizing their papers using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), so that service providers, such as OAIster, can help make those papers accessible (Norris et al., 2008).

When metadata is retrieved but full texts are missing, copies of materials, when available, can be obtained and uploaded, and regular checking of links – particularly for materials stored in author websites or other external OA sites – can help to ensure continued access. It may also be useful to take advantage of redundancy, and maintain full text copies of materials within the IR even if they are also freely available elsewhere. When full texts can be retrieved but come with a subscription or other cost, librarians may be able to find ways of including alternative versions of full texts within their IRs or may be able to negotiate with vendors, publishers, faculty members, or academic societies so as to make them accessible. Finally, ongoing initiatives to educate faculty about their rights related to their own work may help to increase the frequency of open access clauses in copyright agreements with non-OA publishers.

This study, although limited to a single case at a specific point in time, suggests that relying on either Google or GS individually cannot ensure full access to scholarly works housed in OA IRs. Counting only those 104 instances that can be fully considered to be OA (i.e., scenarios one and two), there is an overlap in results between the two search engines of only 57.5%, with Google providing links to 20 items not found via GS, and GS providing links to 25 items that are inaccessible via Google. Thus, it is necessary to use the two search tools together to find materials deposited in IRs. In terms of general availability, GS appears to have the edge, especially for metadata records. On the other hand, Google does a better job of supporting access, requiring fewer mouse clicks overall to get to DigiNole materials than GS because of the way GS clusters multiple copies of an item into a single initial link.

Conclusion

While availability is a dimension of accessibility, making a distinction between them enabled us to identify possible impediments to the success of IR. In this case study we examined the degree of availability and physical accessibility of a collection of limited size housed in a particular open access IR, via known-item title searches in Google and Google Scholar. Overall, the findings suggest that items in the collection are, for the most part, both available and accessible,

although a bit more than 30% of items, falling into scenarios four and five, could not be retrieved at all; further, although an additional 6% of items could be retrieved, that retrieval came with the cost of subscription or other charges to either the institution or individual searchers, limiting the degree to which their accessibility could truly be considered to be "open". Considering those items, impediments to open access generally fall into the following two broad categories:

1. Impediments related to contractual agreements between authors, publishers, and vendors, including costs related to institutional subscriptions, item embargo, etc.
2. Impediments related to the policies, practices, and technologies governing the IR itself, including outdated links, file upload errors, and internal search engine shortcomings.

It should be noted that other approaches to this study had the potential of leading to different findings – for instance, author searches would likely have looked very different. Similarly, this study presents a slice of time in the life of a specific IR; a replication of the study after a period of months or longer might turn up quite different results, whether due to possible changes in Google and GS search and relevance algorithms, changes to the IR itself, or other factors. In addition, it is important to note that the full dataset analyzed in this study was limited to only 170 items, a number that does not even come close to representing the full research productivity of the FSU faculty. This final fact suggests what is, perhaps, the most important impediment to the success of OA is rooted in faculty attitudes toward OA and, most important, lack of faculty participation in IRs that are available to them; however, such impediments fall into the realm of intellectual and/or social access, and are thus beyond the scope of the current study.

The FSU Library system is continuing development of services related to its IR, including efforts to enhance and "clean up" metadata records of items housed there. The authors of the current study plan further study in collaboration with the library, and will initiate additional outreach efforts, outlining the benefits of archiving full-text versions of articles in the repository. The campus Office of Scholarly Communication is also working with several new offices on campus, including the Office of Proposal Development and the Office of Sponsored Research, to make more effective connections between open access scholarly objects online and their metadata records as presented in DigiNole Commons. This study is an essential part of the monitoring, testing, assessing and adapting of the repository platform as one facet of the scholarly communication initiative at Florida State University.

As noted earlier, this current study is part of an ongoing investigation into OA issues and IR effectiveness; future work will look at several issues beyond simple questions of availability and physical accessibility, investigating other types of accessibility (intellectual and social) in relation to open access IRs, and will include projects designed to increase faculty awareness and participation.

References

Antelman, K. (2004). Do open-access articles have a greater research impact? *College and Research Libraries, 65*(5), 372-382.

Bailey, C. W. (2010). A short introduction to open access. In *Transforming Scholarly Publishing through Open Access: A Bibliography.* Retrieved from http://digital-scholarship.org/tsp/w/introduction.htm

Björk, B-C. (2004). Open access to scientific publications: An analysis of the barriers to change. *Information Research, 9*(2) paper 170. Retrieved from http://InformationR.net/ir/9-2/paper170.html

Björk, B-C., Laako, M., Welling, P., & Paetau, P. (2014). Anatomy of green open access. *Journal of the Association for Information Science and Technology, 65*(2), 237-250.

Björk, B-C., Roos, A., & Lauri, M. (2009). Scientific journal publishing: yearly volume and open access availability. *Information Research, 14*(1) paper 391. Retrieved from http://InformationR.net/ir/14-1/paper391.html

Budapest Open Access Initiative. (2002). Read the Budapest open access initiative. Retrieved from http://www.opensocietyfoundations.org/openaccess/read

Bullinger, H.-J., Einhäupl, K. M., Gaehtgens, P., Gruss, P., Henkel, H.-O., Kröll, W., & Winnacker, E.-L. (2003). Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. Retrieved from http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf

Burnett, G., Jaeger, P. T., & Thompson, K. M. (2008). Normative behavior and information: The social aspects of information access. *Library & Information Science Research, 30*(1), 56-66.

Chan, L. (2004). Supporting and enhancing scholarship in the digital age: The role of open-access institutional repositories. *Canadian Journal of Communication, 29*, 277-300.

Coutts, C. (2009). Multiple case studies of the influence of land-use type on the distribution of uses along urban river greenways. *Journal of Urban Planning and Development, 135*(1), 31-38.

Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact? A critical review of the literature. *Journal of Informetrics, 1*, 239-248.

Crow, R. (2002). *The case for institutional repositories: A SPARC position paper*. Retrieved from http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf

Cullen, R., & Chawner, B. (2011). Institutional repositories, open access, and scholarly communication: a study of conflicting paradigms. *Journal of Academic Librarianship, 37*(6), 460-470.

Culnan, M. J. (1985). The dimensions of perceived accessibility to information: Implications for the diversity of information systems and services. *Journal of the American Society for Information Science, 36*(5), 302-308.

Davis, P. M., & Connolly, M. J. L. (2007). Institutional repositories: Evaluating the reasons for non-use of Cornell University's installation of DSpace. *D-Lib Magazine, 13*(3/4). Retrieved from http://www.dlib.org/dlib/march07/davis/03davis.html

Davis, P. M., & Fromerth, M. J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics, 71*(2), 203-215.

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., & Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: randomized controlled trial. *BMJ, 337*:a568. doi:10.1136/bmj.a568

Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology, 4*(5), e157.

Falk, D., Lepore, F. E., & Noe, A. (2013). The cerebral cortex of Albert Einstein: A description and preliminary analysis of unpublished photographs. *Brain, 136*, 1304-1327. doi:10.1093/brain/aws295

Fidel, R., & Green, M. (2004). The many faces of accessibility: engineers' perception of information sources. *Information Processing and Management, 40*, 563-581.

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., & Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE, 5*(10):e13636.doi:10.1371/journal.pone.0013636

Hargittai, E., & Hinnant, A. (2006). Toward a social framework for information-seeking. In A. Spink & C. Cole (Eds), *New Directions in Human Information Behavior* (pp. 55-70). Dordrecht, The Netherlands: Springer.

Harnad, S., & Brody, T. (2004). Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine, 10*(6). Retrieved from http://www.dlib.org/dlib/june04/harnad/06harnad.html

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C., & Hilfet, E. (2008). The access/impact problem and the green and gold roads to open access: An update. *Serials Review, 34*(1), 36-40.

Hart, S., A., Taylor, J., & Schatschneider, C. (2013). There is a world outside of experimental designs using twins to investigate causation. *Assessment for Effective Intervention, 38*(2), 117-126.

Jacso, P. (2005). Google Scholar: the pros and the cons. *Online Information Review, 29*(2), 208-214.

Koler-Povh, T., Južnič, P., & Turk, G. (2014). Impact of open access on citation of scholarly publications in the field of civil engineering. *Scientometrics, 98*, 1033-1045.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., & Murray, S. S. (2005). The effect of use and access on citations. *Information Processing and Management, 41*, 1395-1402.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature, 411*, 521.

Markland, M. (2006). Institutional repositories in the UK: What can the Google user find there? *Journal of Librarianship and Information Science, 38*(4), 221-228.

McCord, A. (2003). Institutional repositories: Enhancing teaching, learning, and research. EDUCAUSE Evolving Technologies Committee white paper. Retrieved from http://net.educause.edu/ir/library/pdf/DEC0303.pdf

Miguel, S., Chinchilla-Rodriguez, Z., & Moya-Anegon, F. (2011). Open access and Scopus: A new approach to scientific visibility from the standpoint of access. *Journal of the American Society for Information Science and Technology, 62*(6), 1130-1145.

Moed, H. (2007). The effect of "Open Access" on citation impact: An analysis of ArXiv's Condensed Matter section. *Journal of the American Society for Information Science and Technology, 58*(13), 2047-2054.

Nicholas, D., Rowlands, I., Watkinson, A., Brown, D., & Jamali, H. R. (2012). Digital repositories ten years on: what do scientific researchers think of them and how do they use them? *Learned Publishing, 25*(3), 195-206.

Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open access articles. *Journal of the American Society for Information Science and Technology, 59*(12), 1963-1972.

Oguz, F., & Assefa, S. (2014). Faculty members' perceptions towards institutional repository at a medium-sized university: Application of a binary logistic regression model. *Library Review, 63*(3), 189-202.

Parks, R. (2002). The Faustian grip of academic publishing. *Journal of Economic Methodology, 3*, 317-335.

Swan, A., & Brown, S. (2005). *Open access self-archiving: An author study*. Retrieved from http://eprints.soton.ac.uk/260999/

Shearer, M. K. (2002). Institutional repositories: Towards the identification of critical success factors. *The Canadian Journal of Information and Library Science, 27*(3), 89-108.

Shearer, B. S., Klatt, C., & Nagy, S. P. (2009). Development of a new academic digital library: a study of usage data of a core medical electronic journal collection. *Journal of the Medical Library Association, 97*(2), 93-101.

Ugah, A. D. (2008). Availability and accessibility of information sources and the use of library services at Michael Okpara University of Agriculture. *Library Philosophy and Practice*. Retrieved from http://www.webpages.uidaho.edu/~mbolin/ugah4.pdf

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested. *Information Processing and Management, 40*, 677-691.

Vaughan, L., & Shaw, D. (2005). Web citation data for impact assessment: A comparison of four science disciplines. *Journal of the American Society for Information Science and Technology, 56*(10), 1075-1087.

Watson, S. (2007). Authors' attitudes to, and awareness and use of, a university institutional repository. *Serials, 20*(3), 225-230.

Xia, J. (2010). A longitudinal study of scholars attitudes and behaviors toward open-access journal publishing. *Journal of the American Society for Information Science and Technology, 61(*3), 615-624.

Xia, J., Myers, R. L., & Wilhoite, S. K. (2011). Multiple open access availability and citation impact. *Journal of Information Science, 37*(1), 19-28.

Xia, J., & Nakanishi, K. (2012). Self-selection and the citation advantage of open access articles. *Online Information Review, 36*(1), 40-51.