2-2014

Academic Libraries and Open Access Strategies

C. Sean Burns
University of Kentucky, sean.burns@uky.edu

Digital Object Identifier (DOI)
http://dx.doi.org/10.1108/S0732-067120140000032003

ACADEMIC LIBRARIES AND OPEN ACCESS STRATEGIES

C. Sean Burns

ABSTRACT

With the rise of alternate discovery services, such as Google Scholar, in conjunction with the increase in open access content, researchers have the option to bypass academic libraries when they search for and retrieve scholarly information. This state of affairs implies that academic libraries exist in competition with these alternate services and with the patrons who use them, and as a result, may be disintermediated from the scholarly information seeking and retrieval process. Drawing from decision and game theory, bounded rationality, information seeking theory, citation theory, and social computing theory, this study investigates how academic librarians are responding as competitors to changing scholarly information seeking and collecting practices. Bibliographic data was collected in 2010 from a systematic random sample of references on *CiteULike*.org and analyzed with three years of bibliometric data collected from *Google Scholar*. Findings suggest that although scholars may choose to bypass libraries when they seek scholarly information, academic libraries continue to provide a majority of scholarly documentation needs through open access and institutional repositories. Overall, the results indicate that academic librarians are playing the scholarly communication game competitively.

Keywords: Open access; collection management; bibliometrics; decision and game theory; bounded rationality; principle of least effort

INTRODUCTION

In 2010, Ithaka S + R published the results of a 2009 survey which asked faculty about their scholarly communication behaviors and attitudes. The survey gives some credence to the following key observation:

> Basic scholarly information use practices have shifted rapidly in recent years, and as a result the academic library is increasingly being disintermediated from the discovery process, risking irrelevance in one of its core functional areas [Emphasis added] (Schonfeld & Housewright, 2010, p. 2).

Contrary to recent studies that suggest increased usage of the academic library (e.g., Budd, 2009), the report suggests that researchers in the sciences, social sciences, and the humanities have moved away from the library building, the librarians, and the library's catalog and databases and have moved toward general purpose search engines and other electronic resources to find and satisfy their document needs. Although search and discovery through electronic services include those to which the library subscribes, the report reveals, at the network level, the heavy use of nonlibrary electronic discovery services. For instance, searching with Google ranks third in the discovery process (~70%), behind searching electronic, full text databases (~90%), and following citations (~90%), a process referred to as chaining (Ellis, Cox, & Hall, 1993). While only 8.6% out of 35,184 faculty who received the survey responded, and

although some have argued that the survey is based on incomplete premises (Nyquist, 2010), the findings warrant additional research about either the central or marginalized role academic libraries play in the work of today's scholars. Thus, the Ithaka report informs the first research question:

> **RQ 1.** Is the current state of affairs, at the network level, such that nonlibrary electronic discovery services marginalize academic libraries?

The state of affairs at the network level may encourage alternate paths to information, but open access content adds an additional problem for academic libraries. Broadly speaking, open access content is the content that is freely accessible to readers with means to the Internet. This is unlike other electronic, scholarly content behind subscription barriers, which requires both Internet and subscription access, such as through a library. Given that open access content is accessible outside a library's collections, if researchers increasingly use nonlibrary electronic discovery services, then nonlibrary electronic discovery services plus the growing availability of open access content make it possible to bypass both the library's services and electronic collections.

Research about the influence and reach of open access content is growing. With its perceived importance for academic libraries, as a publishing model that librarians hope will counteract the growing and unsustainable costs of serials, such influence and reach require examination and inform the second research question:

> **RQ 2.** Does open access content, in conjunction with nonlibrary electronic discovery services, marginalize academic libraries?

Framing these research questions in this way seems to suggest an argument against open access publishing, but that is neither the purpose nor the intent of this study. Rather, the objective of this study is to understand how trends in information seeking practices (e.g., searching for information outside the library with services such as Google and *Google Scholar*) in conjunction with the increasing availability of open access content (e.g., the ability to acquire a growing amount of quality information outside the library from open access entities such as PLOS ONE, PeerJ, and others) will change the fundamental notion of what an academic library is and will be in the 21st century.

The unit of analysis in this study involves both the information seeking and information use practices of scholars and researchers (hereafter just researchers). In order to frame this study, we can think of information seeking as a type of decision making and of acquiring information as a type of payoff. Addressing these questions from this perspective allows us to draw from a framework built on a theory of decision making and competition, or more properly, decision and game theory. This becomes clear when we think of the whole scholarly game itself, where the practices of these researchers are placed in the context of the services and the content provided by academic librarians. That is, any time a researcher seeks information, the researcher engages in a series of decisions. Any time a researcher acquires a relevant and salient piece of information (such as a journal article), the researcher receives a payoff. Likewise, if academic libraries measure their value and receive their payoff by the quality, quantity, and use of their collections, then any time a researcher does not use the academic library in favor of some other route where he or she still acquires a payoff, then the academic library declines in value. In the whole game, it is important to know how the academic librarian responds to the researcher's complete information seeking strategy.

The analysis in this study is based on a systematic random sample of bibliographic references collected by users of *CiteULike*, a social computing bibliographic reference management web site. Using these references' bibliometric data, collected from Google Scholar, the objective is to identify where and how these users have collected their journal article references. Using logistic regression, the second objective is to determine what factors predict or explain open access availability. Finally, using Bayes' theorem, the third objective is to build a hypothetical probability profile that illustrates the likelihood that a library's collections have been used given the use of other documents that may be sourced at other locations, such as those held in subject or institutional repositories and which may be found through a service such as *Google Scholar*. This process allows a determination of whether using nonlibrary discovery services to retrieve open access or freely available content is a relevant alternative to using the library's services to retrieve subscribed content. If the relevant alternative is viable, then the process allows for a determination to be made about the competitiveness of the alternative.

Framing the terms nonlibrary discovery services, alternate discovery services, relevant alternative, or third party discovery services with respect to what Ithaka S + R (Schonfeld & Housewright, 2010) describe as "A general purpose search engine on The Internet or World Wide Web such as Google or Yahoo" (p. 4), we can propose two hypotheses:

> **H1.** Using a third party discovery service to retrieve open access or freely available content is a relevant alternative to using the library's services to retrieve subscribed content.
>
> **H2.** The relevant alternative is a competitive alternative; that is, the relevant alternative entails an outcome where the payoffs are greater than the decision to use the academic library's services and subscribed content.

The overall goal of this project is to understand the implications that researchers' information seeking, retrieving, and collecting practices have on academic libraries. The hope is that the analysis will help academic librarians and library and information science researchers devise strategies that serve their communities' needs given a world where users have many choices for searching and retrieving information.

Furthermore, the significance of this issue involves the impact that open access content and alternate discovery services will have on the academic library's core function and purpose. While the existence of nonlibrary options to the information seeker is nontrivial, what gives the entire search and source domain its real value lies with how and why people make decisions or accomplish their information tasks. The information needs of the user are not met simply by providing relevant collections but by also addressing their decision matrices and by developing an understanding of how their decision matrices might be rational. An introduction to these decision issues is presented in the following section.

PREFERENCE, UTILITY, RISK, AND PRIOR INFORMATION

The Ithaka S + R report (Schonfeld & Housewright, 2010) reveals something about information seekers' and users' preferences. More generally though, library and information science research has excelled in identifying the preferences of those engaged in information seeking and use. These preferences are often used to help both librarians and information seekers acquire more skills at handling the complex information and knowledge systems that our society is built upon

(Julien & Genuis, 2011). However, a list of these user preferences can also be applied by librarians to devise appropriate strategies that respond to users' information seeking related actions (e.g., Mullen & Hartman, 2006). In this sense, the preferences that library and information science research have identified serve as a rich source of information for devising and responding to what users want or need in terms of information services and sources and also in terms of organizational needs (see e.g., Theng & Sin, 2012).

Decision and game theory use preferences to rank the payoffs one would expect to receive by applying a decision or strategy (Dixit & Skeath, 2004). The theories help either to explain or to prescribe courses of action for single individuals or agents or between two or more people or agents whose decisions take into consideration the others'. For example, given an agent's preference to act in a certain way, such as a tacit preference to acquire as much as possible or as much as is needed for as little cost as possible, decision theory provides an analytical framework that describes how an agent makes a decision among a set of relevant alternatives, with the intention of receiving a maximum payoff. In the context of this study, the decision may involve the use of a library's or a nonlibrary's search service as a research starting point.

Game theory describes how an agent selects a strategy in response to an opposing player's strategy selection. For example, given a user's preference for little effort and much gain, it could be asked what is a librarian's best strategic response. In this research, there is the abstract view that librarians function as one player and researchers, as information seekers and users (in general), function as an opposing player. This relationship is motivated by a simple explanatory heuristic (Abbott, 2004) which places front and center the notion that a strategic interaction exists between librarians and members of their communities. This is due to the librarians' attempts to offer the best search and retrieval services and the information seekers' attempts to satisfy their search efforts using whatever relevant search services are available to them.

George Kingsley Zipf (1949) termed the principle of least effort to describe what he derived as a natural tendency among individuals not simply to minimize their work but their probable average rate of work. He used the phrase principle of least effort to describe this tendency but in doing so, the focus on the probable aspect of the principle sometimes gets lost. In reemphasizing this, it becomes apparent that actions to minimize the probable average rate of work are based on the information we have regarding those probabilities or, lacking complete information, the predictive expectations (Nickel, 2009) or beliefs we have about them. This implies, though, that if we intend to minimize our probable average rate of work, we may or may not be successful given what we expect or believe will help do so (c.f., Savolainen, 2012).

Although Zipf describes the principle of least effort as a natural human behavior or tendency, the framework used in this study takes the view that the principle of least effort can also be thought of as a preference of least effort. The semantic substitution simply places extra emphasis on the notion that what explains our tendencies and choices are varied (Hausman, 2005). In the sense that we intentionally act on those preferences, then they are actionable too. Despite the terminology, we might posit that some choose *Google Scholar* as a research starting point because their preferences for locating information include maximizing their success for finding information while minimizing their probable rate of effort to do so. At the same time, some may choose the library for the same reason.

The important question for librarians concerns what users are doing in the aggregate. If researchers tend to select a third party search service as a research starting point as often as a library's search service (e.g., Niu et al., 2010), then it may not be because these researchers believe that the library's services cannot satisfy their information needs; rather, it may very well be because these researchers believe that using the library's search service requires greater effort, or greater cost, given both the possible outcomes or payoffs with respect to the other options available to them. The question then is how much of a payoff does one need to pursue a decision when it is believed to be costly? Or what incentives are needed to encourage maximizing and not just satisficing (Simon, 1955), where to maximize indicates acquiring the highest possible payoff? Or, alternatively, how can the use of an academic library, or the conscious decision to choose the academic library as a research starting point, be viewed or believed to be a satisficing function and not a maximizing function? These alternate choices are always in opposition to the other; hence, while the principle of least effort is an interesting concept alone, it is even more interesting when placed alongside relevant alternatives. When evaluating a library's services or its collections in order to determine, for example, a return on investment (e.g., Tenopir, 2012), that value cannot be determined in isolation from the value of a relevant alternative, just as the value of real estate cannot be determined without tracking adjacent property values (e.g., Farber, 1998). Thus, for example, we could ask what the academic library's value is given the existence and the popularity of a thing like *Google Scholar* which can be used to retrieve free content.

Consider a hypothetical. If someone guarantees Adam $10 to perform a task involving minimal effort or $20 to perform a task involving great effort, which task will Adam select? This depends on several factors. One, it depends on Adam's current need and wealth (Brandstätter & Brandstätter, 1996). If Adam has no wealth and is trying to determine how ¨ to purchase his next meal, it may be more likely that he will choose the more difficult task for $20 in order to increase his payoff. However, if Adam has a few hundred dollars in hand, then the law of diminishing returns suggests it is likely that he will choose the easy task since the difference between $20 and $10, minus the cost of effort, is less important to him.

The subjective utility or payoff of either task may depend on Adam's risk attitude (Rabin, 2000). Let us stipulate that a payout is guaranteed only if Adam succeeds in accomplishing the task, and let us define the minimal and maximal efforts by the likelihood of successful completion. The risk might involve Adam's belief about whether he can accomplish the task. For example, let us say that he believes the task that involves minimal effort will be less risky with a probability of success at 0.70, while the task that involves greater effort will only have a probability of success at .30. In a case with few qualifications, such as this, only the risk-seeking person chooses the path of greater effort. Both the risk-averse and the riskneutral persons will likely prefer the path of least effort (see also Kahneman & Tversky, 1979; Tversky & Kahneman, 1974).

A third factor involves prior information (Schmeidler, 1989), which can be illustrated with a story (Grune-Yanoff & Schweinzer, 2008). Imagine we are on a quest to seek the Holy Grail and as we walk down a road surrounded by a dark forest, we find ourselves at a fork in the road and thus have a choice between going left or going right. If we have no prior information, then we cannot make an informed choice. Our choice is random. However, suppose we do have prior information. We recall that we met a mysterious knight at a tavern in the last town we visited and over a pint of ale, the knight recounted a poem that we now believe is a clue about

which path to select. Based on this information, we decide to take the path on the left, the road less traveled. However, we find that it is less traveled because it is underdeveloped. As such, it requires greater effort to traverse it. Since we expect the payoff to be great, we take it.

The story illustrates that when we deviate from our natural tendency to reduce our probable average rate of work, we may do so only if the expected payoff is greater, and we may only have such an expectation if we have the requisite prior information and a proper risk attitude. The problem is that we know that researchers do have prior information when they make decisions about which choice they are going to make when they initiate a search. Since we know that, we are left with the notion that, if researchers, in aggregate, more often choose one path over another, they do so because they perceive the payoff to match the risk and cost involved.

Problem Statement

The preferences, utilities, risk attitudes, and prior information held by information seekers and users all play a role in the choices made among a set of relevant alternatives. In order to influence those user choices, librarians have responded by teaching users certain skill sets or ways of thinking critically about information and its sources. This response is most representative in the drive to promote and teach information literacy skills (ACRL, 2000). While possessing information literacy skills may encourage the critical evaluation of sources and help ensure the use of good, quality information, it does not entail the use of the library to acquire those sources, and it does not necessarily encourage that use. As more scholarship and data migrates to online databases or is born digital, if it remains freely accessible at zero marginal cost to the information user and can be discovered using nonlibrary discovery services, then a problem exists if librarians define themselves as primarily about the tools and collections they provide. This is especially problematic if library tools and collections are used less than others that are available outside the bounds of the academic library. The consequences are strategic and can be illustrated with the following set of inferences.

Inference 1:

> **P1.** If academic libraries are places where, historically, scholars have acquired most of their scholarly documentation, then academic libraries are places that have had a monopoly on scholarly documentation (Hamlin, 1981; Sapp & Gilmour, 2002, 2003; Shiflett, 1981; Wiegand, 1990).
>
> **P2.** Scholars can now acquire scholarly documentation from any of a number of places that are readily available (Tenopir, King, Spencer, & Wu, 2009).
>
> **C.** Therefore, academic libraries no longer have a monopoly on providing scholarly documentation.

Inference 2:

> **P1.** If academic libraries no longer have a monopoly on providing scholarly documentation, then academic libraries are in competition with other places or entities that scholars use to acquire scholarly documentation (Sennyey, Ross, & Mills, 2009).
>
> **P2.** Scholars are using these other places or entities as or more frequently than academic libraries for acquiring scholarly documentation (Niu & Hemminger, 2012; Schonfeld &

Housewright, 2010).

C. Therefore, these other places (or other entities) may be competing successfully with academic libraries as providers of scholarly documentation.

Inference 3:

P1. If other places are out-competing academic libraries as providers of scholarly documentation, then these other places have dominating strategy profiles.

P2. Successful competition is largely determined by the choice of a dominating strategy profile (Binmore, 2007; Dixit & Skeath, 2004).

C. Therefore, academic libraries are competing with dominated strategy profiles.

The academic library has played a central role in the life of researchers for most of the 20th century, but today researchers have other options available to them, and these options provide competing services and sources of information. The first two inferences illustrate this, and the conclusion expressed in the third explains the actions made by researchers and scholars who actively choose these other services and sources of information instead of those provided by librarians. If academic libraries must compete, or are competing, then it is important to understand the strategies they are using to meet the challenge.

Research Questions

Based on the availability of nonlibrary discovery services such as *Google Scholar*, the availability of freely accessible content such as open access journal articles, as well as an aggregate preference for least effort and other decision-making factors such as subjective utility, this study asks and addresses the following two research questions:

1. Is the current state of affairs, at the network level, such that nonlibrary electronic discovery services marginalize academic libraries? The first research question has a strategic dimension, which is highlighted in the following forms:

(a) **R_1:** Using a third party discovery service to retrieve open access or freely available content is a relevant alternative to using the library's services to retrieve subscribed content.

(b) **R_2:** The relevant alternative is a competitive alternative.

The second research question, by acknowledging the existence of open access content, grants viability to the strategic dimension of the first research question:

2. Does open access content, in conjunction with nonlibrary electronic discovery services, marginalize academic libraries?

The research questions are answered by deriving two operational questions, where the first operational question addresses research question 1 and the second operational question addresses research question 2.

i. What is the probability that any given researcher can use *Google Scholar* to retrieve a relevant full text document without the benefit of an academic library's proxy or similar service?

ii. What bibliometric or publishing characteristics are driving full text access to journal articles that users collect?

Limitations and Delimitations

Open access is defined, for the purposes of this paper, as anything that is freely available in full text format via an alternative discovery network such as *Google Scholar*. This weaker definition is used simply due to the difficulty in determining the strictly defined open access status of each full text document found in this study when that document may come from any of a variety of sources, including publisher web sites or personal, academic web sites. It also does not consider the quality of the full text document or whether that document is a preprint, postprint, a copy of a published article, or a word processed document.

Furthermore, the subject content of the study is largely limited to the scientific disciplines and does not attempt to control for variations used in specific fields of study, for speed of communication, or for obsolescence of the product of study. Instead, it randomly samples from a single community of researchers, most of whom however come from the life, computer, and information sciences.

Although the unit of observation is the bibliographic reference and although this study employs methods from citation analysis and bibliometrics, the context of this analysis is not based on the social act of citing a bibliographic reference. Instead, it is based on the social act of collecting a bibliographic reference. In most citation analyses, the object under study concerns citations of references by citing authors, but in this study, the reference is collected by a potential reader. Although at least one study has been conducted on the social collecting of bibliographic references and what this activity means with respect to scholarly communication (Borrego & Fry, 2012), and although the altmetrics movement argues for evaluating additional sources of influence (Priem & Hemminger, 2010), there is no strong theoretical study that compares the collecting of a reference to the citing of a reference. This research will offer theoretical leads to the behavior and meaning involved in collecting bibliographic records, including whether the actions involved in collecting a bibliographic reference are theoretically comparable to the actions involved in citing a bibliographic reference (e.g., Narin & Moll, 1977).

Lastly, a note about the data sources used in this study. *CiteULike* (http://www.citeulike.org/) is a specialized social bookmarking service particularly tailored to meet the document management needs of researchers and scholars (Hull, Pettifer, & Kell, 2008; tbogers, 2009), and it has been available since November 2004. Unlike other social bookmarking services that encourage users to capture and tag a link to any web page, *CiteULike*'s focus is scholarly bibliographic references. Essentially, it "is a Web-based tool to help scientists, researchers and academics store, organise, share and discover links to academic research papers" (Emamy & Cameron, 2007, para. 2). Users maintain digital libraries of their collected references, attach memorable tags to these references, and upload articles for later access. Personal libraries are public by default, although users can make their bibliographic references private, and users may form groups based on research interests or projects. These libraries are also indexed by search engines, such as Google and *Google Scholar*.

Google Scholar is the second data source used in this study. It is a bibliographic database owned by the Google search company. As a bibliographic database, it is similar to Elsevier's Scopus and Thompson Reuter's Web of Knowledge, the latter having origins in work done by Eugene Garfield (1955). *Google Scholar*'s strengths and weaknesses are debated, but research suggests that its ability to retrieve links to a wide range of scholarly communication sources is as

strong as its subscription counterparts (Chen, 2010; Howland, Wright, Boughan, & Roberts, 2009). Other researchers have found that it can retrieve high numbers of open access materials (Norris, Oppenheim, & Rowland, 2008).

LITERATURE REVIEW

Introduction

This study examines the impact that two states of affairs, nonlibrary resource discovery services, and freely available content, have on academic libraries. Since these two states of affairs raise the possibility of bypassing an academic library's services and collections, they have the potential to marginalize academic libraries in at least two of their core functions: collection development and user services. The issue highlights the nature of what it means for a library to collect and to disseminate its collection.

In order to study the issue, the first section of this chapter explores some historical aspects of the academic library and seeks to explain how perspectives of the academic library have shifted in the last century and a half. Since the common perception of libraries is very much intertwined with the collections librarians build, store, and manage, particular emphasis is placed on the significance of library collections and on librarianship as a profession.

The second section reviews *Google Scholar* and outlines how it has become a viable scholarly information discovery service. This involves reviewing the literature that has examined Google Scholar's ability to locate and retrieve scholarly information as well as its ability to retrieve open access content. This review is followed with a discussion of issues in scholarly communication and publishing relating to rising journal costs and the move to digital formats which has made scholarly communication freely accessible. This will entail a discussion of the open access movement including an outline of its characteristics and why both researchers and librarians consider it important.

The third section outlines the theoretical and methodological dimensions of this study. Since scholarly information behavior is simply another way to refer to the choices scholars make in searching and using scholarly information, and since these choices influence the choices made by others, this section begins with a discussion of decision and game theory based on bounded rationality assumptions. The theoretical framework is explored using bibliometric methods and so an overview of the use of bibliometrics and citation analysis for the study of scholarly communication follows.

This is closely followed with a discussion of the use of social computing and the theoretical characteristics that allow or afford scholarly communication behavior on the web. Particular attention is paid to web-based data sources that are citation based, such as *Google Scholar*, and web-based data sources that are socially driven, such as *CiteULike*.

The Purpose of the Academic Library

The definition of the academic library is an evolving and contested issue, and this is largely due to two issues: the role the library has played in the development of the modern university and the role of the librarian in that setting, and the development of librarianship as a profession (Hamlin,

1981; Shiflett, 1981). In this section, some of the historical discussions related to the development of the academic library are described. This entails an explanation of the meaning of the library's collections, as it has been understood and discussed in the last century, and an explanation of librarianship as a profession. These two factors, collections and the profession, contribute the most substantial practical and theoretical considerations in defining the academic library, largely because the development and meaning of the academic library's collection is closely intertwined with the development and meaning of librarianship as a profession. One does not make sense without the other.

The Academic Library and its Collections

Historically, the rise of the academic library in the United States began in the late 19th century. Wiegand (1990) argues that during this time an ideology of reading, of how and what to read, although often associated with early American public libraries (Ross, 2009), fostered the shape of scholarly communication and academic life. Through the first three quarters of the 19th century, college curricula remained fairly static. It demanded that students engage, memorize, and translate Greek and Latin works. For those managing libraries at the time, generally faculty and not librarians, this meant that collections need to only support a limited canon (Hamlin, 1981). According to Wiegand, this changed after two events: when Charles Darwin published the Origin of Species in 1859 and when the United States passed the 1862 Morrill Act, which set aside lands for colleges to study agriculture and the mechanical arts. In addition to the research library movement (Shiflett, 1981), these two events upset previous pedagogy and curricula, challenged established assumptions about the purpose of the academy, and contributed to a "culture [which] consisted of experts whose job it was to find new truths to replace the old authority patterns" (Wiegand, 1990, p. 74). Hence, the revolution involved developing and exploring new sources of data and methodologies, which led to an emphasis on the creation of new knowledge, which further led to new journals and eventually to new responsibilities for librarians, such as collection development.

For academic libraries, the focus on developing comprehensive collections continued through most of the 20th century, and the purpose assigned to academic libraries has rested on fundamental questions about what a collection is and how the items in the collection are transmitted, stored, and retrieved. In 1978, F. W. Lancaster published the controversial and discussion-provoking work Toward Paperless Information Systems (Lancaster, 1978). Lancaster predicted that by the end of the 20th century, automation and other technological developments would lead to a society where the primary mode of communication, and especially scholarly communication, would be electronic. Lancaster's argument, in part, arose from certain trends in academic libraries and scholarly publishing at the time. He noted that, through the early 1970s academic libraries were able to keep pace with the amount of published scholarship, at least in terms of titles if not volumes, but as the cost of serials and book titles and the personnel required to select, process, and maintain collections rose, this system could not be sustainable.

As a result of the creation of the web in the early 1990s and the rise of the delivery of published scholarship via this medium in the intervening years, Lancaster's prediction about a paperless society has turned out to be mostly true, in a complicated fashion, and has led to the formulation of the notion of digital collections. At the heart of the issue is the idea of a paperless society and the ubiquitous availability of personal search, retrieval, and storage devices, as

envisioned by Vannevar Bush (1945) and J. C. R. Licklider (1965). The question raised is whether this development renders the academic library obsolete if the need to develop and maintain comprehensive print collections is diminished.

The dawn of library automation in the 1930s (Black, 2007; Kilgour, 1939; Parker, 1936) launched an era of predictions about the future of academic libraries. After Licklider and others warned librarians about the potential implications of a paperless society and what that meant for libraries, Sapp and Gilmour (2002, 2003) noted that the literature written by librarians and library and information scientists began to shift focus away from collections to users. Instead of a future where "Libraries could not and should not expect to retain a monopoly over information" (Sapp & Gilmour, 2002, "The Next Decade in Academic Librarianship," para. 3), librarians should adjust to a future where information is decentralized and where other information agencies, including for-profit ones, have much more direct control over the dissemination of content to end users. Sapp and Gilmour (2002) write that, in 1985 Allen B. Veanor, a library consultant commissioned by the Association of College and Research Libraries (ACRL), argued that "The breakup of the academic library's monopoly on information inevitably would result in competition from external, non-academic entities. This would cause an increasing number of information resources to be marketed directly to the user" (para. 5).

Arguments about the competitive role of the academic library have been made more recently by others. Sennyey et al. (2009) describe changes for the academic library as it moves into a landscape dominated by digitized and digital collections. They note that digital and digitized content, and especially open access content, "creates a growing corpora that is accessible outside of the aegis of the library" (p. 254) and puts the academic library into a relationship with publishers and others in the scholarly communication system where they are expected to compete for patrons.

The rise of digital content and the orientation toward the library user have had an impact on what it means to collect. Harloe and Budd (1994) argue that content, and not packaging, should drive collection management. They make the case that the needs of the community are paramount, and, quoting Sheila Dowd (1990), write that,

> Bits and bytes of information are important only if the mind can link them with other pieces of information to build the orderly patterns that are fabric of knowledge. Hence the mission of the library is more properly identified as the provision of access to organized information, for the fostering of knowledge [emphasis added]. (p. 87)

Despite the cognitive and epistemological emphasis on what a collection means by authors such as Harloe and Budd (1994), others in the field continued to emphasize the importance of the physical collection. Carrigan (1995) argues that the primary purpose of the library is to offer certain benefits to its users, and the greatest of these benefits is its collection. He writes that "Libraries have multiple functions but all functions presume ultimate use of libraries' collections" (p. 100). This view highlights perhaps the most important premise held by academic librarians—that building collections is a library's primary duty.

Though the academic library has a contested definition and purpose, what is clear is that a balancing act exists between the role of developing or managing collections and the role played by librarians in the life of the user. Akeroyd (2001) argues that "It is all about becoming more

user centered and less collection focused or function dominated" (p. 82). In the same vein, Michalak (2012) describes the library at the University of North Carolina at Chapel-Hill as "outward facing" (p. 412), meaning that both collections and services have become less the purview of the library as content has become digitized and decentralized. Librarians there now spend more time going to the academic library user instead of waiting for the academic user to come to the library. Michalak finds that service follows the collection, and as the collection has become digitized and decentralized, so has the "service dynamic" (p. 413). Others have observed that the academic library is becoming more of a learning organization (Senge, 1990), and this not only has had an effect on the services offered but also on the organizational structure of the library, which is becoming more grounded in "information sharing, team-based structure, empowered employees, decentralized decision making and participative strategy" (Moran, 2001, p. 108).

Librarianship as Profession

Moran's (2001) and Michalak's (2012) observations reflect the changing role of the librarian in the academic library. While automation and digitization have had a substantial impact on what it means to collect and what the nature of a collection is, they have also influenced what it means to be a librarian. When Ralph H. Parker (1936) implemented the first library automation project in 1936, the goal was to pursue "a new day of no mistakes, no nervous strain, and much less manual labor for the library worker" (p. 905). Parker's motivation was to create a better working environment for the librarian, one that had a stronger intellectual base with fewer mundane tasks.

Despite such motivation, librarians have faced considerable obstacles in establishing themselves as a professional class. Part of the issue has been blamed on society's biases toward the feminization of the work (Mitchell, 2007) or the lack of self-esteem in a faculty dominated environment (Oberg, Mentges, McDermott, & Harusadangkul, 1992). Carpenter (1996) proposes that librarians have received less stature than faculty because their work has primarily been about the dissemination of knowledge rather than its creation. Carpenter's view reinforces Wiegand's (1990) discussion, cited earlier, of the 19th century change in the role assigned to higher education from colleges designed to disseminate classical knowledge to their students to universities charged to create new knowledge: "the more 'pure,' the more highly esteemed" (Carpenter, p. 87). In essence, knowledge creation replaced knowledge dissemination as the primary virtue of the academy.

The decentralization of digital collections and their accessibility outside the aegis of the library, the importance of the content of the collection rather than the format, reaching out to users rather than passively waiting, and the desire to professionalize librarianship imply that the competition for the attention of the patron lies with librarians and their ability to serve their communities in a way that addresses their mission rather than in a race to build bigger collections. Plutchak (2012) describes the strategic necessity of developing skills that best serve librarians' communities. He also argues that the tendency to personify the library, emphasizing the role of the library rather than the librarian, diminishes the importance of the librarian's role in identifying and disseminating information. In this context, it will be important to discern how important the role of disseminator of knowledge will be in the age of Google, whose role is not dissimilar, and in an academic system that is based on access to decentralized storage of content.

Alternate Discovery Services and the Decentralization of Collections

This section describes *Google Scholar*, which is used increasingly by researchers to search for relevant information. It also reviews the open access movement, outlining some of the reasons why it has become an important topic for both librarians and researchers. These are important elements in the scholarly communication system because it now seems possible and rational to use *Google Scholar* to acquire open access content in lieu of the academic library to acquire content from its collections.

Google Scholar: Alternate Discovery Services

Google Scholar (*GS*) has become an important bibliographic database and citation index as it has become more capable of indexing a comprehensive amount of scholarly documentation. Unlike Scopus or Web of Science, it is freely available to any user with an Internet connection. Furthermore, studies show that *GS* is perceived to be useful to end users (Cothran, 2011), is becoming a growing presence on academic library web sites (Neuhaus, Neuhaus, & Asher, 2008), and is becoming a preferred choice among academic library users, though more among those in the sciences and the social sciences rather than those in the humanities (Herrera, 2011).

According to a study conducted by Baldwin (2009), *GS* "indexes publisher web sites, PubMed Central (PubMed), institutional repositories, preprint archives, etc. It also locates full text results from research groups posting articles online for their own use and failing to make access proprietary" (Introduction section, para. 8). Baldwin's study suggests variability in sources used to retrieve full text documents depending on the type of article and subject matter being searched. For example, in a comparison between *GS* searches for mechanical and chemical engineering, a small percentage of the mechanical engineering full text articles were sourced from PubMed, whereas nearly half of the chemical engineering articles searched originated from there. Institutional repositories provided a nearly even balance between the two subject-based searches and a small percentage of the found mechanical engineering articles were sourced from publishers' open access sites compared to nearly a third of the found chemical engineering articles (p. 6). Additionally, Meho and Yang (2007), as cited by Harzing and Wal (2008), found a small overlap between the subscription databases Web of Science and Scopus with *GS*, suggesting that *GS* indexes material not found on either of the other major bibliographic databases.

In a study to evaluate "the breadth and scope of available content" on *GS*, Howland et al. (2009) found that "Google Scholar actually contained 76 percent of all the citations found in the library databases, while the library databases contained only 47 percent of the citations found in Google Scholar" (p. 231). While such studies suggest the strengths of *GS*, it should be noted that coverage of all disciplines is not universal. Kirkwood and Kirkwood (2011) find mixed results in *GS*'s coverage of historical scholarship, and institutional repositories using the Dublin Core Metadata Element Set can be overlooked by *GS* given certain deficiencies in the ability of Dublin Core to appropriately describe scholarly content (Arlitsch & O'Brien, 2012).

In an interesting and perhaps, within its very limited framework, successful attempt to measure recall and precision in *GS*, within the scope of the subject area searched ("later-life migration"), Walters (2009) found that "*GS* performs better than many subscription databases" (p. 16). In this study, involving a comparison of *GS* and 11 subscription databases, relevance was

defined as an assessment of "subject matter, importance of findings, innovativeness of methods or approach, number of other studies published on the topic, accessibility of content (readability), and accessibility of the document itself (availability to students and scholars)" (p. 7). One hundred and fifty five papers were selected for the recall and precision study. *GS* placed fourth in both recall and precision when evaluating the first 10 hits and moved to first place after 75 result hits. For the most part, the differences between first, second, third, and fourth places were trivial.

Open Access: The Decentralization of Collections

For the last 30-- 40 years, journal prices have increased at a rate that has been difficult for libraries to match. The end result is a situation librarians refer to as the "serials crisis" (Greco, Wharton, Estelami, & Jones, 2006). Some have argued or pointed out that part of the reason for the increased cost in journal prices is due to the costs involved in publishing both print and online formats (Fidczuk, Beebe, & Wallas, 2007; Kling & Callahan, 2003). Others have argued that copyright law creates a monopoly that allows publishers to charge exorbitant fees (Bergstrom & Bergstrom, 2006). While there are certainly other causes, the end result is a system that many believe is unsustainable.

Although academic libraries command a seemingly large budget for the acquisition of materials, the average annual price for serials has increased at a much faster rate than library acquisition budgets. For the 2010À2011 year, the Association of Research Libraries (ARL) reports that "total library expenditures of all 126 member libraries ... was slightly more than $4.6 billion" (Kyrillidou, Morris, & Roebuck, 2012, p. 5). Although this represents a billion dollar increase from six years earlier (Kyrillidou & Young, 2006), the portion of those funds that are spent on library materials is increasing to nearly half of all expenditures. Academic libraries receive more money, but a greater percentage is committed to materials (Bosch & Henderson, 2012; Budd, 2002; McGuigan & Russell, 2008; Romero, 2008).

Proponents of open access (OA) as a publishing model have argued that it can help alleviate the burden on academic libraries' serials and acquisitions budgets (Albert, 2006; Corrado, 2005). The ARL statistics highlight how this may yet be the case, and it may also depend on what type of OA model is pursued. OA exists in two broad forms: Gold OA and Green OA. Lewis (2012) describes the types of Gold OA models. "Direct Gold OA" pertains to journals that publish articles that are freely accessible to readers at the time of publication. Journals that provide access to articles after an embargo period are considered Delay Gold OA journals. Hybrid Gold OA journals give authors an option to pay a submission or publication fee. When authors pay this fee, their articles will be immediately accessible to readers even in journal issues that have articles that are not OA because other authors did not pay a fee.

Green OA, on the other hand, "sits alongside the subscription journal system and does not attempt to replace it" (Lewis, 2012, p. 494). This model is primarily about self-archiving the publication. Authors who take advantage of Green OA have several options for self-archiving. They may deposit a copy of the article's preprint or postprint version either on their personal web site or in an institutional or subject repository. Preprints are versions of the article that have yet to be peer-reviewed and postprints are versions of the article that have been peer-reviewed. Chan (2004) distinguishes between Gold and Green OA as open access publishing (OAP) and open access archiving (OAA), respectively (cf., Harnad et al., 2008). Both OAP and OAA models are

original definitions in the Budapest Open Access Initiative, released in February 2002, which provides the core definition of open access (Bailey, 2007). Other OA characteristics noted by Bailey include content that is freely available, is online, and has minimal restrictions for reuse. The reuse factor relates to copyright, which is often held by the author(s) of an OA work, and may be assigned a Creative Commons license.

Open access research largely focuses on three areas: the benefits to libraries in the form of journal cost-saving, the benefits to the public and to scholars in the form of increased access, and the influence of open access in terms of citation counts or number of downloads. While the first two types of research focus on the implications of open access for libraries and readers, those implications are often one-sided. That is, it is assumed that the benefits outweigh any costs, where the costs might be the marginalization of academic libraries, in terms of the decentralization of content storage, or some other unnamed implication. Drott (2006), for example, illustrates that "the emergence of the discussion of open access as a viable alternative to traditional publishing rests on developments in three main areas: economics, technology, and social justice" (p. 81). Thus, while OA's impact on libraries' budgets is often a major component of the discussion, the impact on the use of the library's collection is not.

Research that focuses on measuring OA's influence by comparing download and citation counts between open access and subscription-only articles or journals includes as its audience other researchers with interest in such measures for various reasons when deciding to publish in open access or subscription-based journals. Generally, this research suggests that open access articles have increased download rates, but there is no agreement that open access articles have a citation advantage—an increased likelihood of citability or an increased citation count. For instance, in a randomized controlled trial involving journals published by the American Physiological Society, Davis, Lewenstein, Simon, Booth, and Connolly (2008) found open access articles led to substantially increased downloads over subscription-only articles, with 89% more full text, open access downloads. However, they found that, after one year, the access level had little to do with citability: 63% of the subscription-only articles were cited and 59% of the open access articles were cited.

This finding is in direct conflict with Eysenbach (2006), whose study of the *Proceedings of the National Academy of Sciences (PNAS)* found that after a mean of 206 days plus 6 months after publication, subscription-only articles were less often cited than open access articles. Specifically, 51% of the subscription-only articles were cited in contrast to 63.2% of the open access articles. Eysenbach also found that open access articles saw a higher citation count as early as four months after publication. Between 6 and 10 months after publication, open access articles received average counts of 6.4 versus 4.5 for subscription-only articles.

However, Gargouri et al. (2010) found an open access citation advantage primarily for higher quality open access articles, which saw nearly an eightfold odds increase in citation counts. The study examined subscription-only articles, mandated institutional repository open access articles, and self-selected open access articles. It specifically compared subscription-only articles against self-selected open access articles, subscription-only articles against mandated institutional open access articles, and self-selected open access articles against mandated institutional repository open access articles. Gargouri et al. concluded that high quality articles see many more citations if the articles are open access. They also ruled out the argument that if open access articles see a citation advantage, it is because authors choose to make their best work

open access. Instead, they infer that there is a "quality advantage" due to "user self-selection" (Discussion section, para. 5) and not author self-selection.

Whether there exists a download or a citation advantage, these studies demonstrate OA's influence on the research front. However, the growing number of OA journals means that academic libraries do not always provide records to open access journals in their catalogs. Additionally, the main bibliographic indexes, including Web of Science, EBSCO Academic Search Complete, ProQuest Research Library, Biological Abstracts, and others do not always list open access journals, and those journals that are listed generally have privileged characteristics, such as high impact factors and high publication output per year; they may also be U.S. based and charge authors publication fees (Collins & Walters, 2010; Walters & Linvill, 2011a, 2011b). This suggests that much OA published content is left to be discovered by less discriminating services like *GS*. Despite the disagreement among the findings and the uncertain accessibility of OA content in library supplied databases, these studies do suggest that OA has an increasingly broader reach than articles that exist behind a pay wall and that this is in large part because services such as *GS* are good at locating OA content.

Theoretical and Methodological Basis for the Study

If researchers use nonlibrary services such as *GS* to acquire documentation such as OA content that does not necessarily have to be collected by libraries, this has implications for the academic library for several reasons. First, decisions about whether one begins a literature search on an academic library's web site or on *GS* are made based on perceptions about the likelihood of success and payoff, areas associated with decision and game theory. The purpose of this section is to show that it can be rational not to use an academic library's services and collections, meaning that the payoff for the scholar who uses nonlibrary services to retrieve nonlibrary collected documents is sufficient to encourage the continued use of those services. If academic librarians are to respond to these actions, the justified rationality of the searcher will have to be taken into consideration.

Since this study gathers data from a social computing web site where users of the web site collect and store bibliographic references, and since these bibliographic references are analyzed using bibliometric data collected from *GS*, theoretical discussions of bibliometrics, social computing, and what it means to collect bibliographic references follows. Essentially, while the act of citing a scholarly document with a bibliographic reference has been a primary object of study in information science for the last 50 years (Narin & Moll, 1977), the act of collecting and saving bibliographic references to scholarly documentation on social computing web sites is a relatively recent phenomena that is just beginning to be explored. However, citation theory may be used to build a framework for outlining what it means to collect a bibliographic reference in terms of the social activity involved with collecting. That is, it may suggest something meaningful about the document that is collected in a way that is analogous to the relationship that is inferred between citing and cited documents. Furthermore, the ability to collect these references on social computing web sites built for such purposes contributes a necessary theoretical part of this study. This ability is only possible and is only acted on because of certain technological affordances offered by these social computing web sites.

Decision, Game Theory, and Bounded Rationality

Decision theory describes those "situations where each person can choose without concern for reaction or response from others" (Dixit & Skeath, 2004, p. 18). Game theory describes those situations where decisions by a player interact with decisions made by other players. It has been used to explain topics in economics, political science, sociology, ethics, and philosophy (Binmore, 1994; De Bruin, 2005). Dixit and Skeath (2004) outline several components of strategic games. Players have strategies where these strategies are simply the relevant "choices available to them" (p. 27). The outcomes of a game are described as payoffs, and these are usually assigned some numerical score (such as the number of dollars awarded for some outcome). Additionally, the players are thought or assumed to be rational in that they seek to achieve the highest payoff. All strategic games have solutions which are described in terms of the game's equilibrium. An equilibrium indicates "that each player is using the strategy that is the best response to the strategies of the other players" (p. 33).

Game theory is applicable in a descriptive way (Cave, 2005). For example, it can assist in the identification of inherent preferences and it can help highlight barriers that prevent best strategic responses. For example, librarians prefer lower subscription rates for serials although they continue to pay higher costs. Game theory suggests either two analyses: (1) librarians are irrational because they choose to play with weaker or dominated strategies; or (2) librarians are rational but forced or coerced into playing with weaker or dominated strategies. If we accept that librarians are rational agents, then it seems likely that the second analysis describes the problem.

Game theory also offers insights into information seeking, but in such cases, it is important to address certain assumptions about rationality (Budd, 2012). Rationality is often assumed to mean that players in a game have complete knowledge of their own preferences and are able to perform "flawless calculation[s] of what actions will best serve those [preferences]" (Dixit & Skeath, 2004, p. 30). Additionally, it generally means that players will remain consistent about their preferences (see Ritzberger, 2002).

Consider, for example, the Ultimatum Game, where two players, a Proposer and a Responder, must decide how to split a pot of money. In this game, the Proposer offers a $20 pot of money to the Responder. If the Responder rejects the offer, neither receive any payoff. If the Responder accepts the offer, they receive a share based on the proposed split. Both players are aware of the rules. The rationality assumption often adhered to by game theorists means that even if the Proposer offers the Responder $1 in order to keep $19 for him or herself, the Responder will accept this offer since receiving some money is better than receiving no money. That is, the Responder is selecting his best strategy given the strategy selected by the Proposer. As such, a $1 and $19 split represents a solution to the game, otherwise referred to as its equilibrium. However, studies show that "the majority of Proposers offer 40--50% of the total sum, and about half of all Responders reject offers below 30%" (Nowak, Page, & Sigmund, 2000, p. 1773). Common explanations for this behavior incorporate notions of fairness, reputation, and retribution even though these represent affective states and social norms, rather than rational attitudes.

The same kind of rationality assumption can be applied to the study of scholarly information seeking. Consider that a researcher requires information about topic X. Simplifying the strategies available to the researcher, suppose that the researcher has two options: one based

on using library resources to acquire a document about X and the other using nonlibrary resources for that purpose. The payoff for either strategy is access to a relevant document about X. Though the payoff is the same for both strategies, the difference between the two strategies lies in the researcher's costs in terms of time, knowledge, or frustration with the retrieval systems. The rational course would have the researcher always using that strategy which will cost him less, given that the payoff is the value of the relevant information minus the cost in acquiring that information.

How a researcher may choose between these two options may depend on what he believes is the best option. Psychological game theory (Dufwenberg, 2010) suggests that "belief-dependent motivations," where the game's payoffs "are defined on beliefs (about actions and beliefs), as well as on which actions are chosen" (p. 272), might shed light on the researcher's strategy profile given his or her prior beliefs about any given strategy. If a researcher believes Google is great, based on past experience, then he or she may be more likely to use Google in the long run (likewise with a library resource). This can be problematic for some library systems if they have failed users in some way (e.g., Kress, Bosque, & Ipri, 2011; Yadamsuren, Paul, Wang, Wang, & Erdelez, 2008).

This problem may further hinge in this case on the researcher's perception of the cost of either service. Zipf (1949) might argue that the perceived cost will be dependent on both the amount of work involved in using either service and the researcher's estimate of the probability that he or she will depend on either service over the long run. For Zipf,

> The most that any individual can do is to estimate what his future problems are likely to be, and then govern his conduct accordingly. In other words, before an individual can minimize his average rate of work-expenditure over time, he must first estimate the probable eventualities of his future, and then select a path of least average rate of work through these.
>
> Yet in so doing the individual is no longer minimizing an average rate of work, but a probable average rate of work; or he is governed by the principle of the least average rate of probable work.
>
> For convenience, we shall use the term least effort to describe the preceding least average rate of probable work. (p. 6)

The least average rate of probable work is expressed by the ability to solve problems and apply search heuristics given our limited computational abilities. Herbert Simon's (1990) notion of bounded rationality is a good extension of Zipf's principle in the sense that our "computational limitations," in tandem with the characteristics of the systems we use to search, result "not in optimizing techniques but [in] methods for arriving at satisfactory solutions with modest amounts of computation" (p. 11). This is simply another way of saying at little cost or "least average rate of probable work." Simon argued that we do not maximize our utilities; rather, due to our limitations and our settings, we simply attempt to satisfy our preferences in whatever way we can to reduce our computational load, thereby incurring less cost.

What is satisfactory simply refers to what is most probable or what is believed to be most probable, given the work involved and the setting of the work. When making a decision, a person

has at least two options to consider, two ways to act, in order to achieve some outcome. If the person is rational, he or she will choose the act that will most likely result in the desired outcome. If a person requires a journal article and has before him or her several paths to acquire it, then it is assumed that person will choose the path that he or she believes will most likely have the desired result with the least amount of effort.

Bibliometrics and Citation Analysis

Broadus (1987) defines bibliometrics as the "quantitative study of physical published units, or of bibliographic units, or of the surrogates for either" (p. 376). White and McCain (1989) note that "bibliometrics is to publications as demography is to peoples" (p. 122). If this is so, then that data that composes the bibliometric study defines and sets its boundaries. Often, researchers gather bibliometric statistics from citation lists generated by bibliographic databases such as those provided by Thompson Reuter's Institute of Scientific Information (ISI) indexes (e.g., Web of Science). More recently, interest has risen in Elsevier's Scopus and Google's *Google Scholar* as sources for both bibliometric and citation analysis (e.g., Falagas, Pitsouni, Malietzis, & Pappas, 2008; Harzing & Wal, 2008; Howland et al., 2009; Noruzi, 2005; Yang & Meho, 2006). While these data sources differ in scope, they both seek to capture formal scholarly communication (Wouters, 1998), authenticated or authorized as such in some standard fashion, and to enhance an understanding of the relationships between authors, journals (or other formats), and their communities.

As methodologies, bibliometrics and citation analysis have been used for a variety of purposes, including developing and testing certain theories (see Borgman & Furner, 2002; Bornmann & Hans-Dieter, 2008; Brookes, 1969; Cronin, 1984). They have an object of study, the publication as a whole and its various components including authorship, the byline (Cronin & Franks, 2006; Cronin, Shaw, & La Barre, 2003), the reference, and the citation. They have a way of going about what they study—their methods, which may include counting citations, examining author co-citations, and analyzing bibliographic coupling relationships. The motivations for these studies may be practical. For example, McCain and Bobick (1981) used citation analysis to study journal use in an academic library. More recently, Enger (2009) used citation analysis to study core book collections in an academic library in order to enhance collection development activities.

Nicolaisen (2003) writes that "in order to understand, explain, and predict the dynamics of citation networks, we need to penetrate the social worlds of individual authors" (p. 18). This is also true of bibliometrics in general. The problem is not uncomplicated. While penetrating the social worlds of scholars and scientists may be difficult, advances in social computing technologies (O'Reilly, 2005) offer insights into these social worlds as well as the variety of research traditions that exist around them. Importantly, these insights may be derived from the "empirical grounding" Nicolaisen seeks from a social theory of citing and, by extension, bibliometrics too. Specifically, this empirical starting point may lie at the intersection where social computing and bibliographic reference collecting converge and may exist to supplement the empirical grounding of more traditional sources such as the Science Citation Index (SCI), as historically outlined by De Bellis (2009). Thus, web-based applications such as *CiteULike*, BibSonomy, and others, where users collect, store, tag, and share bibliographic references, serve

as likely candidates of attention. As Cronin (2001) noted, “the web has challenged, and may revolutionize, many of the assumptions that have underpinned the established scholarly communication system” (p. 3) while enabling us “to detect early signs of emerging trends” (p. 6).

Social Computing

If the web revolutionizes assumptions about scholarly communication, alerts us to emerging trends, and alters our actions, habits, and behaviors, then it does this most effectively through social computing in general, and, in particular, to two important attributes of this phenomenon: affordance and place (see also Pomerantz & Marchionini, 2007). Dourish (2001) defines affordance with regards to social computing, humanÀcomputer interaction, and system design as a “a property of the environment that affords action to appropriately equipped organisms” (p. 118). Affordance theory suggests that a social computing application functions as an “artifact,” or more broadly, as an “environment,” that offers those features that enable and “afford particular sorts of actions” (p. 185). Affordance is fostered by a social computing application’s use of place, defined as a social environment in contrast to its locational characteristics. Thus, affordance theory allows us to understand how the environment plays a role in researchers’ decisions to use library and/or nonlibrary discovery services to obtain OA documents.

According to Dourish (2001), the concept of place leads to several substantial sociological consequences. The first consequence is highlighted by the difference between the terms place and space. A place directs our attention away from the environment as simply a structure and toward the environment as a social sphere. Hence, the structure of the surroundings disappear into the background as the space becomes used. Often a “‘place’ reflects the emergence of practice” (p. 90), by which Dourish means that a place is customized and shaped just as we may rearrange the chairs in a room according to how we use the room. In the same vein, a place may mean different things to different community of practice, so one particular setting may have multiple meanings depending on how it is used.

These insights about social computing provide the necessary framework for understanding scholarly communication. In particular, a social computing application’s structure and functionality affords the tools necessary to create a space where users converge through common practice. When these events overlap at a place where the practice concerns scholarly and scientific bibliographic references, the social worlds of authors, scholars, scientists, and readers become more accessible to researchers interested in the sociological aspects of scholarly communication as well as the quantitative techniques used to measure it.

Collecting Bibliographic References: Social Computing and Bibliometrics

White and McCain (1989) write that “bibliometrics is grounded in the patterned behavior of human beings—the authors, editors, and indexers on the production side of the world of learned publications. Specifically, it is grounded in the linguistic choices by which they associate indicators of content” (p. 123). They mark a distinction between authors, editors, and indexers on the production side and readers or users on the consumption side. While they also write that “bibliometrics can deal only with explicit data” (p. 164), the data provided by bibliographic

reference management social computing applications, about what is collected and possibly read by scholars, makes explicit what was previously unavailable in quantitative aggregate. Essentially, the bibliographic references and papers scholars collect provide new insights into how traditional bibliometric data is used after it has been extracted from subscription databases and the newer, nontraditional, more complicated sources traced and predicted by Cronin (2001).

The online availability of bibliographic records along with the growth in interactive digital libraries has resulted in a new blend of these facets of information science. Users, by providing content, become producers in some sense, and in the scholarly setting, the production and consumption of bibliographic records merges with the authors, editors, and indexers on the publication side and with the readers and users on the consumption side. That is, the readers or users of published scholarly and scientific literature now also produce "the linguistic choices by which they associate indicators of content" with articles and other writings, which are the "true unit of analysis in many bibliometric studies" (White & McCain, 1989, p. 124).

Readers and users contribute to the production side in two significant ways: by selecting, saving, and building second-tier databases of bibliographic records and by tagging them with keywords. The outcome of this activity is the creation of systems, such as *CiteULike* or Mendeley, that highlight different aspects of information retrieval and information needs and uses as identified by White and McCain (1989). These databases are different from other databases that are traditionally used in bibliometric studies like the ISI indexes, Scopus, and, more recently, Google Scholar. Rather than attempts at storing, organizing, or simply linking to the entirety of scholarly and scientific publications, or some authenticated set of it, these databases (or indexes) are the result of user and/or reader production and therefore consumption-side aggregated value. It is this phenomenon of readers as indexers and what it may reveal about the social world of scholarly communication that is the indirect fuel for this study and the bibliographic references produced by these tools that is its object.

It is important to note that users collecting, storing, sharing, and tagging bibliographic references in such web-based social computing applications are not instances of citing behavior. Citing is a norm which acknowledges "the work of those who have gone before" (Budd, 1992, p. 348), and citations may be seen, metaphorically, as "signposts" (Smith, 1981, p. 85). In contrast, there is no such permanence involved in adding bibliographic references to online personal yet public digital libraries which may later be deleted. While these bibliographic references do act as a sort of acknowledgment, they do not necessarily act as a sort of acknowledgment in the sense that a citation does, given that they are not situated within published discourse, specifically grounded in argument, or directly serve to promote scientific or scholarly progress based on traditional forms of inquiry.

Adapting and modifying three of Smith's (1981) list of five assumptions of citation analysis, we wonder whether (1) collecting a bibliographic reference to a document implies use, or potential use, of that document by the person collecting it; (2) collecting a bibliographic reference to a document reflects the merit of that document; and (3) users are collecting bibliographic references to the best possible works. With regards to Smith's third point, she writes that a number of other factors influence citing behavior and these may include access to the document and awareness of the document. If access to a document is a factor in whether that document gets cited, then an examination of access levels in an OA world is important.

CONCLUSION

While many perceive the purpose of academic libraries to be collecting, organizing, and providing access to the scholarly record, not all within the profession or the research community agree on the specifics. However, even if collecting, organizing, and providing access to information is the primary purpose of the academic library, the academic library is no longer the sole or primary actor with this function. New sources to discover scholarly information and new publishing models place the academic library in competition for the attention of users.

This study merges several theories to answer its research questions. Collecting bibliographic references using bibliographic reference management services such as *CiteULike* allows us to work with new data types. Although these data exist in the familiar form of a bibliographic reference, they represent a different activity. Rather than being instances of citing, they are instances of collecting, and studying them is possible because of advances in social computing. Using decision and game theory, as well as notions of rationality, we can infer from this activity the strategic impact these collecting actions have on academic libraries while still holding some of the assumptions of citation analysis true.

PROCEDURES

Introduction

This study proposes examining the properties of bibliographic references scholars and researchers collect and using a freely accessible bibliographic database to examine additional statistics about these references. The research questions are:

> **RQ 1.** Is the current state of affairs, at the network level, such that nonlibrary electronic discovery services marginalize academic libraries?
>
> **RQ 2.** Does open access content, in conjunction with nonlibrary electronic discovery services, marginalize academic libraries?

The first section of this chapter describes the sources of data, in this case, *CiteULike* and *Google Scholar*, both tools used for the bibliometric and regression analyses. The next section describes the logistic regression method, used here to determine what predictor variables predict access to full text documents outside of a library's proxy. The third section describes the Bayesian probability method, used with the findings of the Ithaka S + R study (Schonfeld & Housewright, 2010) to determine a hypothetical probability that a library's discovery services and collections were used to find information or documents even though an option to use an alternate discovery service and an alternate collection was available. The fourth section describes the data collection process. This is followed with a description of the variables used from the *CiteULike* and *Google Scholar* data. This chapter concludes by outlining the plan of analysis.

Data Sources

The bibliometric and regression analyses are conducted on data collected from *CiteULike* and *Google Scholar*. *CiteULike* provides the bibliographic references and *Google Scholar* provides bibliometric data. This includes citation counts and item sources.

CiteULike

CiteULike has been an object of study and a source of data for studies. It has primarily been used by those interested in folksonomies and tagging (Capocci & Caldarelli, 2008; Kipp, 2011). As of October 2008, less than two years before collecting data for this study, *CiteULike*.org had "885,310 unique items, annotated by 27,489 users with 174,322 unique tags" (Bogers & van den Bosch, 2008). At least one study used *CiteULike*, along with two other social bibliographic reference managers, as a source to analyze journal usage (Haustein & Siebenlist, 2011).

CiteULike users may add bibliographic references to their libraries either manually or automatically. In the latter case, adding a bibliographic reference to a personal library is accomplished either via a JavaScript bookmarklet for the browser or through a social bookmarking link on a scholarly document's web page (CiteULike, 2010b). The bookmarklet or bookmarking link will extract bibliographic data from an appropriate web page and import the bibliographic details into its database. Users can assign tags to their references, and these will function as a type of "flexible filing system" (Emamy & Cameron, 2007, para. 6). Users may also assign additional metadata, and this includes noting whether the reference refers to the user's own publication (authored), the priority assigned to the publication, and whether collecting the reference is public or private (default is public) information. Users may also add notes via a simple text editor in the browser and write a review of the publication. Users may view related articles based on the tags that have been assigned by the user. *CiteULike* will generate a formatted reference on command in a number of styles including APA, Chicago, IEEE, Harvard, and others. Finally, users may export their libraries in various formats, either for generating formatted references or for importing into other bibliographic reference manager applications.

CiteULike offers a number of social functions. Users may connect with other users and join groups of users who may be interested in similar research or who are working together on a research project. Users can share bibliographic references and write blog entries about those references within the site. Users may also create personal profiles of themselves where they can provide details such as their name, email, location, job title, affiliation, web page, and research fields.

Google Scholar

Google Scholar was introduced in 2004 and has since grown in popularity on several fronts. Research has been conducted on its use and popularity as a search tool among students (Cothran, 2011; Herrera, 2011) and by librarians (Neuhaus et al., 2008), its ability to index content in institutional repositories (Arlitsch & O'Brien, 2012) or to locate open access content (Norris et al., 2008), and its scope (Chen, 2010) and coverage in various subject areas such as history (Kirkwood & Kirkwood, 2011) and engineering (Baldwin, 2009).

Some studies have used *Google Scholar* as a bibliometric or informetric tool, where the latter methodology refers to a broader notion of bibliometrics and means "the quantitative study of recorded discourse" in any medium (Wolfram, 2003, p. 39). Kousha and Thelwall (2007) compare *Google Scholar* to the ISI indexes. Noruzi (2005) provides an introduction to *Google Scholar*'s use as a citation analysis tool. Harzing and Wal (2008) describe Google Scholar as a citation analysis tool and offer a free program that uses *Google Scholar* to compute alternative journal impact scores and other citation measures (see Publish or Perish at

http://www.harzing.com/ pop.htm). However, Aguillo (2012) conducted a webometric analysis and concluded that *Google Scholar* is a problematic source for bibliometrics because its coverage lacks quality control.

Despite Aguillo's (2012) concerns about the quality of sources *Google Scholar* indexes, *Google Scholar* is a useful bibliometric tool in this study for two main reasons: (1) because it is used to locate known bibliographic references that have been saved by users in *CiteULike* and (2) because *Google Scholar* functions as the relevant alternative to using the academic library as a research starting point. This study, therefore, depends on *Google Scholar*'s increased coverage over subscription bibliographic databases such as Scopus and Web of Science since the references that *CiteULike* users save may themselves be more comprehensive than what the more selective bibliographic databases cover.

Google Scholar offers a number of functions including the ability to locate scholarly works, either through simple or advanced searching, export citations to those works, provide total counts of citations, search within works that cite other works, and link to the full text of works if the full text is available and indexed by *Google Scholar*. In the latter case, the hostname providing the full text is provided by *Google Scholar* as a hyperlink to the full text. For example, a full text document with a link to the hostname umsystem.edu likely refers to the University of Missouri's institutional repository at mospace.umsystem.edu.

Libraries can use a link resolver to allow *Google Scholar* to provide access to subscribed content (Google, n.d.). When libraries configure and use this service, *Google Scholar* seamlessly integrates with the library's collections. This works for the users of a particular library who use *Google Scholar* within an authenticated Internet Protocol (IP) range, usually that of a university's network. In such cases, it will be necessary for patrons to use *Google Scholar* on campus or, if off campus, through a virtual private network (VPN) connection.

Logistic Regression

One of the variables in this study is whether *Google Scholar* points to full text article copies of the bibliographic references in the *CiteULike* sample. This variable is a binary or dichotomous data type (Yes/No) and is a candidate as a dependent variable in a logistic regression. A logistic regression tests how a set of predictor variables affects or is related to a binary or dichotomous variable (Harrell, 2001). Logistic regression does not assume a normal distribution or linear relationships between the variables (Sin & Kim, 2008). However, a logistic regression requires meeting four conditions: multicollinearity, independence of errors or cases, linearity of the logit, and no complete separation, which means any one variable should not completely predict any of the other variables (Field, Miles, & Field, 2012). However, separation is generally only a problem when there are multiple categorical or dichotomous variables (Boslaugh, 2012). When the independent (predictor) variables are of the same data type (e.g., ratios), multicollinearity becomes a concern (Adkins & Bala, 2004; Sin & Kim, 2008). There is no test for independence of errors, which assumes that variables are not related. Testing for the linearity of the logit requires modeling the logistic regression and including an interaction between any continuous predictor variables and the log of itself (Field et al., 2012).

The predictor variables may include both categorical and continuous data (King, 2008), and this study includes the number of authors for each bibliographic reference (author count), the

year the bibliographic reference was posted to *CiteULike* (post year), the publication year of the reference (pub year), and the citation counts. The post year variable is unique to this study and to a bibliometric analysis. It represents the social computing nature of *CiteULike*. Based on these variables and the more general theoretical motivations described in this study, the logistic regressions address whether the variables in the data set predict full text availability in *Google Scholar*. The model produces an odds ratio (OR) for each of the independent variables in relation to the dichotomous dependent variable. This reflects an overall effect size (Harrell, 2001).

The OR is perhaps the most important statistic, at least for interpretation, resulting from a logistic regression. It is the result of dividing the odds of one group by the odds of a second group and is interpreted by reference to the numerator. For example, "odds ratios of 2, 0.5, and 1 indicate, respectively, that the odds of the group in the numerator are 100% larger (doubled), 50% smaller (halved), and neither larger nor smaller than the odds of the group in the denominator" (King, 2008, p. 366).

Bayesian Analysis

The 2009 Ithaka S + R faculty survey (Schonfeld & Housewright, 2010) found that 38% of scientists claim to begin their information seeking with Google, and from that statistic and others like it, the authors concluded that academic libraries are increasingly being disintermediated from the discovery process. The problem is that this claim does not take into consideration the alternate route. That is, if 38% of scientists use Google as a starting point for their research, then we might say, broadly speaking, that 62% of scientists use the academic library as a research starting point. Although the complement claim is a simplification and a broad assumption and the real world choice or sample space certainly does take into consideration other discovery mechanisms, such as invisible colleges (Price, 1986), the decision between the two represents a near world scenario. Contrasting them provides a way to outline the theoretical upper and lower bounds of the model.

Additionally, a set of conditionals relating to the success rate of either the academic library or *Google Scholar* in retrieving relevant full text documents is necessary in order to make a claim about the disintermediation of the academic library. That is, before a valid claim about the disintermediation of the academic library can be made, we must determine the probability of retrieving a relevant full text document. So, the meaningful question is, given that 38% of scientists use Google as a research starting point, what percentage of those scientists hypothetically experience successful retrieval events of relevant documents outside of a university's proxy? Bayes' theorem allows us to invert this question in order to determine the probability that a scientist who used an academic library (or *Google Scholar*) as a starting point then retrieved a full text document. If we address that question, then we address the claim about the disintermediation of the academic library.

More pointedly, we can ask what is the probability of having used an academic library as a research starting point given having retrieved a relevant full text document. Bayes' theorem does not allow us to compute this without taking into consideration all the relevant decisions or events. Such that, we have to know the joint probability of having retrieved a relevant full text document outside of a university's proxy by using *Google Scholar*. We also have to know the joint probability of having retrieved a relevant full text document having used an academic

library. It is not enough to know how successful the academic library is in aiding a searcher in retrieving a relevant full text document without taking into consideration how successful *Google Scholar* is also, given that these are the two broad options available to researchers, as the Ithaka report suggests.

Data Collection

After receiving approval on May 18, 2010 from *CiteULike* to access their data, their entire data set was downloaded on May 19, 2010 in two separate files. These files contained identification numbers for each of the references in the *CiteULike* library and amounted to 2,419,452 unique bibliographic references (*CiteULike*, 2010a).

These identification numbers were sorted and used for the systematic random sampling (Vaughan, 2001; Vaughan & Shaw, 2008). The count of the unique bibliographic references was divided by 000. This resulted in the number 2419. A random number was generated (4438), and starting at this number, which indicated the 4438th bibliographic reference in the data, every 2419th identification number was harvested. This resulted in a sample size of 999 bibliographic references.

Each identification number in the sample was manually used to retrieve the bibliographic reference from the *CiteULike* web site in the BibTeX format, a format that provides standard bibliographic data. Four of the 999 references in the sample were irretrievable for indeterminate reasons.

Using *Google Scholar*, bibliometric and publication data was retrieved on July 14, 2010, July 17--19, 2011, and July 14À16, 012. *Google Scholar* was used to collect data on the following variables: found (yes/no), citation count, full text access (yes/no), and full text source. Some of the bibliographic references referred to simple web pages and there were some instances when *Google Scholar* found a citation one year but not the next. Also, the search was conducted outside of the university's proxy or network. This insured that full text sources, found outside the subscription pay wall, are truly full text sources. However, not all links were tested, and it is possible that some of these links were broken. This is a limitation of the study.

Description of Variables

The data sources are *CiteULike* and *Google Scholar*. *CiteULike* provided the initial data set of bibliographic references. The variables from *CiteULike* include:

1. Document type: Includes the type of document found in the sample of bibliographic references. This includes the common formats: journal articles, proceeding articles, and books.
2. Posted year: The year the bibliographic reference was posted to *CiteULike* by a *CiteULike* user.
3. Published year: The year the bibliographic reference indicates the source was published.

The variables from *Google Scholar* include:

1. Citation count: The number of citations *Google Scholar* shows for each bibliographic reference.
2. Found: This variable indicates whether *Google Scholar* was able to find the bibliographic reference and return a link or a citation to it. The result is either true or false.

3. Full text access: Whether *Google Scholar* was able to find a full text copy of the source. We use the term full text and not open access because we do not make any assumptions about the licensing status of the document. The result is either true or false.
4. Full text source: If a full text document was found for the bibliographic reference, this indicates the source providing the full text. Such sources may include institutional repositories, open access journals and databases, academic portfolio web sites, preprint archives, or others.

Plan of Analysis

The majority of the sample of bibliographic references pointed to the journal article document type and most of the analysis is on this document type. The analysis begins with a description of the overall sample. This is followed by a bibliometric analysis, which includes *CiteULike*'s coverage, *Google Scholar*'s full text retrieval rate, sources providing full text retrieval, and a citation analysis. Then two logistic regression models are built. These models test for factors that explain the full text availability of articles found using *Google Scholar* based on the sample drawn from *CiteULike*. Finally, the analysis ends by applying Bayes' theorem to assess the hypothetical probability that the academic library or *Google Scholar* was used as a research starting point.

The bibliographic references collected from *CiteULike* were saved in a spreadsheet file. Data collected from *Google Scholar* was added to this file under additional columns. The data was then cleaned and exported to a comma separated value (CSV) file and imported into RStudio (http://www. rstudio.com/), an integrated development environment (IDE) for the R programming language (R Core Team, 2012). The R programming language was used for the analysis along with several packages that extend its functionality. These packages include ggplot2 (Wickham, 2009), reshape2 (Wickham, 2007), and lubridate (Grolemund & Wickham, 2011). All software used is free and open source software.

RESULTS

Introduction

The purpose of this analysis was to determine whether the academic library is being disintermediated by researchers' information discovery processes and the decentralization of scholarly content, and consequently, risks marginalization. It is true that scholarly information seekers have many tools available to them to query information systems and retrieve the documents they need. Since not all these services or collections are provided by the academic library, as was the case for much of the 19th and 20th centuries, the data analyzed in this chapter should shed light on the impact both libraries and other services and sources have on the current state of affairs.

Bibliometric Analysis

CiteULike's Coverage

CiteULike users appear to collect a great variety of document types including journal articles, books, proceeding articles, and so forth. However, some document types are more abundantly collected than others. As seen in Table 1, a majority of documents retrieved in the sample are journal articles (69.45%), followed by books (8.94%), and proceedings articles (8.94%). Since the article document type dominates the sample, and because issues with open access largely concern journals (although not necessarily), much of the analysis that follows focuses on the references to articles.

CiteULike users have collected articles from as early as 1904, but most of the articles were published in the last 10 years. Additionally, starting with articles published in 2007, the frequency of freely available or open access articles in the sample is greater than those that are not available (Fig. 1).

[INSERT TABLE 1 ABOUT HERE]

Google Scholar's Coverage

Since the validity of *Google Scholar* as a bibliographic database is pertinent to this study, it is important to know how well *Google Scholar* located the items in the sample. In 2010, Google Scholar located 648 out of the 691 references to journal articles. In 2011, the retrieval rate increased to 663 and dropped to 662 in 2012.

Full Text Access

Controlling for the relative yearly increases in the bibliographic references that are discoverable through *Google Scholar*, the increase in full text access from 2010 (345/648) to 2012 (381/662) is 8.10%. By the year 2012, when 381 out of 662 full text access articles, or over 57%, were found by *Google Scholar*, the difference between freely available and not became statistically significant. Essentially, holding a sample of bibliographic references to articles constant, the probability that a user will be able to retrieve a full text copy from *Google Scholar* without the benefit of a university's proxy increases by 2012.

[INSERT FIGURE 1 ABOUT HERE]

Full Text Sources

The number of full text articles that are freely available through *Google Scholar* appears to be a function of the number of unique sources providing full text access. In 2010, 176 unique sources provided full text access to 345 articles via *Google Scholar*. In 2011, the number of unique sources increased to 190 and these sources provided access to 364 of the articles in the sample. In 2012, 229 unique sources provided access to 381 articles. Overall, this represents a 29.94% increase in the number of unique sources providing full text access, from 2010 to 2012, and a 8.10% increase in the full text articles that are available, after controlling for differences for each

year's total sample. Dividing these numbers by the three-year time period implies that for every .98% point increase in the number of unique sources, there is a .70% point increase in the number of full text articles that are available. Thus, as scholarly sources of information become more decentralized and grow in number, the probability increases that full text material (e.g., open access articles) identified in *Google Scholar* will be accessible outside of a university's proxy.

All sources providing full text access to the articles in the sample were examined by frequency of source and by type of source. For example, in 2010 *Google Scholar* linked to CiteSeerX to provide the majority of full text access to articles, but by year 2012, *Google Scholar* linked to CiteSeerX for just five articles. The remaining unique sources hold fairly steady across the time period. Lastly, 4 of the top 10 sources reference full text articles under the Green OA publishing model (e.g., open access institutional repositories) while 6 link to full text articles under the Gold OA model (i.e., open access journals) (Table 2).

Classifying these sources involved decision-making. For example, NIH. gov was classified as a government source and France's multidisciplinary open archive HAL (http://hal.archives-ouvertes.fr/) was classified as a national source. If the source was affiliated with a university, it was classified as a university source. However, if it was a faculty's vanity web site that was hosted on a university's server, it was classified under personal files. The Universities category includes institutional repositories, subject repositories that are operated by universities or university libraries (e.g., arXiv.org), and departmental or research group sites. For-profit and nonprofit journal publishers were classified as publisher files. If the source was affiliated with an academic or professional association, such as the American Psychological Association, it was also classified as a publisher file. In order to maintain consistency, all sources for all three years of data were classified at the same time, in mid-January 2013 (see Burns, 2013, appendices A, B, and C for the full list of sources).

[INSERT TABLE 2 ABOUT HERE]

The classification shows that universities, primarily including institutional and subject repositories, remain significant points of access for full text documentation. Table 3 provides a breakdown of the unique sources providing full text access. Most significantly, universities account for 56.82% of the unique sources providing full text access to articles in 2010. By 2012, this had increased to 63.32%.

Table 4 provides a breakdown of the number of documents to which each unique source provides access. Although it could be true that a small number of unique source types provide access to a majority of the documents, it does not hold true here. For example, although government agencies only account for a small percentage of the unique source types providing full text access, it would be possible that this source type provides a large percentage of the documents. However, the data indicates varied relationships. For example, Tables 3 and 4 show that in 2010 four unique government sources provided full text access to 39 articles but 100 universities provided access to 183 articles.

[INSERT TABLE 3 ABOUT HERE]
[INSERT TABLE 4 ABOUT HERE]

Citation Analysis

The median citation counts show fairly substantial increases over the threeyear time period, from a median of 23 in 2010 to a median of 37 in 2012. Table 5 shows that most of the references to articles that *CiteULike* users collect may be considered low to moderately influential, with respect to citation counts. Table 6 illustrates this further and shows that most articles have a citation count equal to less than half the cumulative percentage of sampled articles. In short, both tables highlight how the majority of articles that *CiteULike* users collect have very few citations in proportion to the highly cited articles. This shows that *CiteULike* users collect references that have a broad range of impact, the majority of which may be considered low impact. Given this, *CiteULike* users tend to function like a library by collecting and curating articles that have a broad range of appeal and not just articles that are popular.

[INSERT TABLE 5 ABOUT HERE]
[INSERT TABLE 6 ABOUT HERE]

A citation difference exists between articles that are available full text via *Google Scholar* and articles that are not available because they may be behind a pay wall. Although the data shown in Table 7 indicates no statistically significant difference between full text availability of article counts for the 2010 measures, as given in Table 8, there is a substantial difference between median citation counts in 2010 when the function is open access status. Specifically, articles that were referenced in the *CiteULike* sample and for which full text was not available via *Google Scholar* in the year 2010 had a much lower citation count compared to articles that were available. Furthermore, the spread widens as the articles age.

[INSERT TABLE 7 ABOUT HERE]
[INSERT TABLE 8 ABOUT HERE]

Logistic Regression

Although it appears that what is influencing full text availability via *Google Scholar* outside of a university's proxy is both the number and type of sources providing full text availability, the citation difference between full text and non-full text documents and the dispersion and growth over the three years suggests a positive relationship between higher citation counts and full text availability. To test whether citation counts predict full text availability, plus other variables that might be a factor, logistic regression was used to model these influences.

The logistic regression models show the influence of several predictor variables on a dichotomous dependent variable. The predictor variables include author count, publication year, post to *CiteULike* year, and citation count. The dependent variable includes full text availability. Although three years of citation data were collected, since high citation counts may indicate the following year's full text availability, only two years were modeled.

Tables 9 and 10 present the summary statistics for both logistic regressions. All assumptions have been met and both models show that they have value predicting outcomes (Field et al., 2012). Table 9 lists the predictor variables on the 2011 full text availability variable. The post year's relationship to the availability of full text in *Google Scholar* is not statistically

significant. Table 10 lists the predictor variables on the 2012 full text availability variable. This time the odds ratios for author count and post year are not statistically significant but the publication year and the citation counts for 2011 are.

[INSERT TABLE 9 ABOUT HERE]
[INSERT TABLE 10 ABOUT HERE]

To determine the influence of the statistically significant variables, the OR was used to calculate the difference between variables at different points (Boslaugh, 2012). For example, the OR for 2011 full text author count is 1.0928, which suggests that the more authors an article has, the more likely the article will be available full text. The predicted change in the odds of an article with an author count of five compared to an author count of one is 1.4261. Although citation counts have a much greater range than author counts, the influence is controlled by the relatively neutral odds ratio for the 2010 citations counts. Consider the predicted change for an article with a citation count of 101 compared to an article with a citation count of one: 1.0015100 = .1617. Thus, citation counts (or high impact articles) do not seem to influence the collecting of open access or freely available journal articles.

Table 11 summarizes the predicted probabilities (Boslaugh, 2012). In essence, when all variables are held constant at the first quartile mark, the 2011 model suggests there is 49.59% probability that the article will be available full text through Google Scholar outside of a university's proxy. This increases by nearly five percentage points for the 2012 model. When the values are held constant at the third quartile mark, the predicted probability increases substantially. In the 2011 model, there is a 60.82% probability that an article will be available full text and 63.34% probability it will be available full text in 2012. Since not all the odds ratios are statistically significant for each model, caution is advised before accepting them wholesale. However, the models do suggest that as each variable increases in count, the probability that an article will become available full text increases over time.

[INSERT TABLE 11 ABOUT HERE]

Bayesian Hypothetical

The data collected from *CiteULike* and *Google Scholar* indicates the probability of retrieving full text documents identified as relevant to the users that have collected references to them. In other words, if we assume that the bibliographic references collected by *CiteULike* users represent documents that they deem relevant and since we can determine how many of those documents can be retrieved from *Google Scholar*, we can infer the probability of retrieving a desired full text article given having used Google Scholar as a research starting point.

The success with information retrieval given the use of a service like *Google Scholar* or the academic library can also be used to determine the likelihood of how many *CiteULike* users started their research with *Google Scholar* or the academic library. This kind of Bayesian inference directly addresses the disintermediation issue. It takes information about our two data points, research starting points and information retrieval rates given the research starting points,

and derives from that a conclusion about the inverse of that conditional: that is, the likelihood of research starting points given information retrieval rates. Given this, what follows is a hypothetical exploration, rather than a statistical analysis, that provides a heuristic to consider the impact on academic libraries of alternate discovery services and decentralized, openly accessible scholarly content.

The Bayesian process is outlined by Phillips (1973). It allows for the ability to make an educated guess about a set of conditionals given two data points. It proceeds first by selecting two hypotheses:

H_1. Use academic library as research starting point.

H_2. Use *Google Scholar* as research starting point.

And adding notation for marking the outcome of either:

D_1. The data marking the retrieval of a full text document.

D_2. The data marking the nonretrieval of a full text document.

Assigning numbers to the prior probabilities, or prior beliefs or knowledge, was based on the 2009 Ithaka S + R faculty survey, which indicated that 38% of scientists use Google as a research starting point (Schonfeld & Housewright, 2010). From this statistic, the complement was inferred, which is that the academic library was used as a research starting point by the remaining 62% (again, a simplification). Thus, the probability of the first hypothesis is $p(H_1)$=0.62 (academic library starting point) and the probability of the second is $p(H_2)$=0.38 (*Google Scholar* starting point).

Likewise, using the data from this study, the probability that a scientist retrieved a full text document D1 after the likelihood of using *Google Scholar* as a research starting point is simply the product of the second hypothesis and the first outcome, $p(H_2) \times p(D_1)$, or 0.38 × 0.58 (see Table 7). Continuing, the probability that a scientist failed to retrieve a full text document D_2 after the likelihood of using *Google Scholar* as a research starting point is $p(H_2) \times p(D_2)$, or 0.38 × 0.42. It follows then that the probability of having retrieved a full text document given having used *Google Scholar*, $p(D_1|H_2)$, is about 0.22 or 22%. And the probability of failing to retrieve a full text document given having used *Google Scholar*, $p(D_2|H_2)$, is about 0.16 or 16%.

The same logic applies to assessing the use of the academic library. Suppose that an academic library can supply 97% of the articles in the *CiteULike* sample and can do so either through its collection on hand, from its collection in storage, from its subscribed content, or through interlibrary loan. In short, it can do so with any relevant means at its disposal. While this is a simplification of the sample space and does not consider other potential research starting points, it emphasizes the reality that using the academic library as a research starting point has a maximal upper bound. Thus, if the probability of having used the academic library as a research starting point p(H1) is 0.62 (the complement of having used *Google Scholar* as a research starting point), then the probability of retrieving a full text copy of one of the articles is $p(H_1)$=0.62 × $p(D_1)$=0.97, or about 60%. Likewise, the nonretrieval $p(D_2)$ resulting from the use of the academic library as a research starting point, $p(D_2|H_1)$, is 0.62 × 0.03 or about or 2%. Fig. 2 highlights these probabilities in a decision tree and shows that:

[INSERT FIGURE 2 HERE]

$p(D_1|H_1)$. The probability of retrieving a full text document given having used the academic library as a research starting point is 60%.

$p(D_2|H_1)$. The probability of not retrieving a full text document given having used the academic library as a research starting point is 2%.

$p(D_1|H_2)$. The probability of retrieving a full text document given having used *Google Scholar* as a research starting point is 22%.

$p(D_2|H_2)$. The probability of not retrieving a full text document given having used *Google Scholar* as a research starting point is 16%.

Expressed as propositions or in the form of the decision tree, the calculations show that it is more rational to use the academic library as a research starting point than it is to use Google Scholar. However, as the story about the knight and the fork in the road at the beginning of this study illustrated, if more researchers continue to use a service such as *Google Scholar* as a research starting point, then it must be concluded that the probable payoff for using *Google Scholar*, which is not null, must be worth more than the higher probable payoff that results from using the academic library.

The disintermediation question, though, uses the above calculations to ask the inverse of this conditional probability. It asks, what was the likelihood that a *CiteULike* user used an academic library or *Google Scholar* given having retrieved (or not retrieved) a relevant full text document. In essence, we ask:

$p(H_1|D_1)$. The probability that a *CiteULike* user used the academic library as a research starting point if she collected a full text document for her bibliographic reference.

$p(H_2|D_1)$. The probability that a *CiteULike* user used *Google Scholar* as a research starting point if she collected a full text document for her bibliographic reference. The above conditional probabilities complete Bayes' theorem, such that, where Bayes' Theorem is:

[INSERT EQUATION 1 HERE]

Then, the academic library as the Research Starting Point

[INSERT EQUATION 2 HERE]

And, *Google Scholar* as the Research Starting Point

[INSERT EQUATION 3 HERE]

Consequently, the following two conclusions are possible:

1. There is an 82% maximal probability that a *CiteULike* user used the academic library as a research starting point if he or she collected a full text document for a bibliographic

reference for an article.

2. There is an 18% minimal probability that a *CiteULike* user used *Google Scholar* as a research starting point if he or she collected a full text document for a bibliographic reference for an article.

CONCLUSION

This chapter applies a bibliometric analysis of a systematic random sample of data collected from *CiteULike* and augmented by data collected from *Google Scholar*. First, the chapter began with an overview of the entire sample and then proceeded to focus on the article document type. This was done to ensure measurement consistency, because the article document type is the most popular document type in the sample, and because open access issues largely pertain to journal articles. It was then shown that Google Scholar was able to provide full text access to a majority of the articles in the sample. While the proportion was not significantly different in the year 2010, it was by the year 2011 and more so by the year 2012. This was due to the increasing number of articles collected in the 2010 sample that became available as full text two years later. Although the sources providing full text access via *Google Scholar* are varied, when classified by type, the data shows that the dominant source providing full text access to journal articles is the university, which is largely composed of two sources: institutional and subject repositories.

The bibliometric analysis of the article type, by publication date, by post date, and by citation count show that the articles exhibit fairly typical characteristics with those in other bibliometric and citation counts. This weakly suggests that *CiteULike* users are not very different from researchers in general, an important consideration in inferring the composition of the *CiteULike* population. A surprising finding was that those articles with full text availability via *Google Scholar* exhibited a rather substantial citation advantage compared to those articles that were not full text accessible via *Google Scholar*. This supported the notion that citations might be a factor of full text availability.

To determine what factors influence full text availability, two logistic regressions were conducted based on a selection of predictor variables that might point to factors influencing full text availability. The models provided overall fits, and the predicted probabilities derived from the models suggest some influence on full text availability; however, statistically significant variables shifted between the two years. Although this warrants additional modeling, the results suggest that the main influence lies outside the variables tested.

Lastly, Bayes' theorem was used to build a hypothetical probability profile that would infer the likelihood of the academic library's use. This profile drew upon a statistic found in the Ithaka S + R 2009 faculty survey report (Schonfeld & Housewright, 2010) that showed that 38% of scientists report the use of Google as a research starting point. Adding that number with the data from this study, two inferences are drawn about the use of both *Google Scholar* and the academic library given the possibility of having retrieved a relevant full text document to an article reference in the sample. These inferences are:

1. There is an 82% maximal probability that a *CiteULike* user used the academic library as a research starting point if she collected a full text document for her article bibliographic reference.
2. There is an 18% minimal probability that a *CiteULike* user used *Google Scholar* as a

research starting point if she collected a full text document for her article bibliographic reference.

If we suppose that a *CiteULike* user is like any researcher (i.e., from comparable populations), then these claims may generalize to the broader scientific community, although further testing is needed before too many generalizations can be drawn.

Based on the analysis, this study suggests that what predicts full text availability is simply the number of sources providing full text access to articles. As these numbers increase, so does the number of accessible full text articles. Based on the classification of sources providing full text access to articles, in 2012 we know that universities (e.g., institutional or subject repositories, largely) provided 52.09% of the documents in the article sample (see Table 4). When this takes into consideration the Bayesian hypothetical assessment, not only is there an 82% maximal probability that a *CiteULike* user used the academic library as a research starting point if she collected a full text document for her article bibliographic reference, but over half of the articles she might have retrieved if she used *Google Scholar* as a research starting point came from the academy. This result has strategic implications for academic libraries, which will be discussed in the following chapter.

DISCUSSION AND CONCLUSION

Introduction

Early in this study, literature was cited highlighting researchers increasing use of alternate discovery services as research starting points in comparison to the services provided by academic librarians. Furthermore, since open access content is retrievable by these search engines, or other alternate discovery services, and since the amount of open access content is growing, then it is likely that many researchers can fulfill much of their informational needs by retrieving open access content with these tools. Similar reasoning has led to the claim that academic libraries will become marginalized by these information seeking practices.

This study applied decision theory and bounded rationality to frame this claim. This project showed that it is rational to begin with an alternate discovery service such as *Google Scholar* when it is possible to retrieve relevant scholarly documentation. Three years of bibliometric data based on a systematic random sample of bibliographic references collected by users on a social bookmarking web site were used to measure how many of the bibliographic references were found by *Google Scholar* and refer to freely available scholarly articles outside of a university's proxy. One key finding was that in 2012, nearly 58% of the bibliographic references to journal articles were freely available from 229 unique sources but that academic libraries provide over half of this content, possibly either through subject or institutional repositories. It was then shown that the number of academic libraries providing access to these journal articles have also increased over the threeyear time period under study. Given the success of these tools and the growing amount of material available as OA, researchers act rationally no matter which of the two broad choices they make to begin their research starting point.

The dominance of the university in providing full text access to material when researchers use *Google Scholar* as a research starting point is evidence that has strong impact on the strategic future of the academic library. Collectively, it implies that academic librarians' use

of institutional repositories to provide open access content appears to be serving them well. It also shows that academic libraries continue to work in the collection building business, as institutional repositories serve as one strategic response to an increasingly open access world. The payoff in this use of its collections is, furthermore, a mutual gain for both information seekers and academic libraries. The larger implication, though, comes from generalizing the strategic response that institutional repositories specifically serve. That is, access to collections should not be dependent on the popular information seeking practices of any specific population. Rather, they should be inherently flexible and be able to meet, without much or any intervention, whatever information seeking practices are in use.

Discussion

The two main research questions in this study explore the claim that academic libraries are being marginalized by the availability of alternative discovery services and by the increased decentralization of scholarly information. While the specific claim made by the Ithaka S + R report is one of the most recent of these claims, the claim itself is not new though the present state of affairs gives it renewed import.

The claim itself is based on the idea that one of the academic library's core functions is to collect scholarly information. The implicit argument is that if academic libraries have competitors in the collection "business," and if the use of their collections is being challenged by these competitors, then academic libraries risk marginalization. Accepting this definition of academic libraries and this argument as it stands, this study shows that even though the storage of scholarly information has become decentralized, academic library collections continue to be used to access scholarly information whatever the research starting point might be. We can therefore reject the argument about the marginalization of academic libraries.

It may make rational sense for a scientist or any researcher to use a nonlibrary electronic discovery service such as Google Scholar. If it takes less effort to use such a service, and if that service does its job well, then such activity can satisfice and is therefore rational under bounds. That rationality must be emphasized in any strategic interaction between librarians and their users or potential users. Still, librarians appear to be responding appropriately by providing open access content, either in the form of subject or institutional repositories, that can be retrieved through alternative services. While using a third party discovery service to retrieve open access or freely accessible content is a relevant alternative to the library's services, i.e., those that it pays for, librarians continue to insert their activities by providing content through open access archiving. The relevant alternative, that is, using *Google Scholar* or the like, thus appears quite challenging, but librarians seem to be, in aggregate, responding in a competitive fashion.

Librarians have at least three types of competitors. The first type includes those who provide alternate collections, the second type includes those who provide the discovery tools to search for and retrieve those collections, and the third type includes the information seekers. A simple heuristic supports these claims but can also be used to compose strategic plans. This heuristic can be framed as: given the actions taken by competitor A, what is the strategic response that maximizes the outcome and equilibrates the game and where the domain of A may include the three types of competitors listed above. If the actions and the agents are relevant to the mission and purpose of the responder, then the heuristic applies. When this heuristic is not

used, either by those who make claims about the importance of academic libraries' role in the scholarly communication system, problems arise. All too often, these claims are based on the idea that new technologies, new players, and new practices will by their existence threaten library use.

These claims are simplistic when they do not take into consideration the relevant alternatives or the conditional likelihoods of choosing these alternatives. In this context, it is not appropriate to value something in and of itself. It is only appropriate to valuate something in comparison to a similar thing and to do so iteratively. Specifically, measuring the value of an academic library must take into consideration measuring the value of comparable entities who provide similar services and tools and whose services and tools are used for similar tasks. While the Ithaka S + R study concluded that the presence and use of alternate tools can undermine the role of the library as an intermediary in the research process, the suggestion offered from this study is that the information discovery process as it relates to research is simply growing more complicated and interconnected as new alternatives become available.

Academic librarians do face challenges. If discoverability and access to their collections are dependent on the use of specific applications, then academic librarians cannot succeed in responding strategically to the popular information seeking practices of the day. As stated in the beginning of this study, such a scenario is not fully capable of taking into consideration the decision matrix of the information seeker. Currently, for instance, online public access catalogs (OPAC) maintain bibliographic records in the deep web making the content discoverable only through their search applications. Consequently, there is generally only one main path to identify that item in the collection, and that one main path is dependent on the use of a specific tool. Limiting access in this way is a poor strategic response to today's most prevalent information practices. If libraries persist in this vein, they may forfeit their role as intermediaries in the information use process, if not also the search process. Current efforts to grow the Digital Public Library of America (DPLA) may resolve this issue by using a platform that allows libraries to coordinate pathways to collections without committing to any one search tool (see Peek, 2012 for a description of the DPLA), but more needs to be accomplished.

Despite the fact that academic librarians are responding competitively as more varied tools for retrieving scholarly documentation emerge and as they become available in varied locations, academic librarians may still face a competitive disadvantage if researchers do not recognize that the materials they collect, read, and use come from academic libraries. That is, academic librarians may suffer from researchers' skewed impression that emerging vehicles for both searching and retrieving information may be superior to the ones that librarians provide, or that open access articles retrieved by these services are not originating from institutional repositories, as they often seem to do, per the data in this study.

Furthermore, while the open access movement offers numerous advantages for many scholarly stakeholders, it also represents an existential shift for academic libraries and for the role and profession of librarianship. It is now impossible for academic librarians to exercise "completeness and control" (Smith, 1990, p. 9) of the scholarly record, and this state of the affairs has significant implications for the library and the profession.

However the future of collection development and management works in practice, academic libraries are, in fact, defined as much by the librarians who work in them as by the

collections they build. The expertise of the people who work in the library make it more than a warehouse of content. As Plutchak (2012) argues, the future of libraries is librarians, as it has long been, and it is good to recognize that. To prosper, librarians need to recognize and effectively respond to changes in scholarly communication with programs and policies that match opportunities to the needs of users. Indeed, as Lingel (2012) writes, the " ... Library reflects the values of its community through its policies, not through its collections" (Policies are politics section, para. 1), and Hill (2009) notes that "Policies guide the organization and the responsibility to create them confers a great amount of power to the creator" (p. 87). These policies, it is important to observe, and within the context of this study, are a reflection of the intent of the librarians who write them, increasing the importance of their response to the environment in which they work and live.

CONCLUSION

Bibliometrics and information seeking studies both aim to understand information behavior using two different approaches. The former furthers our understanding about general patterns of behavior while the latter offers methods for gaining deeper understanding of the various personal dimensions of the seeking and gathering processes. Using one to build on the other is a complimentary process. Additionally, the availability of personal collections of reading material offers an attractive means for inquiring into both the scholarly communication system and the information seeking and gathering behavior of researchers. However, this study has focused less on overall behavior and concentrated more on the inherent decisions and implications of information seekers and their strategic outcomes.

The theory and material used in this study provided a guide to understand the rich source of data—how context influences, constrains, and binds such behavior. This material offered important insights into the decisions users make when searching for and saving scholarly content. Lastly, the study sought to identify theories and develop a strategy for understanding the impact that various alternatives have on academic libraries, something that has either has largely been ignored or, when it has been addressed, has been studied based on incomplete premises that led to incomplete conclusions. Future inquiry into the future of academic libraries should always take into consideration the entirety of the system and not focus on the isolated actions of any set of people or any single type of service.

REFERENCES

Abbott, A. (2004). *Methods of discovery: Heuristics for the social sciences. Contemporary societies*. New York, NY: W. W. Norton and Company.

ACRL. (2000). *Information literacy competency standards for higher education*. Retrieved from http://www.ala.org/acrl/standards/informationliteracycompetency

Adkins, D., & Bala, E. (2004). Public library outreach as a function of staffing and metropolitan location. *Library and Information Science Research, 26*, 338-350. doi:10.1016/j.lisr.2004.01.001

Aguillo, I. F. (2012). Is Google Scholar useful for bibliometrics? A webometric analysis. *Scientometrics, 91*, 343-351. doi:10.1007/s11192-011-0582-8

Akeroyd, J. (2001). The future of academic libraries. *Aslib Proceedings, 53*(3), 79-84. doi:10.1108/EUM0000000007041

Albert, K. M. (2006). Open access: Implications for scholarly publishing and medical libraries. *Journal of the Medical Library Association, 94*(3), 253-262. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1525322/

Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratio of IRs in Google Scholar. *Library Hi Tech, 30*(1), 60-81. doi:10.1108/07378831211213210

Bailey, C. W. (2007). Open access and libraries. *Collection Management, 32*, 351-383. doi:10.1300/J105v32n03_07

Baldwin, V. A. (2009). Using Google Scholar to search for online availability of a cited article in engineering disciplines. *Issues in Science and Technology Librarianship, 56*. doi:10.5062/F4WM1BBC

Bergstrom, C. T., & Bergstrom, T. C. (2006). The economics of ecology journals. *Frontiers in Ecology and the Environment, 4*(9), 488-495. doi:10.1890/1540-9295(2006)4[488:TEOEJ]2.0.CO;2

Binmore, K. (1994). *Game theory and the social contract: Playing fair* (Vol. 1). Cambridge, MA: MIT Press.

Binmore, K. (2007). *Game theory: A very short introduction*. Oxford: Oxford University Press.

Black, A. (2007). Mechanization in libraries and information retrieval: Punched cards and microfilm before the widespread adoption of computer technology in libraries. *Library History, 23*, 291-299. doi:10.1179/174581607x254785

Bogers, T., & van den Bosch, A. (2008). Recommending scientific articles using CiteULike. In *Proceedings of the 2008 ACM conference on recommender systems* (pp. 287-290). New York, NY: ACM.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology, 36*, 2-72. doi:10.1002/aris.1440360102

Bornmann, L., & Hans-Dieter, D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45-80. doi:10.1108/00220410810844150

Borrego, A., & Fry, J. (2012). Measuring researchers' use of scholarly information through social bookmarking data: A case study of BibSonomy. *Journal of Information Science, 38*(3), 297-308. doi:10.1177/0165551512438353

Bosch, S., & Henderson, K. (2012, April 30). Coping with the terrible twins: Periodicals price survey 2012. *Library Journal*. Retrieved from http://lj.libraryjournal.com/2012/04/funding/coping-with-the-terrible-twins-periodicals-price-survey-2012/

Boslaugh, S. (2012). *Statistics in a nutshell* (2nd ed.). Beijing: O'Reilly.

Brandstätter, E., & Brandstätter, H. (1996). What's money worth? Determinants of the subjective value of money. *Journal of Economic Psychology, 17*(4), 443-464. doi:10.1016/0167-4870(96)00019-0

Broadus, R. N. (1987). Toward a definition of "bibliometrics". *Scientometrics, 12*(5-6), 373-379. doi:10.1007/BF02016680

Brookes, B. C. (1969). Bradford's law and the bibliography of science. *Nature, 224*(5223), 953-956. doi:10.1038/224953a0

Budd, J. M. (1992). Bibliometrics: A method for the study of the literature of higher education. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8, pp. 345-378). New York, NY: Agathon Press.

Budd, J. M. (2002). Serials prices and subscriptions in the social sciences. *Journal of Scholarly Publishing, 33*(2), 90-101. doi:10.3138/jsp.33.2.90

Budd, J. M. (2009). Academic library data from the United States: An examination of trends. *Libres: Library and Information Science Research Electronic Journal, 19*(2), 1-21.

Budd, J. M. (2012). Scholarly communication's mess: Can economic analysis help? *Libres: Library and Information Science Research Electronic Journal, 22*(1), 1-17.

Burns, C. S. (2013). *Free or open access to scholarly documentation: Google Scholar or academic libraries*. Doctoral dissertation. Retrieved from MOspace, BurnsS-051413-D1323. Retrieved from http://hdl.handle.net/10355/37582

Bush, V. (1945, July). As we may think. *Atlantic Monthly*. Retrieved from http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

Capocci, A., & Caldarelli, G. (2008). Folksonomies and clustering in the collaborative system CiteULike. *Journal of Physics A: Mathematical and Theoretical, 41*. doi:10.1088/17518113/41/22/224016

Carpenter, K. E. (1996). A library historian looks at librarianship. *Daedalus, 125*(4), 77-102.

Carrigan, D. P. (1995). Toward a theory of collection development. *Library Acquisitions: Practice & Theory, 19*(1), 97-106. doi:10.1016/0364-6408(94)00056-2

Cave, E. M. (2005). A normative interpretation of expected utility theory. *The Journal of Value Inquiry, 39*(3), 431-441. doi:10.1007/s10790-006-7525-2

Chan, L. (2004). Supporting and enhancing scholarship in the digital age: The role of openaccess institutional repositories. *Canadian Journal of Communication, 29*(3). Retrieved from http://www.cjc-online.ca/index.php/journal/article/view/1455/1579

Chen, X. (2010). Google Scholar's dramatic coverage improvement five years after debut. *Serials Review, 36*(4), 221-226. doi:10.1016/j.serrev.2010.08.002

CiteULike. (2010a). Available datasets [Accessing CiteULike datasets]. Retrieved from http://www.CiteULike.org/faq/data.adp

CiteULike. (2010b). How to post a paper to CiteULike [Instructions for browser bookmarklet]. Retrieved from http://www.CiteULike.org/post

Collins, C. S., & Walters, W. H. (2010). Open access journals in college library collections. *The Serials Librarian, 59*(2), 194-214. doi:10.1080/03615261003623187

Corrado, E. M. (2005). The importance of open access, open source, and open standards for libraries. *Issues in Science and Technology, 42*(Spring). Retrieved from http://istl.org/05spring/article2.html

Cothran, T. (2011). Google Scholar acceptance and use among graduate students: A quantitative study. *Library & Information Science Research, 33*(4), 293-301. doi:10.1016/j.lisr.2011.02.001

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.

Cronin, B. (2001). Bibliometrics and beyond: Some thoughts on web-based citation analysis. *Journal of Information Science, 27*, 1-7. doi:10.1177/016555150102700101

Cronin, B., & Franks, S. (2006). Trading cultures: Resource mobilization and service rendering in the life sciences as revealed in the journal article's paratext. *Journal of the American Society for Information Science and Technology, 57*(14), 1909-1918. doi:10.1002/asi.20407

Cronin, B., Shaw, D., & La Barre, K. L. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology, 54*(9), 855-871. doi:10.1002/asi.10278

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

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Lanham, MD: Scarecrow Press.

De Bruin, B. (2005). Game theory in philosophy. *Topoi, 24*(2), 197-208. doi:10.1007/s11245005-5055-3

Dixit, A. K., & Skeath, S. (2004). *Games of strategy*. New York, NY: W. W. Norton & Company.

Dourish, P. (2001). *Where the action is: The foundations of embodied interaction*. Cambridge, MA: MIT Press.

Dowd, S. T. (1990). Library cooperation: Methods, models to aid information access. *Journal of Library Administration, 12*(3), 63-81.

Drott, M. C. (2006). Open access. *Annual Review of Information Science and Technology, 40*, 79-109. doi:10.1002/aris.1440400110

Dufwenberg, M. (2010). Psychological games. In S. N. Durlauf & L. E. Blume (Eds.), *Game theory* (pp. 272-278). New York, NY: Palgrave MacMillan.

Ellis, D., Cox, D., & Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentation, 49*(4), 356-369.

Emamy, K., & Cameron, R. (2007). CiteULike: A researcher's social bookmarking service.

Ariadne, 51. Retrieved from http://www.ariadne.ac.uk/issue51/emamy-cameron

Enger, K. B. (2009). Using citation analysis to develop core book collections in academic libraries. *Library & Information Science Research, 31*(2), 107-112. doi:10.1016/j.lisr.2008.12.003

Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology, 4*(5), e157. doi:10.1371/journal.pbio.0040157

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal, 22*(2), 338-342. doi:10.1096/fj.07-9492LSF

Farber, S. (1998). Undesirable facilities and property values: A summary of empirical studies. *Ecological Economics, 24*(1), 1-14. doi:10.1016/S0921-8009(97)00038-4

Fidczuk, R., Beebe, L., & Wallas, P. (2007). Today's journal cost: Print vs. online. *Serials Librarian, 52*(3-4), 341-348. doi:10.1300/J123v52n03_15

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles, CA: Sage.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(3159), 108-111. doi:10.1126/science.122.3159.108

Gargouri, Y., Hajjem, C., Lariviere, V., Gingras, Y., Carr, 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

Google. (n.d.). Library support. Retrieved from http://scholar.google.com/intl/en-US/scholar/libraries.html

Greco, A. N., Wharton, R. M., Estelami, H., & Jones, R. F. (2006). The state of scholarly publishing: 1981-2000. *Journal of Scholarly Publishing, 37*(3), 155-214.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software, 40*(3), 1-25. Retrieved from http://www.jstatsoft.org/v40/i03/

Grüne-Yanoff, T., & Schweinzer, P. (2008). The roles of stories in applying game theory. *Journal of Economic Methodology, 15*(2), 131-146. doi:10.1080/13501780802115075

Hamlin, A. T. (1981). *The university library in the United States*. Philadelphia, PA: University of Pennsylvania Press.

Harloe, B., & Budd, J. M. (1994). Collection development and scholarly communication in the era of electronic access. *The Journal of Academic Librarianship, 20*(2), 83-87. doi:10.1016/0099-1333(94)90043-4

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., ... Hilf, E. R. (2008). The access/impact problem and the green and gold roads to open access: An update. *Serials Review, 34*(1), 36-40. Retrieved from http://dx.doi.org/10.1016/j.serrev.2007.12. 005

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer.

Harzing, A. W. K., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics, 8*, 61-73. doi:10.3354/esep00076

Hausman, D. M. (2005). Sympathy, commitment, and preference. *Economics and philosophy, 21*(1), 33À50. doi:10.1017/S0266267104000379

Haustein, S., & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics, 5*(3), 446-457. doi:10.1016/j.joi.2011.04.002

Herrera, G. (2011). Google Scholar users and user behaviors: An exploratory study. *College & Research Libraries, 72*(4), 316-330.

Hill, H. (2009). *Outsourcing the public library: A critical discourse analysis*. Unpublished doctoral dissertation, University of Missouri. Retrieved from http://hdl.handle.net/10355/6126

Howland, J. L., Wright, T. C., Boughan, R. A., & Roberts, B. C. (2009). How scholarly is Google Scholar? Comparison to library databases. *College & Research Libraries, 70*(3), 227-234.

Hull, D., Pettifer, S. R., & Kell, D. B. (2008). Defrosting the digital library: Bibliographic tools for the next generation web. *PLoS Computational Biology, 4*(10), e1000204. doi:10.1371/journal.pcbi.1000204

Julien, H., & Genuis, S. K. (2011). Librarians' experience of the teaching role: A national survey of librarians. *Library and Information Science Research, 33*(2), 103-111. doi:10.1016/j.lisr.2010.09.005

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-292. doi:10.2307/1914185

Kilgour, F. G. (1939). A new punched card for circulation records. *Library Journal, 64*(4), 131-133.

King, J. E. (2008). Binary logistic regression. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 358-384). Thousand Oaks, CA: Sage Publications.

Kipp, M. E. I. (2011). User, author and professional indexing in context: An exploration of tagging practices on CiteULike. *Canadian Journal of Information and Library Science, 35*(1), 17-48. doi:10.1353/ils.2011.0008

Kirkwood, H. P., & Kirkwood, M. C. (2011). Historical research: Historical abstracts with full text or Google Scholar. *Online: Exploring Technology & Resources for Informational Professionals, 35*(4), 28-32.

Kling, R., & Callahan, E. (2003). Electronic journals, the Internet, and scholarly communication. *Annual Review of Information Science and Technology, 37*, 127-177. doi:10.1002/aris.1440370105

Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology, 58*(7), 1055-1065. doi:10.1002/asi.20584

Kress, N., Bosque, D. D., & Ipri, T. (2011). User failure to find known library items. *New*

Library World, 112(3À4), 150-170. doi:10.1108/03074801111117050

Kyrillidou, M., Morris, S., & Roebuck, G. (2012). ARL statistics: 2010–2011. Washington, DC: Association of Research Libraries. Retrieved from http://www.arl.org/

Kyrillidou, M. C., & Young, M. C. (2006). *ARL statistics, 2004–05*: A compilation of statistics from the one hundred and twenty-three members of the association of research libraries. Washington, DC: Association of Research Libraries. Retrieved from http://www.arl. org/

Lancaster, F. W. (1978). *Toward paperless information systems*. New York, NY: Academic Press.

Lewis, D. W. (2012). The inevitability of open access. *College and Research Libraries, 73*(5), 493-506.

Licklider, J. C. R. (1965). *The library of the future*. Cambridge, MA: MIT Press.

Lingel, J. (2012). Occupy Wall Street and the myth of technological death of the library. *First Monday, 17*(8). doi:10.5210/fm.v17i8.3845

McCain, K. W., & Bobick, J. E. (1981). Patterns of journal use in a departmental library: A citation analysis. *Journal of the American Society for Information Science, 32*(4), 257-267.

McGuigan, G. S., & Russell, R. D. (2008). The business of academic publishing: A strategic analysis of the academic journal publishing industry and its impact on the future of scholarly publishing. *Electronic Journal of Academic and Special Librarianship, 9*(3). Retrieved from http://southernlibrarianship.icaap.org/content/v09n03/mcguigan_g01.html

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58*(13), 2105-2125. doi:10.1002/asi.20677

Michalak, S. C. (2012). This changes everything: Transforming the academic library. *Journal of Library Administration, 52*(5), 411-423. doi:10.1080/01930826.2012.700801

Mitchell, B. A. (2007). Boston Library catalogues, 1850À1875. In T. Augst & K. Carpenter (Eds.), *Institutions of reading: The social life of libraries in the United States* (pp. 119-147). Amherst, MA: University of Massachusetts Press.

Moran, B. B. (2001). Restructuring the university library: A North American perspective. *Journal of Documentation, 57*(1), 100-114. doi:10.1108/EUM0000000007079

Mullen, L. B., & Hartman, K. A. (2006). Google Scholar and the library web site: The early response by ARL libraries. *College and Research Libraries, 67*(2), 106-122.

Narin, F., & Moll, J. K. (1977). Bibliometrics. *Annual Review of Information Science and Technology, 12*, 35-58.

Neuhaus, C., Neuhaus, E., & Asher, A. (2008). Google Scholar goes to school: The presence of Google Scholar and university web sites. *The Journal of Academic Librarianship, 34*(1), 39-51. doi:10.1016/j.acalib.2007.11.009

Nickel, P. J. (2009). Trust, staking, and expectations. *Journal for the Theory of Social Behaviour, 39*(3), 345-362. doi:10.1111/j.1468-5914.2009.00407.x

Nicolaisen, J. (2003). The social act of citing: Towards new horizons in citation theory. *Proceedings of the American Society for Information Science and Technology, 40*(1), 12-20. doi:10.1002/meet.1450400102

Niu, X., & Hemminger, B. M. (2012). A study of factors that affect the information-seeking behavior of academic scientists. J*ournal of the American Society for Information Science and Technology, 63*(2), 336-353. doi:10.1002/asi.21669

Niu, X., Hemminger, B. M., Lown, C., Adams, S., Brown, C., Level, A., … Cataldo, T. (2010). National study of information seeking behavior of academic research in the United States. *Journal of the American Society for Information Science and Technology, 61*(5), 869-890. doi:10.1002/asi.21307

Norris, M., Oppenheim, C., & Rowland, F. (2008). Finding open access articles using Google, Google Scholar, OAIster and OpenDOAR. *Online Information Review, 32*(6), 709-715. doi:10.1108/14684520810923881

Noruzi, A. (2005). Google Scholar: The new generation of citation indexes. *Libri, 55*, 170-180. doi:10.1515/LIBR.2005.170

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science, 289*, 1773-1775. doi:10.1126/science.289.5485.1773

Nyquist, C. (2010). An academic librarian's response to the "ITHAKA faculty survey 2009: Key strategic insights for libraries, publishers, and societies". *Journal of Interlibrary Loan, Document Delivery & Electronic Reserve, 20*(4), 275-280. doi:10.1080/ 1072303X.2010.508419

Oberg, L. R., Mentges, M. E., McDermott, P. N., & Harusadangkul, V. (1992). The role, status, and working conditions of paraprofessionals: A national survey of academic libraries. *College and Research Libraries, 53*(3), 215-238.

O'Reilly, T. (2005). What is web 2.0: Design patterns and business models for the next generation of software. Retrieved from http://oreilly.com/pub/a/web2/archive/what-is-web-20. html

Parker, R. H. (1936). The punched card method in circulation work. *The Library Journal, 61*, 903À905.

Peek, R. (2012). *Digital public library of America*. Information Today, 29(2), 24.

Phillips, L. D. (1973) *Bayesian statistics for social scientists*. London: Nelson.

Plutchak, T. S. (2012). Breaking the barriers of time and space: The dawning of the great age of librarians. *Journal of the Medical Library Association, 100*(1), 10-19.doi:10.3163/1536-5050.100.1.004

Pomerantz, J., & Marchionini, G. (2007). The digital library as place. *Journal of Documentation, 63*(4), 505-533. doi:10.1108/00220410710758995

Price, D. J. D. S. (1986). *Little science, big science ... and beyond*. New York, NY: Columbia University Press.

Priem, J., & Hemminger, B. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social web. *First Monday, 15*(7). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/2874/2570

R Core Team (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica, 68*(5), 1281-1292. doi:10.1111/1468-0262.00158

Ritzberger, K. (2002). *Foundations of non-cooperative game theory*. Oxford: Oxford University Press.

Romero, L. (2008). Confirming suspicions: An analysis of original communication studies journal price data. *Collection Management, 33*(3), 189-218. doi:10.1080/01462670802045525

Ross, C. S. (2009). Reader on top: Public libraries, pleasure reading, and models of reading. *Library Trends, 57*(4), 632-656. doi:10.1353/lib.0.0059

Sapp, G., & Gilmour, R. (2002). A brief history of the future of academic libraries: Predictions and speculations from the literature of the profession, 1975 to 2000—Part one, 1975 to 989. *Portal: Libraries and the Academy, 2*(4), 553-576. doi:10.1353/pla.2002.0086

Sapp, G., & Gilmour, R. (2003). A brief history of the future of academic libraries: Predictions and speculations from the literature of the profession, 1975 to 2000—Part two, 1990 to 000. *Portal: Libraries and the Academy, 3*(1), 13-34. doi:10.1353/pla.2003.0008

Savolainen, R. (2012). Expectancy-value beliefs and information needs as motivators for taskbased information seeking. *Journal of Documentation, 68*(4), 492-511. doi:10.1108/00220411211239075

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica, 57*(3), 571-587.

Schonfeld, R. C., & Housewright, R. (2010). Faculty survey 2009: Key strategic insights for libraries, publishers, and societies. Retrieved from www.sr.ithaka.org/researchpublications/faculty-survey-2009

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization*. New York, NY: Doubleday.

Sennyey, P., Ross, L., & Mills, C. (2009). Exploring the future of academic libraries. *The Journal of Academic Librarianship, 35*(3), 252-259. doi:10.1016/j.acalib.2009.03.003

Shiflett, O. L. (1981). *Origins of American academic librarianship*. Norwood, NJ: Ablex Publishing Corporation.

Simon, H. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*(1), 99-118.

Simon, H. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1-19.

Sin, S.-C. J., & Kim, K.-S. (2008). Use and non-use of public libraries in the information age: A

logistic regression analysis of household characteristics and library services variables. *Library and Information Science Research, 30*, 207-215. doi:10.1016/j.lisr.2007.11.008

Smith, E. (1990). *The librarian, the scholar, and the future of the research library*. New York, NY: Greenwood Press.

Smith, L. C. (1981). Citation analysis. *Library Trends, 30*, 83-106.

tbogers (2009). Science papers that interest you. Retrieved from http://blog.CiteULike.org/?p=11. Accessed on December 1.

Tenopir, C. (2012). Beyond usage: Measuring library outcomes and value. *Library Management, 33*(1-2), 5-13. doi:10.1108/01435121211203275

Tenopir, C., King, D. W., Spencer, J., & Wu, L. (2009). Variations in article seeking and reading patterns of academics: What makes a difference? *Library & Information Science Research, 31*(3), 139-148. doi:10.1016/j.lisr.2009.02.002

Theng, Y.-L., & Sin, S.-C. J. (2012). Analysing the effects of individual characteristics and self-efficacy on users' preferences for system features in relevance judgment. *Information Research, 17*(4). Retrieved from http://informationr.net/ir/17-4/paper536.html#.UQr4qVn1SJg

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. doi:10.1126/science.185.4157.1124

Vaughan, L. (2001). *Statistical methods for the information professional: A practical, painless approach to understanding, using, and interpreting statistics*. Medford, NJ: Information Today.

Vaughan, L., & Shaw, D. (2008). A new look at evidence of scholarly citation in citation indexes and from web sources. *Scientometrics, 74*(2), 317-330. doi:10.1007/s11192-0080220-2

Walters, W. H. (2009). Google Scholar search performance: Comparative recall and precision. *Portal: Libraries and the Academy, 9*(1), 5-24. doi:10.1353/pla.0.0034

Walters, W. H., & Linvill, A. C. (2011a). Bibliographic index coverage of open-access journals in six subject areas. *Journal of the American Society for Information Science and Technology, 62*(8), 1614-1628. doi:10.1002/asi.21569

Walters, W. H., & Linvill, A. C. (2011b). Characteristics of open access journals in six subject areas. *College and Research Libraries, 72*(4), 372-392.

White, H. D., & McCain, K. W. (1989). Bibliometrics. *Annual Review of Information Science and Technology, 24*, 119-186.

Wickham, H. (2007). Reshaping data with the reshape package. J*ournal of Statistical Software, 21*(12), 1-20. Retrieved from http://www.jstatsoft.org/v21/i12/. Accessed on February 3, 2013.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.

Wiegand, W. A. (1990). Research libraries, the ideology of reading, and scholarly communication, 1876-1900. In P. Dain & J. Y. Cole (Eds.), *Libraries and scholarly*

communication in the United States: The historical dimension (pp. 71-87). New York, NY: Greenwood Press.

Wolfram, D. (2003). *Applied informetrics for information retrieval research*. Westport, CT: Libraries Unlimited.

Wouters, P. (1998). The signs of science. *Scientometrics, 41*, 225-241. doi:10.1007/ BF02457980

Yadamsuren, B., Paul, A., Wang, J., Wang, X., & Erdelez, S. (2008). Web ecology: Information needs of different user groups in the context of a community college website. *Proceedings of the American Society for Information Science and Technology, 45*(1), 1-4. doi:10.1002/meet.2008.14504503112

Yang, K., & Meho, L. I. (2006). Citation analysis: A comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology, 43*, 1-15. doi:10.1002/meet.14504301185

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley Press.

Table 1. *CiteULike* Sample Composition.

Document Type	**Count**	**Percentage (%)**
Article	691	69.45
Book	89	8.94
In proceeding articles	89	8.94
Misc	39	3.92
Electronic	18	1.81
Proceedings	17	1.71
In collection (e.g., standalone book chapter)	15	1.51
Tech report	15	1.51
PhD thesis	9	0.90
In book (e.g., book chapter)	6	0.60
Unpublished	3	0.30
Master's thesis	2	0.20
Booklet	1	0.10
Manual	1	0.10
Total	995	99.99

Table 2. Top 10 Full Text Article Sources: 2010-2012.

Full Text Source	**2010**	**2011**	**2012**	**OA Type**
CiteSeerX	40	38	5	Green
NIH	35	42	40	Gold
arXiv	27	28	26	Green
Oxford Journals	12	13	12	Gold
PNAS	11	11	11	Gold
BioMed Central	7	10	11	Gold
PLoS	5	4	5	Gold
Harvard University	5	5	5	Green
Rockefeller University	4	-	4	Green
American Meteorological Society	-	-	4	Gold

Table 3. Full Text Sources by Type with Count and Percentage of Unique Source Types.

Type	**2010**	**2011**	**2012**
Activist organizations	-	-	1 (0.44%)
Business	7 (3.98%)	8 (4.21%)	10 (4.37%)
Government	4 (2.27%)	4 (2.11%)	4 (1.75%)
National	3 (1.70%)	3 (1.58%)	5 (2.18%)
Other organization	1 (0.57%)	-	1 (0.44%)
Personal files	5 (2.84%)	7 (3.68%)	9 (3.93%)
Publisher files	40 (22.73%)	40 (21.05%)	46 (20.09%)
Universities	100 (56.82%)	117 (61.58)	145 (63.32%)
Other	16 (9.09%)	11 (5.79%)	8 (3.49%)
Total	176 (100%)	190 (100%)	229 (100%)

Table 4. Full Text Source by Type with Count and Percentage of Number of Articles Provided by Type.

Type	**2010**	**2011**	**2012**
Activism	–	–	1 (0.26%)
Business	7 (2.03%)	8 (2.20%)	11 (2.88%)
Government	39 (11.30%)	46 (12.64%)	46 (12.04%)
National	5 (1.45%)	5 (1.37%)	6 (1.57%)
Other organization	1 (0.29%)	–	1 (0.26%)
Personal files	5 (1.45%)	7 (1.92%)	9 (2.36%)
Publisher files	88 (25.51%)	87 (23.90%)	100 (26.18%)
Universities	183 (53.04%)	200 (54.95%)	199 (52.09%)
Other	17 (4.93%)	11 (3.02%)	9 (2.36%)
Total	345 (100%)	364 (100%)	382 (100%)

Table 5. Distribution of Articles and Citations, Ordered by Cumulative Percentage of Articles.

	Cumulative Percentage of Articles (%)	**Cumulative Sum of Articles**	**Citation Count**	**Cumulative Percentage of Citations (%)**
2010	25.00	162	4	0.02
	50.46	327	23	0.47
	75.62	490	72	4.27
	100.00	648	6156	100.00
2011	25.19	167	7	0.04
	50.38	334	28	0.58
	75.26	499	83	4.70
	100.00	663	7062	100.00
2012	25.04	166	11	0.07
	50.23	333	37	0.79
	75.41	500	102	5.04
	100.00	663	8374	100.00

Table 6. Distribution of Articles and Citations, Ordered by Cumulative Percentage of Citations.

	Cumulative Percentage of Articles (%)	**Cumulative Sum of Articles**	**Citation Count**	**Cumulative Percentage of Citations (%)**
2010	25.11	597	285	92.13
	20.24	628	736	96.91
	76.38	644	1591	99.38
	100.00	648	6156	100.00
2011	25.44	612	348	92.31
	51.08	643	838	96.98
	75.22	658	1702	99.25
	100.00	663	7062	100.00
2012	25.05	605	372	91.25
	50.85	642	937	96.83
	75.51	658	2145	99.25
	100.00	663	8374	100.00

Table 7. Article Count with Google Full Text Access: 2010-2012.

	Full Text	Count	Estimate	χ^2	*df*	*p*	95% CI
2010 (n = 648)	No	303	46.76%	2.5941	1	0.1073	[42.87%, 50.69%]
	Yes	345	53.24%	2.5941	1	0.1073	[49.31%, 57.13%]
2011 (n = 663)	No	299	45.10%	6.178	1	0.0129	[41.28%, 48.98%]
	Yes	264	54.90%	6.178	1	0.0129	[51.02%, 58.72%]
2012 (n = 662)	No	281	42.45%	14.8051	1	0.0001	[38.66%, 46.32%]
	Yes	381	57.55%	14.8051	1	0.0001	[53.68%, 61.34%]

Table 8. Article Citation Counts by Full Text Access: 2010-2012.

Year	Full Text	n	Median	Min	Max
2010	No	303	12	0	1662
	Yes	345	32	0	6156
2011	No	299	15	0	1833
	Yes	364	37	0	7062
2012	No	281	20	0	2048
	Yes	381	49	0	8374

Table 9. Logistic Regression on Full Text Dichotomous Variable: 2011 Article Full Text Access with Exponentiated Coefficients and Confidence Intervals.

					95% CI for Odds Ratio		
Variable	*B*	*SE*	*Wald t*	*p*	Lower	Odds Ratio	Upper
Authors	0.0887	0.0318	2.795	0.0052	1.0301	1.0928	1.1665
Pub year	0.0425	0.0101	4.201	0.0000	1.0238	1.0434	1.0653
Post year	−0.1023	0.0646	−1.582	0.1136	0.7946	0.9028	1.0241
Citations 2010	0.0015	0.0005	2.984	0.0028	1.0006	1.0015	1.0025

Note: *B* = parameter estimate; *SE* = standard error of the parameter estimated; *CI* = confidence interval.

Table 10. Logistic Regression on Full Text Dichotomous Variable: 2012 Article Full Text Access with Exponentiated Coefficients and Confidence Intervals.

					95% CI for Odds Ratio		
Variable	*B*	*SE*	*Wald t*	*p*	Lower	Odds Ratio	Upper
Authors	0.0100	0.0198	0.503	0.6148	0.9761	1.0100	1.0583
Pub year	0.0473	0.0098	4.828	0.0000	1.0294	1.0484	1.0697
Post year	−0.0911	0.0644	−1.415	0.1571	0.8038	0.9129	1.0350
Citations 2011	0.0016	0.0006	3.345	0.0008	1.0007	1.0016	1.0025

Note: *B* = parameter estimate; *SE* = standard error of the parameter estimated; *CI* = confidence interval.

Table 11. Summary of Predicted Probabilities of Full Text Access for 2011 and 2012 Logistic Regression Models

Range	**2011 Model**	**2012 Model**
First quartile	49.59%	54.06%
Median	56.27%	59.84%
Third quartile	60.82%	63.34%

Equation 1:

$$p(H_1|D_1)=\frac{p(H_1)\times p(D_1|H_1)}{p(H_2)\times p(D_1|H_2)+p(H_1)\times p(D_1|H_1)}$$

Equation 2:

$$p(H_1|D_1)=\frac{(0.62\times 0.6014)}{(0.38\times 0.2204)+(0.62\times 0.6014)}=0.8165=82\%$$

Equation 3:

$$p(H_1|D_1)=\frac{(0.38\times 0.2204)}{(0.38\times 0.2204)+(0.62\times 0.6014)}=0.1834=18\%$$

Fig. 1. 2012 Full Text Article Access Trends by Publication Year

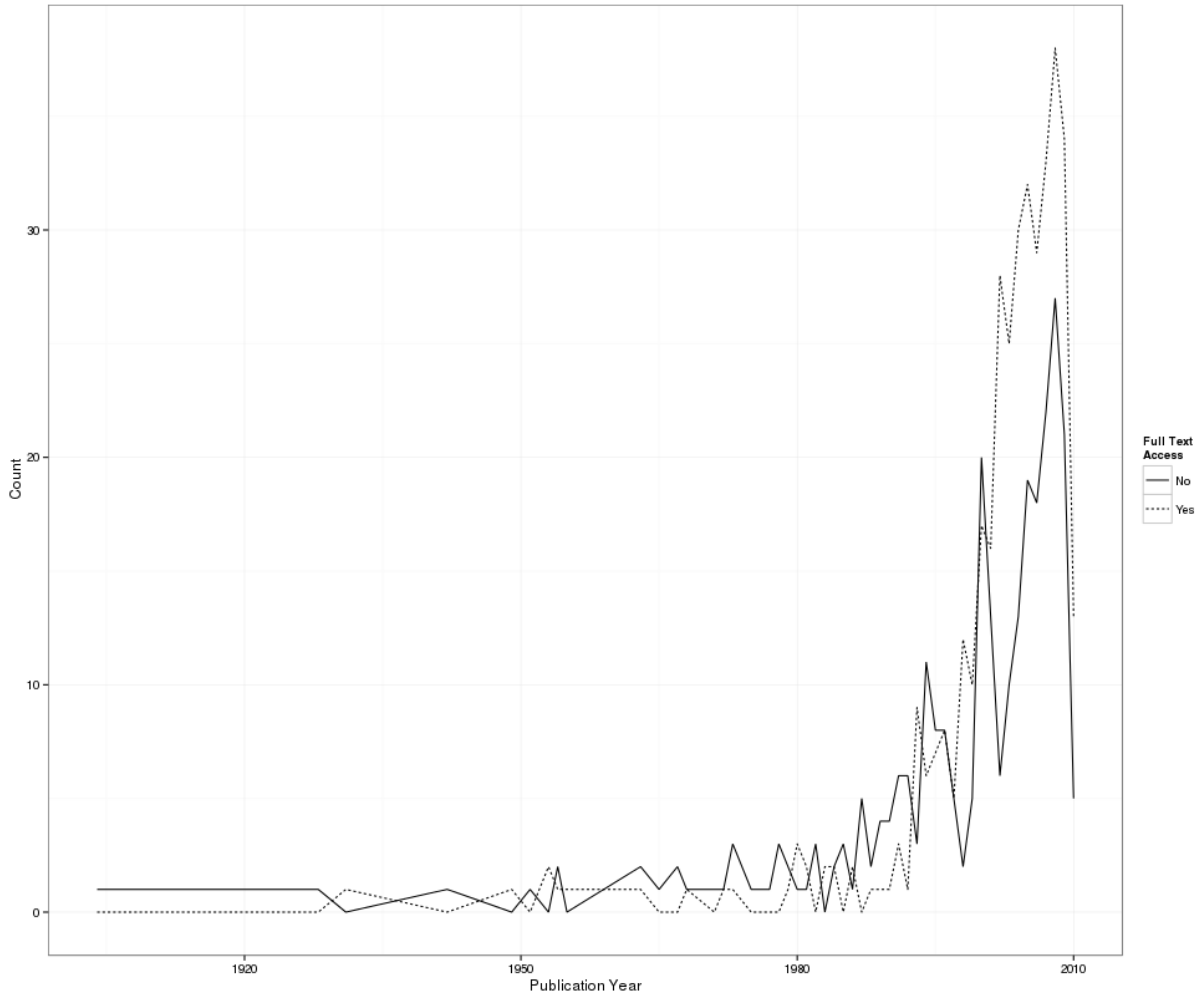

Fig. 2. 2012 Article Data.