Measuring the Value of Library Content Collections

Daniel M. Coughlin
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802
dmc186@psu.edu

Mark C. Campbell
College of Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802
mcc171@psu.edu

Bernard J. Jansen
College of Information Sciences and Technology
The Pennsylvania State University
University Park, Pennsylvania 16802
jjansen@ist.psu.edu

ABSTRACT

This research uses the methodology of web analytics to examine the usage of subscription databases at a major academic library. Our research goal is the development of key performance indicators from which academic libraries can evaluate the business value of their content collections. There are 1,447 databases to which this academic library provides access, and these databases received nearly 2.5 million customer visits in 2012 via the library's meta-search application, which is used for searching these databases. As such, these visits represent a substantial subset of the total traffic to the university's academic databases. The first level analysis shows that the top 20 most used databases represent over half the traffic to these academic databases. The second level analysis compared these heavily used databases (20) categorizing them by provider, and quantifying them with the remaining databases (428) from these providers. These results show the inequality of traffic generated by the top databases relative to the remaining databases from these providers in the context of search. The implications of this inequality illustrate the extreme usefulness of select databases and the possibility of the dispensability of less popular databases. The third level of analysis is a temporal evaluation of demand of databases over the course of two semesters (spring and fall 2012). This evaluation displayed the lack of increased demand throughout a semester beyond the top 300 databases. We used this analysis as the beginnings of the formulation of a set of web analytic metrics tailored for academic libraries.

INTRODUCTION

As a case study, the major academic library used for this research spends more than 11 million dollars on electronic collections annually (Furlough, 2012) in order to give their students and faculty access to a large set of peer-reviewed research articles. Access to these articles is provided in a number of ways. Users can access a journal or a database of journals and search for material within that journal (or database) by performing a search directly on the provider's web site. However, the university has implemented an aggregated search capability, providing access to discovering journal content from a more broad perspective. This advancement provides a full suite of discovery tools and features such as faceted search, filters, and sorting mechanisms. This new functionality provides a much-needed ability to search more than one provider, aggregator, and journal at a time.

The library provides access to over 1,447 databases that received nearly 2.5 million clicks via the meta-search alone in 2012. Each click represents successful access to a database from meta-search discovery tools. It is clear that the meta-search is a heavily used tool; what is less certain is the value of the databases represented and aggregated within this broad search. Are all of these databases used? If so, how often are they used compared with other databases? Ultimately, are all of these databases necessary and worth their subscription fees? This analysis begins to quantify the value of particular databases by defining key performance indicators based on web analytics metrics and measures (Jansen, 2009).

Most academic libraries provide access to journal databases, for which libraries pay publishers. One could make an analogy of journal subscriptions (hereafter, we use the term *journals* to mean journal articles, conference proceedings, or other subscription content) in subscription databases to the way cable/satellite subscriptions are managed, basically in packages or bundles of channels. However, in the world of cable providers, there seems to only be a handful of providers, whereas there are many more database providers and packages offered by each provider. The large number of providers makes it difficult to keep up with what each provider offers, leading to potential duplication of content across databases.

Providers sell access to bundles of journals they refer to as databases. If a researcher at a university needs access to an article in a journal, then the researcher's university must have access to that journal via a database, or they can pay a one-time fee to download that particular article, similar to a pay-per-view event or "on-demand" access. However, if someone is continually ordering on-demand shows from a channel such as HBO, at some point, it makes more sense to pay a monthly fee to access that channel and all the content provided by it. The same is true of articles from a journal.

Conceptually, this provides a nice analogous framework to understand academic subscription deals with libraries; however, the deals libraries have with the providers of this content, in practice, is not this simple. Montgomery *et al.* (Montgomery & King, 2002) point out "subscription" in the electronic world is not a simple payment for the annual content of a journal title. An electronic subscription often brings with it several years of back files. The price models and electronic content vary so radically that it is necessary to define four electronic journal subscription types to try and categorize the levels of complexity: Individual Subscriptions, Aggregator Journals, Full-Text Database Journals, and Publishers' Packages (Montgomery & King, 2002).

In order to provide access to a particular outlet, libraries frequently need to purchase a bundled package of journals. Continuing with this analogy of cable providers, if you would like to watch *The Daily Show* on Comedy Central, then you don't just buy *The Daily Show*, or even the channel Comedy Central, but rather you buy a package of channels that includes Comedy Central plus other channels.

An additional concern, because of complex packaging offers and large volumes of data to track, is duplication of content and the unfeasible costs associated with managing various formats, subscriptions, and indexes (Maple, Wright, & Seeds, 2003). Finally, understanding which content provides value becomes a concern as "different digital formats, interfaces, pricing structures, and access restrictions complicate our ability to evaluate journal resources using consistent measures" (Mercer, 2000).

Because of this situation, academic libraries spend large amounts of money on database subscriptions annually. Are all of these necessary? As state funds for many public institutions continue to diminish, libraries need to provide hard numbers to help define their business contribution (Conyers & Payne, 2011). If there is a known cutoff on the number of databases that have an impact, then strategic decisions can be made concerning the databases of lesser impact. Additionally, it is important for libraries to demonstrate the value of continued subscriptions for their patrons.

This research presents a web analytics methodology that permits libraries to evaluate their content collections via the usage (both cumulatively and temporally) of the databases to which they subscribe. Showing the use of these databases quantifies their value as the value only exists if titles are used (Montgomery & King, 2002).

LITERATURE REVIEW

Libraries are a critical part of academic institutions. Research shows the importance of libraries on the ability to retain students and increase classroom performance (Nackerud, Fransen, Peterson, & Mastel, 2013). In addition to providing an area to congregate for the purpose of study, distributing access of scholarly resources is a critical function of the library. As libraries deliver an increasing proportion of their services through the Web, the need to accurately and comprehensively track the use of library websites, online resources, and services is more important than ever (Conyers & Payne, 2011). An investigation on innovations and how widely they are adopted, as well as their measure of cost effectiveness, provides insight into how library resources can be allocated to better serve their users (Conyers & Payne, 2011).

Over the past 15 years, one such innovation in the library domain is publications moving online. 'Going digital' changes the landscape with regards to providing remote access to publications and discovery of information becoming more robust in the form of online search engines. There are various operational costs that are incurred when shifting management from print to digital, such as (but not limited to) training staff and providing the needed digital infrastructure. So, the costs associated with libraries providing access to digital journals goes beyond subscription fees, which are many times substantial in themselves.

Some figures show operational costs being twice as much as subscriptions costs, so they still represent significant costs that need to be understood (Montgomery, 2000). The shifts in costs are not entirely easy and many times don't make providing overall journal access entirely cheaper. Montgomery points out the common misconception that digital access is cheaper, as this has yet to be proven. Therefore, understanding the total costs associated with journal access has become imperative for libraries to communicate in order to financially survive and maintain adequate funding as the number of competitors in the digital world for access is greater than the number of competitors existing in the previous print domain (Montgomery, 2000).

At the time of this study, the pricing of journal subscriptions remains high, they are rising higher than inflation, as they have for some time (McCabe & Snyder, 2005). As costs rise (both operationally and subscriptions), it becomes essential to understand the organizational and business value of the databases to which a library subscribes. The value (Agarwal, Wang, Xu, & Poo, 2007) of a subscription is based on an understanding of what customers are using; price along with demand often determines the value (Holmström, 2004). It is important to understand this value because of the significant costs associated with many subscriptions.

In many science and technology libraries, much of their material budgets are associated with subscription costs (Montgomery, 2000). Users of these databases are likely to have particular journals and conference proceedings that align more closely with their research. If libraries are not aware of which journals are being accessed, they may potentially cancel journals of high consequence. In addition, removing journals could affect the quality of knowledge being shared and ultimately the output of the researcher (McCabe & Snyder, 2005; Montgomery, 2000).

Cox indicates that an expensive journal well used may be a better value than an infrequently used low-priced title (Cox, 2003). Although there are various methods for library collection management (e.g., Slote analysis, CREW (continuous review, evaluation, and weeding), MUSTY (misleading, ugly, superseded, trivial, and no use in the library collections), Taylor's 15/15 rule), Holmström indicates understanding the costs is a difficult task because of consortia deals, bundling of electronic-only journals, and bundling of print and e-journals that can make cost allocation more complex, as it was easier to comprehend and subsequently predict the costs of single journal subscriptions during the print era (Holmström, 2004).

There are a large number of publishers and an increasing variety of journals that have complex pricing models and bundled journals that make the comparability of these journals rather elaborate (Holmström, 2004). Bundled journal subscriptions appear to have a cost savings, particularly when evaluating on cost per title basis; however, this cost savings may only be near term. "Potential savings in subscriptions, subscription service fees, and serial processing costs may be tempting, but there are cascading consequences to these actions to be considered. Whatever these libraries hope to save in the short run may be lost in the long run, as publishers and aggregators combine to break this model with higher database licensing fees and embargoes in their efforts to recoup lost revenue." (Bell, 2001).

Understanding the full operational cost of online collections (academic databases, journals, etc.) is a difficult objective because it is not always easy to understand the demand for an online collection of what the usage (Agarwal, Xu, & Poo, 2011) is for the online content. Understanding the demand of a collection is important because it helps justify the cost; the business rationale becomes self-evident when we know a collection is highly used. One method to understanding the usefulness is to analyze the digital demand. The ability to determine how often a journal is accessed and comparing that number to other journals can demonstrate which journals have a better cost justification.

We acknowledge that these statistics highlight frequency and demand of a journal, but they fail to determine the motivation behind the demand. While some might argue there is a danger in believing these numbers tell the full story, even the critics of these numbers believe there is value in using these numbers as elemental to a more comprehensive analysis (Gatten & Sanville, 2004).

In many ways, getting access to statistics that demonstrate the demand of collection usage is seemingly easy — we can get more numbers than ever before with digital access to collections. In theory, generating these statistics for online collections should be easier than with print. In practice, getting these statistics from across dozens of providers and hundreds of databases is no simple task and making sense of the statistics becomes a hurdle due to the complexity (Maple *et al.*, 2003; Montgomery, 2000). This complexity exists in the form of finding data logs, and understanding what gateways exist to these collections, as many collections can be reached from multiple methods. Finding a single intersection of this information can be daunting, if not nearly impossible. Statistics provided by vendors might seem to be the closest method to a single point of aggregation in many cases. Unfortunately, statistics provided by suppliers can be flawed and are inconsistently logged across publishers (Montgomery & King, 2002), although they are still considered many times to be the best source of usage information (Ashcroft, 2002). In cases where the vendors provide the ability for this data to be exported, there is the ability to further aggregate, manipulate, and evaluate to tell a new story on the value of these collections, which is, again, no simple task.

There was an abundance of studies on the value of digital journals compared to that of print journals in the years following many journals going online (Degener & Waite, 2000; King, Boyce, Montgomery, & Tenopir, 2007; Montgomery & King, 2002). This discussion has pivoted toward the complexity and cost of bundled package subscriptions from content providers and the value of the provided services. One metric used to help determine value has been usage statistics (Chen, 1972). Increasing digital content and digital services such as meta-search have made it critical to understand usage of both services and content to better understand the value provided.

There is little research for meta-search engines, like the meta-search tool, to indicate content demand in libraries. This may be due to the newness of the technology rather than a lack of perceived importance on the role of search and discovery of information retrieval. However, this type of analysis can help provide an understanding of the current and potential value of a tool, such as meta-search, as well as the databases indexed, by showcasing the usage of the tool and the popularity of databases used from within it. As such, a meta-search tool can provide indicators of usage and database value, while simplifying the data aggregation process.

Without an understanding of what is being used and the value certain content collections provide, libraries are unable to get a grasp of what content is expendable and what content is essential. In the wake of no measure for journal demand, libraries may buy as much content as they can afford instead of focusing on the type and level of content they may need (Scigliano, 2002). Despite the difficulties associated with subscription models, content models, and usage statistics, it is important for libraries to develop a more complete understanding of the value their subscriptions provide. It is with this knowledge that both libraries and providers can justify the price tag. This is the motivation for our research.

RESEARCH QUESTION

In the effort to create a descriptive framework in order to measure value based on demand, for databases in search results our research questions are:

Research Question 1 (RQ1):

What is the distribution of database usage?

Providing access to 1,447 databases is costly not only due to subscription fees and operational costs, but also from indirect corollaries such as keeping up with the deluge of content that is made available. If these databases receive a fair share of traffic relative to costs, these costs are more easily justified. Our premise is that low access to multiple databases from these providers might not be necessary. For these databases, there may be other pricing models to provide access to this online content. Although some of these databases might come with no subscription fee, the operational costs as well as the unintentional negative effect of creating an unmanageable amount of data make them dispensable.

Research Question 2 (RQ2):

What is the relative demand among databases by provider?

For example, if a provider gives access to a set of databases, how does the demand of its most popular database compare to that of the least popular database, and based on this comparison is it able to be determined which databases are providing little value in this context? The scope of this analysis was limited to the providers of the 20 most heavily used databases, which generate over half of the traffic from the meta-search tool.

Research Question 3 (RQ3):

What is the temporal usage trend?

A temporal analysis allows us to distinguish database variations in usage and uncover variations in demand that may be hidden solely via an aggregate analysis. Databases that are seldom used overall and receive little to no increase in traffic during the course of a semester may not provide enough value for the costs (either direct or indirect) associated with them to maintain access to them.

METHODOLOGY

Web Analytics

This research uses web analytics as the method to quantitatively examine databases used by the library patrons. Web analytics is the measurement, collection, analysis, and reporting of Internet data for the purposes of understanding and optimizing Web usage (Jansen, 2006). We aim to modify web analytics techniques from the ecommerce domain to academic libraries. For the purpose of this study, web analytics were used to quantify the usage of each database by leveraging a search engine that indexes metadata from the library database subscriptions, and then further categorize and quantify the usage based on the provider of the database.

One of the weaknesses of this method is the inability to track user motivation and satisfaction (Conyers & Payne, 2011). It is not possible, explicitly, through a usage analysis, to determine why someone is using a particular database over another, or if they are satisfied with the results provided by this database. This analysis focuses on the ultimate strength of analyzing web logs via usage to determine the use of databases, the frequency with which they are used, and comparing that frequency with other data from the web logs (i.e. how often other databases are being used in comparison).

Definition of Terms

In this analysis three terms (provider, database, and usage) warrant further clarification. *Providers* are companies that provide digital content to the library in various formats (digital, print, etc.), subscriptions rates (open access, annual subscription, multi-year subscription), and content packages (databases, journals). A *database* is an aggregation of multiple content types distributed by a provider. A simplified way to think of a database is an accumulation of journals from one provider. Finally, u*sage* implies a click within a meta-search results page that subsequently brings a user to the provider of the desired content's web site.

Data Collection and Analysis

This analysis examines the usage for the 1,447 databases, and their corresponding providers, over the course of a full year from January 2012 through December 2012 at this university library. This investigation used the search logs from the meta-search tool. We first generated a listing of all databases, the corresponding providers, and the usage the database received (via the meta-search tool) for the year 2012. Based on the results of that report, we generated other subset reports, for example, database usage by top 20 providers, databases for individual providers and the corresponding click count, as well as temporal reports breaking down usage by month and week. These reports allowed for a more granular comparison of the complete list of databases from the top providers as well as how one provider's many databases compare against each other.

This data set could be used from a holistic perspective to evaluate the demand of any number of databases (RQ1), or the demand of a particular database when compared with other databases from that provider over the course of a full year (RQ2). Additionally partitioning this dataset by weeks allowed this evaluation to include spring and fall semesters and assess the demand of databases during the semester in comparison to before the semester (RQ3).

RESULTS AND ANALYSIS

What is the Demand of the Most Popular Databases?

This investigation explains where the largest digital impacts are being made, and where there is a drop-off in the value based on database access via meta-search. Understanding this gives insight to the demand of the most popular databases (RQ1). There were nearly 2.5 million clicks via meta-search (2,437,603) in 2012, and the top 20 databases received 1,221,482 of those clicks -- just over half (50.11%) of the traffic to the entire collection of this libraries databases (See Table 1). The remaining 1,427 databases received just the remaining traffic (49.89%).

Top Databases	Click Total	Percent of Usage via the meta-search interface
1	204,273	8%
20	1,221,482	50%
30	1,422,725	58%
60	1,741,069	71%
100	1,949,878	79%
120	2,017,018	82%
300	2,269,517	93%

Table 1. The most popular databases, by click, in 2012 and percent of meta-search traffic for which they are accountable.

Figure 1 illustrates the usage from meta-search across databases flattens off after a certain number of databases and at this number of database subscriptions the remaining databases receive a similar number of hits throughout the year.

The next step was to find the spot where, as defined by database number N, it was reasonable to perceive this flattening off. The method used to determine where the demand flattened off was to start with a small number of databases, and assuming the annual demand was noticeably changing, grow in number of databases until the flat line was more evident (a.k.a., a usage of the elbow method). The flat line represents a lack of increased annual demand for the remaining databases.

Figure 2 illustrates how the demand starts to taper off -- demand approaches a more straight curve in Figure 3; and Figure 4 clearly displays little variation in demand. These iterations were chosen to analyze demand due to the percentage of traffic they represent in meta-search (See Table 1). The top 30 databases were chosen as a starting point, because they represent nearly 60% of the usage via meta-search; the top 100 represent nearly 80%; and the top 300 represent over 93% of the traffic.

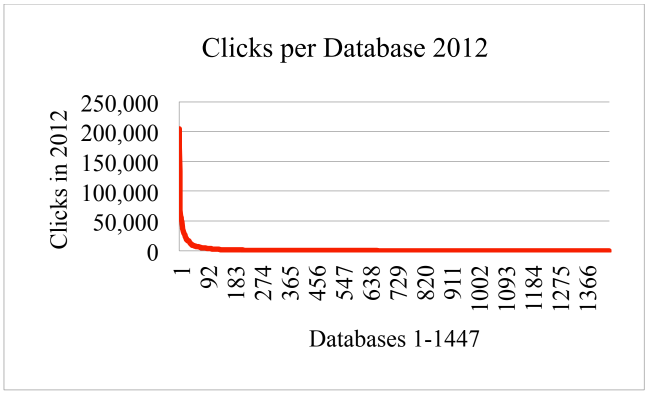


Figure 1. The total number of clicks, in meta-search, for each database, in order from most used to least used, in 2012.

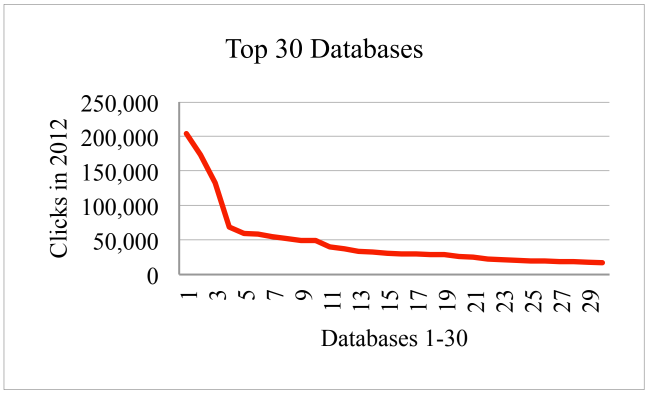


Figure 2. The total number of clicks, in meta-search, for the top 30 databases, in order from most used to least used, in 2012.

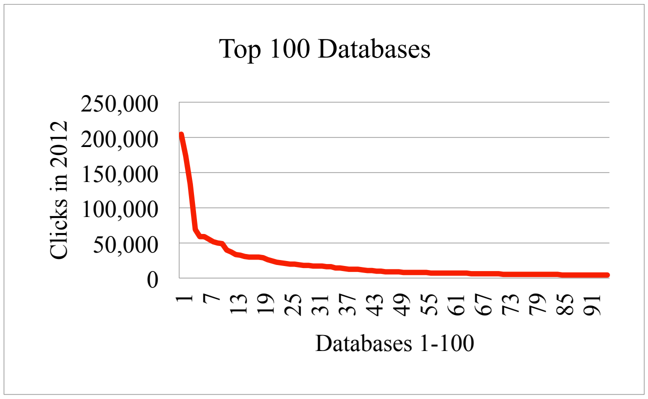


Figure 3. The total number of clicks, in meta-search, for the top 100 databases, in order from most used to least used, in 2012.

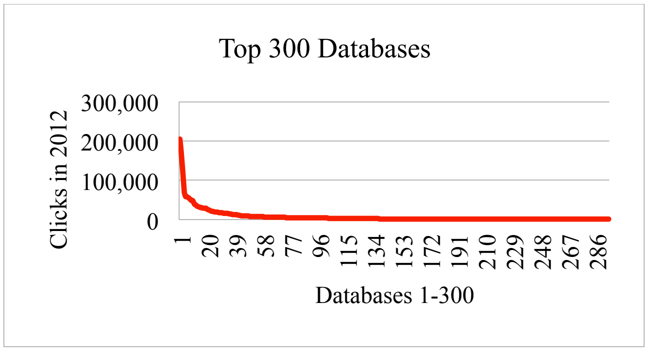


Figure 4. The total number of clicks, in meta-search, for the top 300 databases, in order from most used to least used, in 2012.

The remaining 1,147 databases generate just under 7% of the usage in meta-search, and each of these databases on average is used once every 2-3 days. The top 300 databases are each used on average 20 times a day. This is more than 50 times as often as the remaining 1,147 databases.

What is the Relative Demand Among Databases by Provider?

This examination aims to understand the distribution of database usage compared with other databases from the same provider to help explain the impact of all databases from a single provider, as subscription services can many times be bundled by the same provider.

Initially, the analysis showed all of the library databases and the respective providers for each database, as well as click totals. Then, the databases were arranged in order of most popular where popularity is defined by the usage of a database following a search. Once the databases were listed in terms of popularity, the next step was to determine the number of databases and subsequent providers to analyze. The goal of RQ2 is to understand the distribution of database usage for providers of the most heavily used databases; thus, focusing the evaluation on a subset of providers and their databases that account for more than half of all the usage in meta-search. Is the usage evenly dispersed, or heavily skewed?

To evaluate RQ2, *What is the relative demand among databases by provider,* the scope was limited to the providers of the top 20 databases (see Table 2) generating more than half the traffic analyzed. These 20 databases represent only 1.4% of the databases provided by the university, but they account for more than 50% of the usage in meta-search.

Out of the top 20 most popular databases there are 15 providers, and three of the 15 providers appear more than once because they provide access to more than one of the top 20 databases. In many cases, providers give access to multiple databases and a database may have access to multiple journals (see Table 2). Providers such as EBSCOhost, Gale, and ProQuest distribute access to multiple databases found in the top 20. In fact, all providers delivering access to a database found in the top 20 indeed furnish access to additional databases.

The 15 providers of the top 20 databases provide access to a cumulative 428 databases (the databases in the top 20, and an additional 408 databases). These providers distribute databases in the range of 5, from National Library of Medicine, to as many as 89 databases from Elsevier.

Database	**Provider**	**Clicks via Meta-search**	**Percent of Meta-search Usage**
1. Academic Search Complete	EBSCOhost	204,273	8.4%
2. ScienceDirect Journals	Elsevier	173,313	7.1%
3. LexisNexis Academic	LexisNexis	133,181	5.5%
4. Wiley-Blackwell 2010 Full Collection	Wiley-Blackwell	68,691	2.8%
5. PubMed Central	National Library of Medicine	59,112	2.4%
6. JSTOR Arts & Sciences I Archive Collection	JSTOR	58,711	2.4%
7. Web of Science	Thomson Reuters	54,931	2.2%
8. ProQuest Direct Complete (Legacy Platform)	ProQuest	51,858	2.1%
9. Business Source Premier	EBSCOhost	49,397	2.0%
10. PsycINFO	ProQuest	49,072	2.0%
11. ABI/INFORM Complete	ProQuest	39,605	1.6%
12. SpringerLink Contemporary (1997 - Present)	Springer-Verlag	37,258	1.5%
13. Gale Virtual Reference Library	Gale	33,760	1.4%
14. SAGE Premier 2012	SAGE Publications	32,634	1.3%

15. American Chemical Society Web Editions	American Chemical Society	30,883	1.3%
16. Access World News Research Collection	NewsBank	29,967	1.2%
17. CQ Researcher	CQ Press	29,700	1.2%
18. WorldCat (OCLC FirstSearch)	OCLC	29,359	1.2%
19. Opposing Viewpoints In Context	Gale	29,196	1.2%
20. Communication & Mass Media Complete	EBSCOhost	26,581	1.1%

Table 2. Listing of top 20 databases, corresponding providers, clicks received, and percent of meta-search usage in 2012

The next step in the analysis was to look at the usage that each database received compared with the total usage that provider received from meta-search. The top three providers in terms of usage for all databases from a provider were EBSCOhost, ProQuest, and Elsevier (see Table 3). Also, as Table 2 indicates, the Academic Search Complete database from EBSCOhost received more usage (204,273 clicks) via meta-search in 2012 than all 89 Elsevier databases received combined (203,061) according to Table 3. This shows a high level of disparity in usage even among the most popular databases. Also, this shows that the additional 88 databases that Elsevier provides access to do not generate much traffic, at least via meta-search.

Provider	Number of Databases	Total Clicks	Percent of Meta-search Usage
EBSCOhost	32	339,086	14%
ProQuest	62	265,131	11%
Elsevier	89	203,061	8%

Table 3. Total number of databases from the top 3 providers based on cumulative usage of all the provider's databases and the percentage of usage each provider receives within meta-search.

Traffic to Elsevier represents nearly 8.3% of all database usage in meta-search. Elsevier is of particular interest to look at further because they have the highest subscription fee and provide access to 89 databases, more than any other provider in the top 20, to generate those 203,061 clicks. Three of their databases account for over 90% (185,221) of their usage and the top nine databases account for over 95% of all usage for Elsevier (see Figure 5). There are 85 database subscriptions provided by Elsevier that receive less than 1% of Elsevier's usage (203,061 clicks). These databases either have little or no value in an aggregated search engine, such as meta-search, or if these databases have value in a search engine such as meta-search, work should be done to increase their visibility in search results beyond the top databases they provide. Conversely, a database such as ScienceDirect Journals has tremendous value in meta-search based on usage because it receives 85% of all the usage for Elsevier traffic in meta-search.

The inequality represented by Elsevier's databases, where so many databases represent so little usage in meta-search, is typical for providers in the top 20 and the databases with which they provide access. In total, more than 300 databases (over 75%) of the 428 analyzed represent less than 1% of the usage in their respective provider's annual click-through numbers.

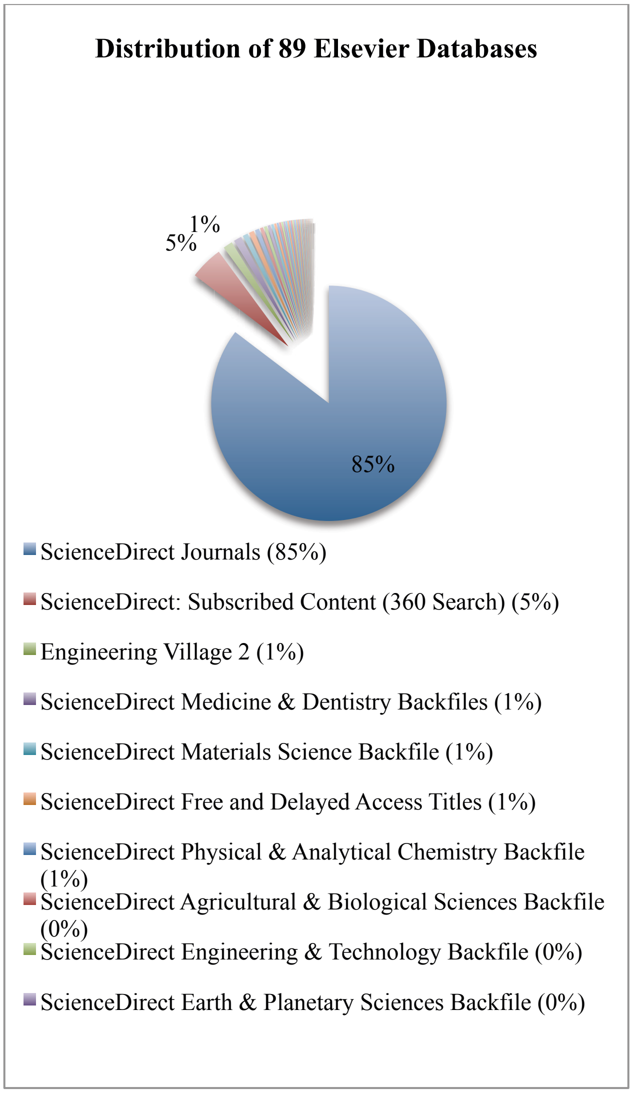


Figure 5. Elsevier 2012 database click distribution in meta-search (legend included is for the top 10 used databases, 0 are ~0).

EBSCOhost provides access to three databases in the top 20, one of which (Academic Search Complete) is the most popular database in meta-search, and generates more usage than any other provider. Figure 6 shows a distribution of usage in the 32 databases from EBSCOhost compare against each other. From Figure 6, it is apparent that the value of the remaining 29 databases, particularly the last 25 receiving less than 1% of the EBSCOhost usage, do not provide nearly as much value to meta-search as do the top databases from EBSCOhost.

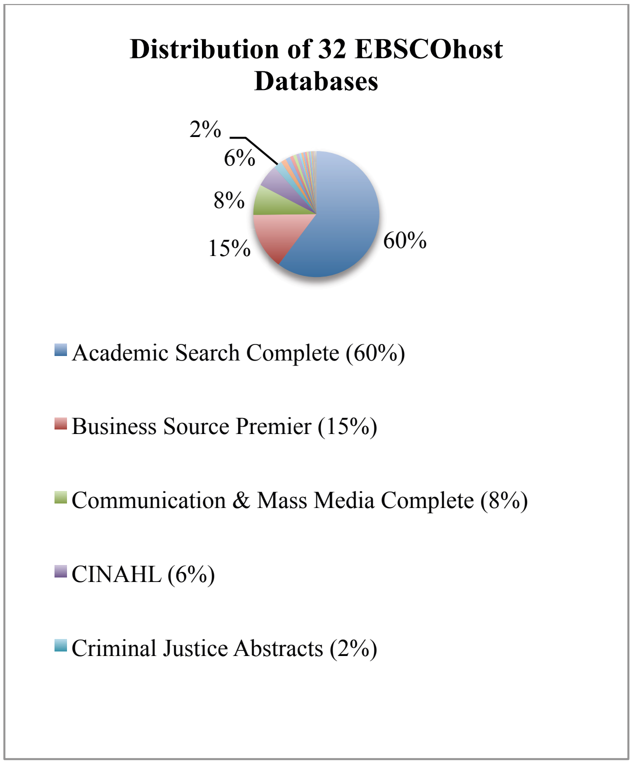


Figure 6. EBSCOhost 2012 database click distribution in meta-search (legend included is for the top 5 used databases).

Another portrayal of this inequity is that ProQuest is the only provider in the top 20 giving access to more than five databases receiving more than 5% of the traffic generated for that provider. However, looking at ProQuest's usage, 48 of the 62 ProQuest databases account for less than 1% of the total ProQuest usage (265,131 clicks) in 2012 (see Figure 7).

EBSCOhost, ProQuest, and Elsevier offer interesting analysis due to the number and popularity of databases they offer. EBSCOhost represents the most popular database, the most popular provider, and has three databases distributed in the top 20 (ProQuest and EBSCOhost both have 3 databases in the top 20, more than any other provider). ProQuest is has the most databases in the top 20 (along with EBSCOhost), but it also has a more evenly distributed click-per-database ratio than other databases in the top 20. Finally, Elsevier furnishes access to more databases than any other provider in the top 20 and makes an interesting case to see the distribution of these databases within meta-search. Also of interest are the providers in the top 20 that provide access to a smaller number of databases.

National Library of Medicine provides access to five databases. This is a fewer number of databases than any other provider that appears in the top 20. There may be an expectation that a smaller number of databases will represent a more evenly distributed demand; however, this is not the case. In fact, National Library of Medicine has an even more narrow distribution. National Library of Medicine receives over 98% of their traffic to PubMed Central Database (see Figure 8). The remaining four databases generate minor usage (920 total clicks) via meta-search in 2012. This is less than a click per day on each database.

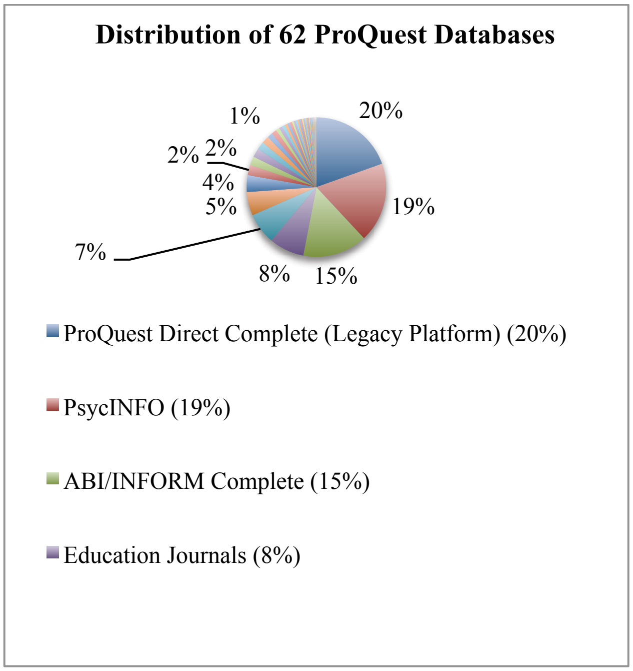


Figure 7. ProQuest 2012 database click distribution in meta-search (legend included is for the top 5 used databases).

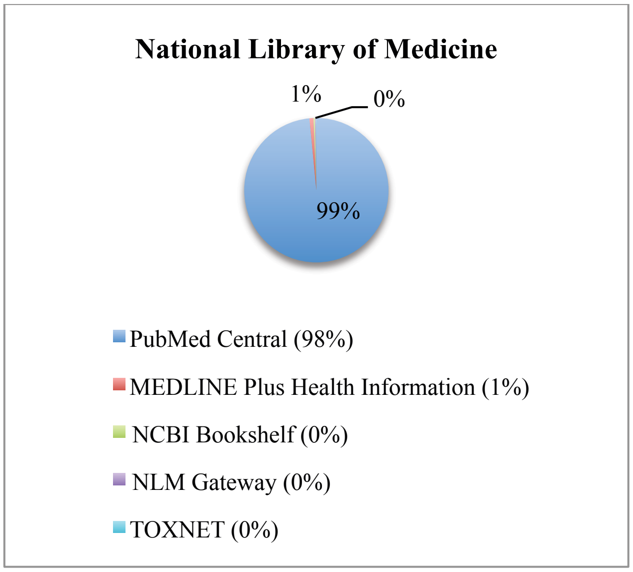


Figure 8. National Library of Medicine 2012 click distribution for their five databases in meta-search. NLM Gateway and TOXNET may not be easily distinguished due to their small click number representation.

These results display the disproportionate usage between the top rated databases and the remaining databases of these providers. Ten of the fifteen providers analyzed supply access to more than ten databases; all providers evaluated distribute more than five databases. These results show that many of these databases, when compared with other databases of the same provider, are rarely used within meta-search and this lack of use brings their value within meta-search into question.

What is the Temporal Usage Trend?

This assessment determines if there are databases that see an increase in demand during a semester compared to the week before a semester starts. Based on the results of this analysis we can answer *what is the temporal usage trend* (RQ3). A temporal analysis will give an evaluation of database demand during the course of the semester and determine what databases are used with more (and less) frequency.

We conducted analysis on the top 100 databases on a week-by-week basis. Comparing the top 100 databases with the remaining 1,347 begins to show a flat line over the course of a semester, but there is still a relative increase (and decrease) in usage (see Figure 9).

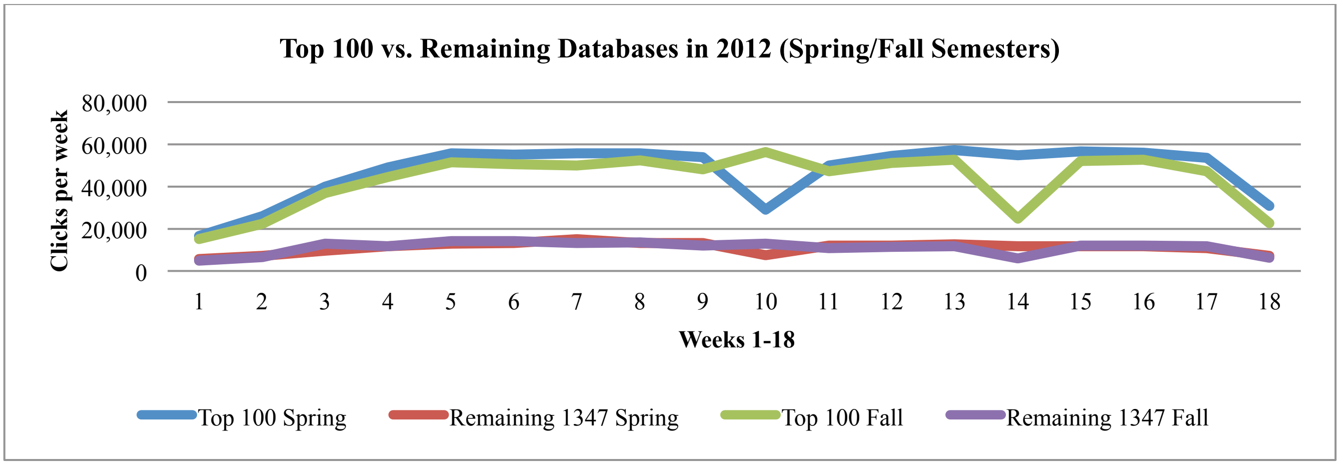
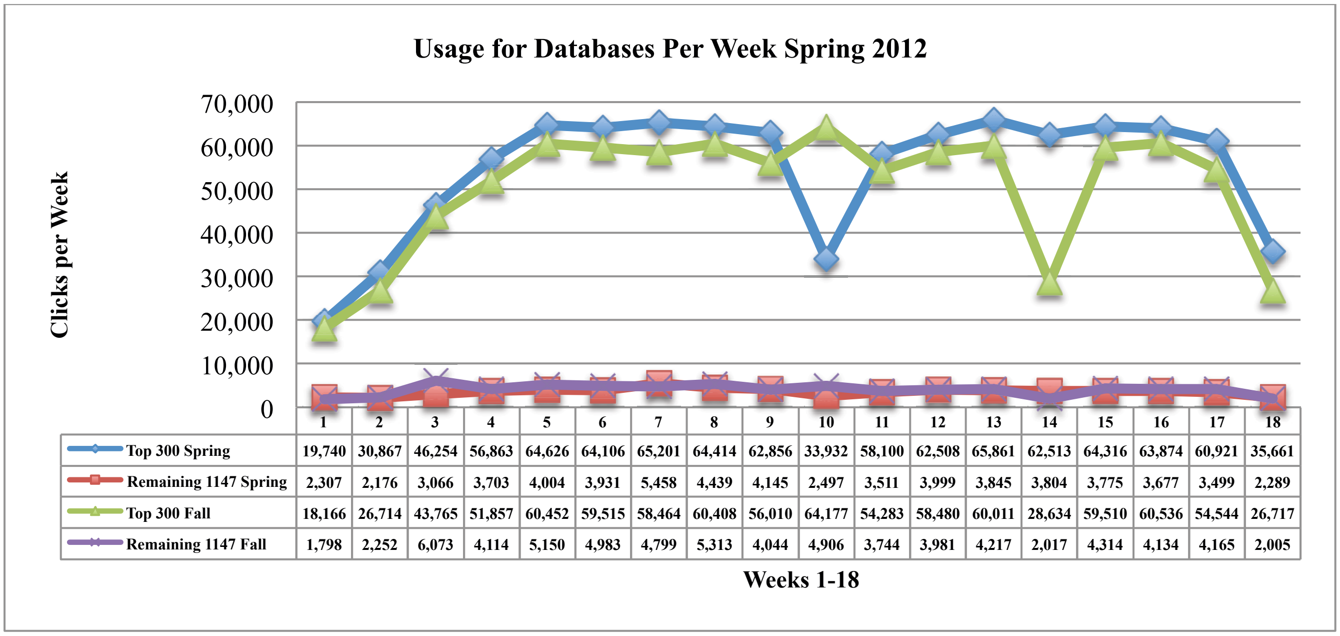


Figure 9. Week-by-week analysis of databases for spring and fall 2012.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Top 300 Spring	19,740	30,867	46,254	56,863	64,626	64,106	65,201	64,414	62,856	33,932	58,100	62,508	65,861	62,513	64,316	63,874	60,921	35,661
Remaining 1147 Spring	2,307	2,176	3,066	3,703	4,004	3,931	5,458	4,439	4,145	2,497	3,511	3,999	3,845	3,804	3,775	3,677	3,499	2,289
Top 300 Fall	18,166	26,714	43,765	51,857	60,452	59,515	58,464	60,408	56,010	64,177	54,283	58,480	60,011	28,634	59,510	60,536	54,544	26,717
Remaining 1147 Fall	1,798	2,252	6,073	4,114	5,150	4,983	4,799	5,313	4,044	4,906	3,744	3,981	4,217	2,017	4,314	4,134	4,165	2,005

Figure 10. Number of clicks (0-70,000) per database on a week-by-week (1-18) basis for the top 300 databases and the remaining 1147 in 2012.

Due to the relative fluctuation in demand at the remaining 1,347 databases (Figure 9), the next phase of this analysis was to do a week-by-week comparison for the top 300 databases and the remaining 1,147 (see Figure 10). Figure 10 shows the week-by-week demand of the top 300 databases, and the remaining 1,147 databases over two separate 18-week spans, both spring and fall 2012. The 18 weeks represent the week before classes start as week 1, weeks 2-17 represent the course of a 16-week semester, and conclude with final exams on week 18. There is a noticeable drop in usage for the top 300 databases in both the spring and the fall that correlate to spring break and fall break respectively. In the spring, the top 300 databases peak at 65,201 hits in week 7. On average each of those databases received 150 more hits that week than the week before the semester started.

We compared these numbers with the remaining 1,147 databases in meta-search that had a combined 5,458 clicks in week 7 (also representing the most significant traffic all spring), and these 5,458 clicks represent less than 3 clicks-per-database increase in usage during the semester than the week before the semester. The numbers in the fall are not much different where the increase ranges from 28-153 clicks-per-database for the top 300 and fewer than 4 clicks-per-database increase for the remaining 1,147 databases. It is also of note for the fall semester, if you remove week one, fall break, and finals week, the range for increased usage per database changes from having a low usage of 28 increased clicks-per-database (28-153) to having a low usage of 85 increased clicks (85-153). These 1,147 databases receive low increase during the semester, and nearly no fall off in usage during the spring and fall breaks whereas the databases with higher usage show large drops in usage (see Figure 10) during these breaks. This lack of drop off for the lesser-used 1,147 databases is another indication of how little their usage fluctuates during the semester.

DISCUSSION AND IMPLICATIONS

The purpose of this research was to investigate the relative value of digital content made accessible through library services. Providers of popular content account for a large percentage of the usage within library discovery services; however, not all the content from these providers receives the same proportion of usage. In fact, some of this content provides such minimal usage it brings the value of this into question. The analysis focused on an aggregation of all database usage (RQ1: *What is the demand of the most popular databases*), a comparative evaluation of usage (RQ2: *What is the relative demand among databases and their providers)*, and a temporal examination based on demand (RQ3: *What is the temporal usage trend*).

Theoretical Implications

This research extends the usage of content as an established metric to help define the impact and the business value of digital collections within the library. There have been many studies in libraries to help determine the value of digital collections, particularly in comparison to print collections (Montgomery & King, 2002; Montgomery, 2000). These studies show that costs go beyond subscription prices and there are indirect costs to consider before attaining more collections. Contemporary [discovery] services present search capability across multiple providers and collections and provide a new context to study the usage of digital collections.

This usage of these collections within meta-search can be extended with other key performance indicators (e.g., cost, citations, authorship, total usage, cost-per-title, cost-per-click, etc.) to provide a comprehensive framework to understand the real value of digital collections in the library. This work is a building block for such a future theoretical framework.

Practical Implications

Understanding the demand of digital collections in the context of library services provides insight into the value of these collections. By examining this demand through three perspectives, we are able to create a number of possible implications for library action. For example, actions could include removal of databases based on cumulative traffic they generate (or don't generate), the amount of traffic they generate compared to other databases from that provider, or those databases that see no increased usage over the course of the semester. Removal of a database might mean removing it from search results within meta-search or removal entirely. However, an alternative implication would be to investigate ways to increase traffic to databases that are used with less frequency yet still considered valuable. There may be digital collections that would provide considerable impact if one could increase their visibility, as cost of a journal goes beyond subscription, value goes beyond demand.

CONCLUSION

There are many key indicators to base the merit of academic collections on such as: cost, demand/usage (clicks), cost-per-click, cost-per-title, duplication of data, citations, authorship, and demand within search. This research provides a foundation for assessing the value of databases within the context of search by measuring demand as the beginning of the development of a web analytics framework for academic libraries.

This analysis provided three lenses to look at the demand of databases in the context of search. The first research question examined *the demand of the most popular* databases. Through this lens, we were able to distinguish where the annual usage of databases leveled off. Next, the analysis focused on *relative demand among databases and their providers*. In this perspective, databases were compared to other databases distributed from that provider. The databases represented include providers of some of the highest subscription fees for the university (Furlough, 2012). It would be interesting to view what percentage of a provider's traffic is generated in a search. This additional analysis provides a metric for the value of a provider's databases, as well as a metric for rationalizing the importance of search. Even so, this comparison made it evident that only the top databases receive consistent usage, while a majority of the databases receive less than 1% of all traffic for that provider and brings the value of these databases into question.

The final evaluation*, the temporal usage* trend, presented a temporal analysis to indicate demand increase as a metric for the value a database offers. This perspective compared the average increase-per-week each database received to describe those databases that receive no increase in use, and therefore, provide little value based on meta-search usage.

Using search as the context, it is evident in 2012 that there existed more content available than was being accessed on a regular basis. This additional content was a business cost to the institution. Libraries play a critical role in research institutions similar to the university and it is important they have the necessary funds to provide access to digital collections needed by researchers. Likewise, there are costs beyond subscription fees, both operational and logistical, that make it detrimental to provide access just for the sake of providing access. Perhaps the most important reason to prevent access to unused content is to allow librarians to have a more manageable amount of data to steward for the benefit of the research they serve.

One difficulty in measuring these numbers is the ability to understand the importance of less frequently used databases. Those that provide small access levels might provide all the database access for a particular area. For example, a database provided by ProQuest might only get < 1% of all the traffic that ProQuest generates; however, an entire department might rely on that database and not be able to survive without access to it. Additionally, it would be interesting to compare hits with competing databases. Perhaps there aren't a lot of hits on some of these databases when compared to the top Elsevier databases, for example, but they still might be the "industry leader" in a niche category. As noted, it is difficult if not impossible to measure complete value or motivation based on demand alone. This study represents one performance indicator of journal values within the context of search. It would be interesting to see what types of patrons use tools such as meta-search and what level of researcher (i.e., academic rank) they are considered.

When considering the analysis based on increased usage during a semester it would be interesting to see what accounts for the relative consistent demand of the remaining 1,147 databases. It is possible that the only demand received by these databases is a web crawler or something similar and that is why the demand stays the same. Average click-per-database over the course of the semester does not show the most popular 1,147 or least popular 1,147 -- only the average. It would be interesting, if not more powerful, to look at additional numbers for these databases that see little to no increase in demand. We reserve this for future research.

REFERENCES

Agarwal, N. K., Wang, Z., Xu, Y., & Poo, D. C. C. (2007). Factors Affecting 3G Adoption : An Empirical Study. *11th Pacific Asia Conference on Information Systems* (pp. 256–270).

Agarwal, N. K., Xu, Y. C., & Poo, D. C. C. (2011). A Context-Based Investigation Into Source Use. *Journal of the American Society for Information Science*, *62*(6), 1087–1104.

Ashcroft, L. (2002). Issues in developing, managing and marketing electronic journals collections. *Collection Building*, *21*(4), 147–154. Retrieved March 17, 2013, from http://www.emeraldinsight.com/10.1108/01604950210447386

Bell, S. J. (2001). The New Digital Divide: Dissecting Aggregator Exclusivity Deals. *D-Lib Magazine*, *7*(7/8). Retrieved March 31, 2013, from http://www.dlib.org/dlib/july01/bell/07bell.html

Chen, C. (1972). The Use Patterns of Physics Journals in a Large Academic Research Library. *Journal of the American Society for Information Science*, (August), 254–270.

Conyers, A., & Payne, P. (2011). Library performance measurement in the digital age. *University Libraries and Digital Learning Environments* (pp. 201–214). Ashgate Publishing.

Cox, J. (2003). Value for Money in Electronic Journals : A Survey of the Early Evidence and Some Preliminary Conclusions. *Serials Review*, *29*(2), 83–88.

Degener, C. T., & Waite, M. A. (2000). When Fools Rush In ... Thoughts About, and a Model for, Measure Electronic Journal Collections. *Serials Review*, *26*(4), 3–11.

Furlough, M. J. (Pennsylvania S. U. (2012). Opening Access to Research: From Concepts to Actions. *Open Access Week* (pp. 1–38). The Pennsylvania State University. Retrieved December 16, 2012, from https://scholarsphere.psu.edu/files/37720c723

Gatten, J. N., & Sanville, T. (2004). An Orderly Retreat from the Big Deal: Is It Possible for Consortia? *D-Lib Magazine*, *10*(10). Retrieved March 31, 2013, from http://www.dlib.org/dlib/october04/gatten/10gatten.html

Holmström, J. (2004). The Cost per Article Reading of Open Access Articles. *D-Lib Magazine*, *10*(1). Retrieved November 29, 2012, from http://www.dlib.org/dlib/january04/holmstrom/01holmstrom.html

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, *28*(3), 407–432. Retrieved March 15, 2013, from http://linkinghub.elsevier.com/retrieve/pii/S0740818806000673

Jansen, B. J. (2009). *Understanding User-Web Interactions via Web Analytics. Synthesis Lectures on Information Concepts, Retrieval, and Services*

(Vol. 1, pp. 1–102). Morgan & Claypool Publishers. Retrieved November 29, 2012, from http://www.morganclaypool.com/doi/abs/10.2200/S00191ED1V01Y200904ICR006

King, D. W., Boyce, P. B., Montgomery, C. H., & Tenopir, C. (2007). Library Trends. *Economics of Libraries*, *51*(3), 376–401.

Maple, A., Wright, C., & Seeds, R. (2003). Analysis of format duplication in an academic library collection. *Library Collections, Acquisitions, and Technical Services*, *27*(4), 425–442. Retrieved March 17, 2013, from http://linkinghub.elsevier.com/retrieve/pii/S146490550300085X

McCabe, M. J., & Snyder, C. M. (2005). Open Access and Academic Journal Quality. *Science (New York, N.Y.)*, *95*(2), 453–458. Retrieved November 29, 2012, from http://www.ncbi.nlm.nih.gov/pubmed/20888544

Mercer, L. S. (2000). Measuring the Use and Value of Electronic Journals and Books. *Issues in Science and Technology Librarianship*, *25*(Winter). Retrieved March 31, 2013, from http://www.istl.org/00-winter/article1.html

Montgomery, C. H. (2000). Measuring the impact of an electronic journal collection on library costs: A framework and preliminary observations. *D-Lib Magazine*, *6*(10), 37–52. Retrieved November 29, 2012, from http://www.tandfonline.com/doi/abs/10.1080/13614570009516951

Montgomery, C. H., & King, D. W. (2002). Comparing Library and User Related Costs of Print and Electronic Journal Collections A First Step Towards a Comprehensive Analysis. *D-Lib Magazine*, *8*(10), 1–17.

Nackerud, S., Fransen, J., Peterson, K., & Mastel, K. (2013). Analyzing Demographics: Assessing LIbrary Use Across the Institution. *Libraries and The Academy*, *13*(2), 131–145. Retrieved March 31, 2013, from http://www.press.jhu.edu/journals/portal_libraries_and_the_academy/portal_pre_print/current/articles/13.2nackerud.pdf

Scigliano, M. (2002). Consortium purchases: case study for a cost-benefit analysis. *The Journal of Academic Librarianship*, *28*(6), 393–399. Retrieved March 17, 2013, from http://linkinghub.elsevier.com/retrieve/pii/S0099133302003464