PREPRINT – Accepted for Publication in Information Processing and Management

The Basis for Bibliomining: Frameworks for Bringing Together Usage-Based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services.

Scott Nicholson
Assistant Professor, Syracuse University School of Information Studies
4-127 Center for Science and Technology, Syracuse, NY 13244
scott@scottnicholson.com

Abstract

Over the past few years, data mining has moved from corporations to other organizations. This paper looks at the integration of data mining in digital library services. First, bibliomining, or the combination of bibliometrics and data mining techniques to understand library services, is defined and the concept explored. Second, the conceptual frameworks for bibliomining from the viewpoint of the library decision-maker and the library researcher are presented and compared. Finally, a research agenda to resolve many of the common bibliomining issues and to move the field forward in a mindful manner is developed. The result is not only a roadmap for understanding the integration of data mining in digital library services, but also a template for other cross-discipline data mining researchers to follow for systematic exploration in their own subject domains.

Keywords: Data mining, data warehousing, digital libraries, bibliomining, evaluation, theory, library measurement, library evaluation

1. Introduction

Digital library services have changed the way patrons seek and access the high-quality information traditionally found only in libraries. Many traditional print-based libraries now spend a significant portion of their budgets on online services. In addition, new organizations not accustomed to solving information needs are providing these digital library services.

Users seek information through digital and physical libraries, sometimes through methods that allow the library to identify certain characteristics and other times as nothing more than an IP address. Administrators of all digital library services are faced with the same management issues: what types of users are accessing their services; which collections of resources are they accessing; and how can the digital library service be improved to better serve their users. Library scientists are also interested in how users are accessing these digital libraries, and because of the nature of the medium, can access detailed logs

about how users wander through these information spaces. One way to understand these issues and harness the large amounts of raw data created through digital library use is through the bibliomining, or the "application of statistical and pattern recognition tools to the data associated with library systems" (Nicholson, 2003, p. 146).

The goal of this paper is to explore the concept of bibliomining. First, the origins of the term and the relationship of bibliomining to its major components, bibliometrics and data mining, are explored. Second, the conceptual placement of bibliomining with other forms of evaluation in two contexts – digital library management and digital library research – are presented. Finally, a research agenda that resolves common issues and encourages for the mindful extension of bibliomining is developed.

Conceptually, the goal of this piece is to extend the range of analysis of the social networks that make up the community of authors and the community made up of the library and its users. Barabási (2003), in his modern classic, *Linked*, focused on the importance of exploring the interconnections between different social networks. Therefore, one primary goal of this bibliomining presentation is to explore the type of data warehouse that would allow for exploration of patterns of connections between authors, works, libraries, and users. An important secondary consideration is that the

goals of practitioners exploring library users and that of researchers exploring patterns of creation are different, and addressing these differences is an important concept in this work.

Researchers working in the application of data mining to other disciplines may find this work to be useful as a framework for their own discipline. Some of the arguments and frameworks presented in this document hold true in other disciplines, especially those that have to strike a balance between researchers and practitioners. The underlying concept of "how does data mining fit in" is one that needs to be addressed in each discipline where data mining tools are applied.

2. Origins and Definition of Bibliomining

Bibliomining is derived from the terms "bibliometrics" and "data mining," as the goal is to take advantage of the social networks that power both bibliometrics and user-based data mining through a single data warehouse. Why create a new term for data mining in libraries? The concept is not new; data mining has been occasionally discussed in traditional library settings since the late 1990's (Banerjee, 1998). The challenge comes from the terminology used; data mining packages contain a *library* of different algorithms. Therefore, searching for articles and scholars in the area produces many items not on the topic, such as *"Data Mining Library Reuse Patterns in User-Selected Applications"* (Michail, 1999). In order to make it easier for scholars working in the area of data mining for libraries to find each other and avoid the confusion of software libraries for data mining, the author coined the term "bibliomining" for Nicholson & Stanton's 2003 work.

In order to better conceptualize bibliomining, it is useful to first conceptualize the data needed for traditional bibliometrics and user-based data mining, then see how they can be combined to form the basis for bibliomining.

2.1. Bibliometrics

Traditional bibliometrics is based on the quantitative exploration of document-based scholarly communication (Borgman & Furner, 2002). Figure 1 demonstrates some of the data used in bibliometric research and the connections between different works. Works have authors and collections (journals, publishers, libraries) associated with them, and the works are connected through citations, authorship, common terms, or other aspects of the creation and publication process.

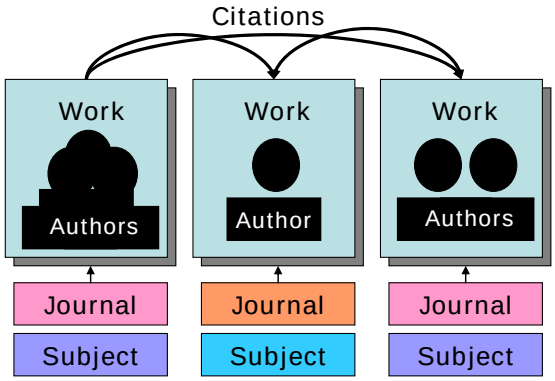

Figure 1: Data for Bibliometrics

Traditional bibliometric data involves information about the creation of the work such as the authorship and the works cited. In addition, metadata associated with the work, such as a general topic area or the specific journal in which it appeared, is connected to the data surrounding the creation of the work. Taking these data together allows the researcher to understand the context in which a work was created, the long-term citation impact of the work and the differences between fields in regard to their scholastic output patterns.

The analyses performed in traditional bibliometrics were frequency-based; however, many newer bibliometric studies are using visualization and data mining to explore patterns in the creation of these materials (Kostoff et. Al, 2001 for example; Börner, Chen, and Boyack, 2003 for a literature review). Some of the concepts explored included frequency of authorship in a subject, commonality of words used, and discovery of a core set of frequently-cited works (Borgman, 1990). Integrating the citations between works allows for very rich exploration of relations between scholars and topics, and these linkages between works are used to aid in automated information retrieval and visualization of scholarship

(White & McCain, 1998) and the social networks (Sandstrom, 2001) between those involved with the creation process. Many newer bibliometric applications involve Web-based resources and hyperlinks that enhance or replace traditional citation information (Cronin, 2001, for example; Borgman & Furner, 2002 for a literature review).

2.2. User-based Data Mining

One popular area of data mining in both the commercial sector and the scholarly literature is the examination of how users explore Web spaces. These studies focus on accesses of different Web pages by a particular user (or IP address). Patterns of use are discovered through data mining and used to personalize the information presented to the user or improve the information service (Wilkinson, Thelwall & Li, 2003; Eirinaki & Vazirgiannius, 2003). Figure 2 contains some of the data types used in this user-based data mining. The goal of this figure is to demonstrate that in user-based data mining, the links between works come from a commonality of use. If one user accesses two works during the same session, for example, then if another user views one of those works then the other might also be of interest. This figure representation links between works that result from the users.

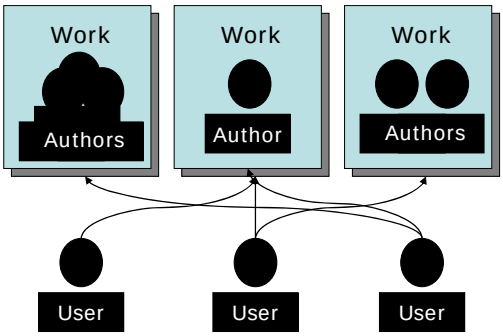

Figure 2: Data for User-Based Data Mining

One record in this data source is a single access of a data resource, and metadata attached to that record includes any identification available about the user, time and date information, and data about the referring Web site. Some studies append metadata about the work accessed in order to improve recommendation algorithms (Geyer-Schulz, et. al. 2003; Kao, Chang, & Lin, 2003). The patterns are focused on understanding how users explore the information space, and if there is some way to identify a user between sessions (through a cookie or a login), then the behavior of the users can be tracked over time. Since many digital library services require some type of login for access to leased and purchased digital materials, this type of usage data mining is possible and would be valuable in decision-making.

The challenge with implementing this type of exploration in a digital library setting is that of user privacy. Privacy of personally identifiable user information is of concern during the bibliomining process. Recent legislation in the United States has given government agencies the power to request any information kept by a library about an individual (American Civil Liberties Union, n.d.), regardless of the original purpose. The response to this by some libraries has been to delete and shred large amounts of data as soon as operationally feasible (Murphy, 2003), which has long-term consequences regarding the management and justification of library services.

Too much data may be being deleted in response to privacy-threatening government regulations, and this destruction has long-term consequences for traditional libraries and grant-funded digital library services called to justify the cost of their services to funding agencies and taxpayers.

One commonly suggested solution is to encode the user identification in the data warehouse. This would allow the management to still track use of items over time while not allowing the immediate identification of users from the data warehouse. Simply replacing the user ID with a code is not appropriate, however, as anyone wishing information about the user's behavior could use the encoding schema to learn the encoded ID for a user and then retrieve information about the user. Therefore, encoding may be a tempting choice, but if the encoding procedure is reversible or an encoding lookup table is kept, the privacy of the patron is still a concern.

2.2.1 Protecting Privacy through Demographic Surrogates

In order to properly protect the user's privacy, a different approach is needed. One route is to preserve only information about the item accessed and the time of access, deleting all information about the user. This would protect the identity of the user, but would also lose some potentially valuable information. Another option is to create a

demographic surrogate. Just as a bibliographic surrogate is a set of fields designed to represent a work, the demographic surrogate is a set of values from predetermined fields selected to represent a user in the bibliomining data warehouse. This set of variables will replace the personal information about a user in the bibliomining data warehouse.

The focus of the data collected for bibliomining is on the materials accessed and the services used and not the identity of the users involved. This task of discarding personal information differs from the steps in data mining for many corporate applications where the goal is to identify individuals. Most library decisions involving users are not made based upon the activities of individuals but rather the activities of groups of individuals. It is hypothesized, therefore, that activities done by large groups of individuals will be represented by patterns that will emerge through the bibliomining process.

One of the current five grand challenges of library research, according to Buckland, is that library managers must better understand their different user communities (Buckland, 2003). These user communities can be identified by demographic variables, which can be used to create a non-unique demographic surrogate for each user. For example, in an academic setting, department, classification, and general location (on-campus vs. off-campus vs. in-library) could make up this demographic surrogate. At the time of access, information about the item accessed can be enhanced with a demographic surrogate with user community information. The demographic surrogate will be a set of demographic values that will replace personally identifiable information about a user. Therefore, the library manager can gain understanding of how user communities are using their library resources without sacrificing the privacy of individuals to do so. Figure 3 is a model of the data for this type of community-based Web usage mining, and represents the same links as in traditional user-based data mining, i.e., links between works, but with demographic surrogates in place of the identifying information about the users.

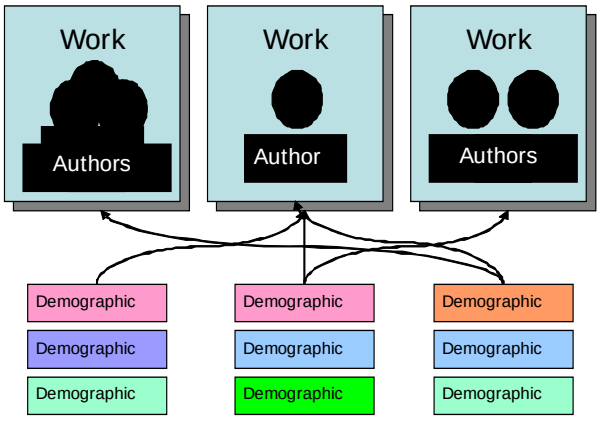

Figure 3: Data for De-Identified Community-Based Web Usage Mining

Replacing individual user identifications with a set of demographic values has significant implications for what types of data mining procedures can be used, and this will be discussed later as part of the research agenda for bibliomining.

3. The Bibliomining Data Warehouse

Both bibliometrics and Web usage data mining have a data field in common – the work accessed. Bibliometrics focuses on the creation of that work, and Web usage mining focuses on the access of the work. Combining these two data sources into one data warehouse allows researchers and library managers to more fully understand the information space they have created. Figure 4 shows the model for the bibliomining data resource, which demonstrates the concept of creating connections between works based upon both the creation process and the user population.

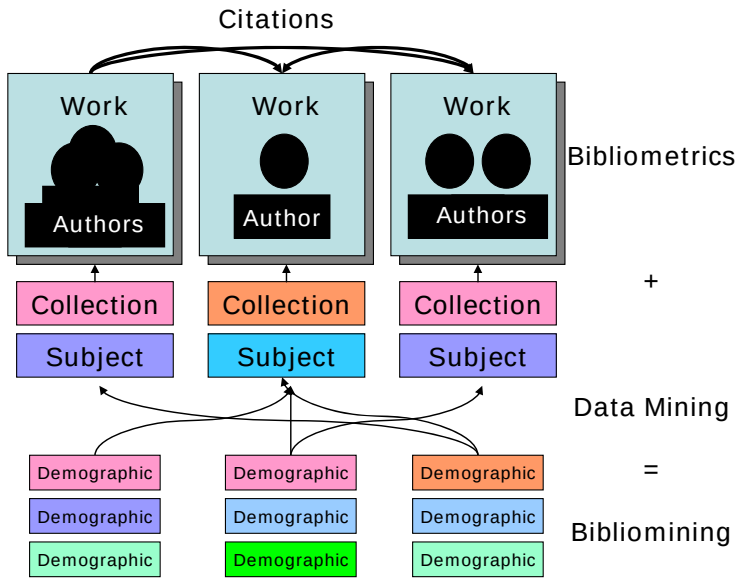

Figure 4: Data for Bibliomining

Bibliomining is defined as the combination of data mining, bibliometrics, statistics, and reporting tools used to extract patterns of behavior-based artifacts from library systems (Nicholson & Stanton, 2003). It has been rarely implemented in the full form as shown here due to the complexity of the data issues. By presenting the model, it is hoped that those developing data warehouses for digital libraries can keep the full bibliomining warehouse in mind as a goal as they develop smaller projects. Integrating bibliomining into current research and evaluation efforts will allow library managers and researchers a more complete idea of the resources contained in their library and how they are being accessed by users.

3.1 A Framework for the Data

The data that can support the linkages from both the creation/publication and use must live in the same data warehouse in order to allow the full bibliomining process to take place. A conceptual framework for these types of data is useful in determining what types of fields to keep from operational systems. There are three parts to this framework – data about a work, data about the user, and data about the service. These three parts will link together to represent one use, which is the base level for the data warehouse. The instance of a use of a library service connects a work (or works), a service, and a user together in the bibliomining data warehouse.

The first section of the data warehouse comes from the work. This will contain three types of fields – fields that were extracted from the work (like title or author), fields that are created about the work (like subject heading), and fields that indicate the format and location of the work (like URL or collection). This information can come from a MARC record, Dublin Core information, or the content management system of the library. This area also can connect into bibliometric information, such as citations or links to other works. This may require extraction from the original source (in the case of digital reference) or linking into a citation database. One challenge with the creation of this link is that the vendors currently report e-resource usage at the title level, and many bibliometric explorations begin at the article level. Standards for article-level reporting akin to the COUNTER aggregate formats are needed; once these are created, vendors can deliver more information about the specific items used at their sites.

The second area of the data warehouse is information about the user. As discussed earlier, this is where the demographic surrogate will be stored. In addition, other fields that come from inferences about the user will be stored here. For example, the IP address of the user can power inferences about the place of use. In an academic library situation, the IP address could be used to infer if someone comes from off-campus, on-campus, or in library. In some cases, the IP address could even tell from what building or computer lab a request came. A similar inference from the public library can come from the zip code. Both of these inference connections may provide a demographic guess as to the user's groups, but neither will provide a definite demographic match. Over a large enough dataset, however, patterns can still be seen from these demographic "best guesses."

The third area comes from one primary reason the library exists – to connect users to information, usually through works. The library service is the most difficult piece of this data to conceptualize because there are many different types

of services. Searching, circulation, reference, interlibrary loan and other library services have fields in common that can be captured in the data warehouse. In addition, each one has a set of fields appropriate for that type of service. A properly-designed data warehouse can handle both types of data; this enables evaluation of a service type or gaining understanding of library use across many services. Fields common to most services include time and date, library personnel involved, location, method, and if the service was used in conjunction with other services.

Each library services also has a set of appropriate fields. For example, searching has the content of the search and the next steps taken. Interlibrary loan will have cost, a vendor, and a time of fulfillment; circulation has information about the acquisition process of the work and circulation length. As with most decisions, the library's need for decision-making and the scholars' needs for research should drive the fields captured while still maintaining patron privacy. In order to aid with exploration, Figure 5 contains many other components and fields to demonstrate the conceptual framework for the bibliomining data warehouse.

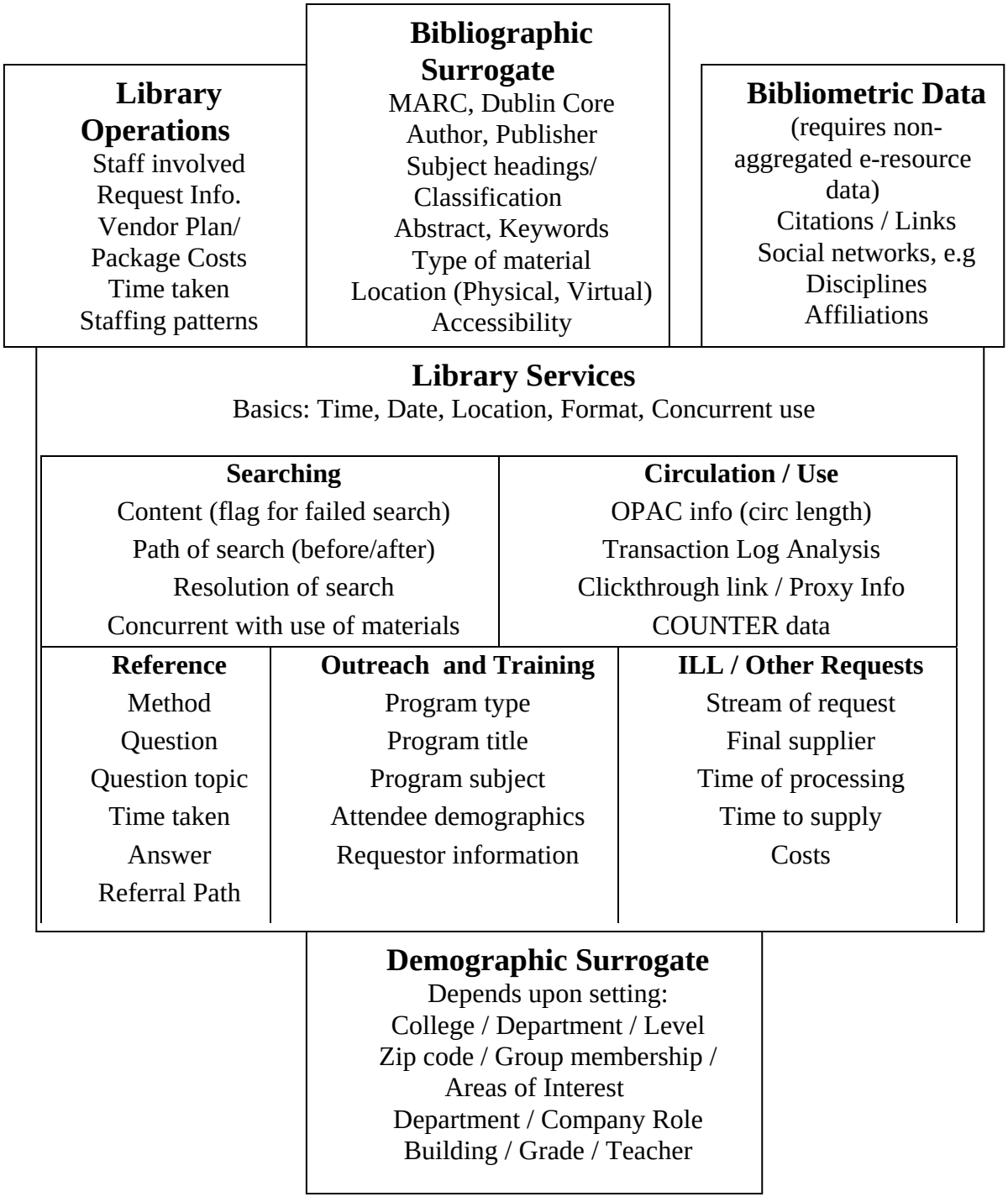

Figure 5. Conceptual Framework for Data types in the Bibliomining Data Warehouse

4. Using the Bibliomining Data Warehouse

Once this data source has been collected, it allows not only traditional measures and reports to be created, but also creates new opportunities for exploration. This section presents different tools useful in exploring the data warehouse; the goal of this section is not to explain how to use each type of tool, but rather to present the tools in context with the bibliomining data warehouse structure and how these tools could be developed and implemented.

4.1 Traditional Reporting

Traditionally, library decision-makers examine aggregates and averages to understand their service use. These measures can still all be created with this data warehouse, but it also has the advantage of empowering the managers and decision-makers to ask other questions. If only aggregates are collected and the underlying data discarded (or never made available), the ability to ask new questions is greatly reduced. Questions must be asked far ahead of time, and once these reports are set up, they can be difficult to change as measures may no longer be comparable over time. One argument for only keeping the aggregates is that it protects the privacy of the users; however, using demographic surrogates allows the library to keep addition data about their services and patrons while still protecting their privacy.

The advantage to the data warehouse is that new questions can be asked not only of the present situation but also, the past. As the data warehouse is a collection of past actions, new aggregates and averages can be collected from past data. This allows those doing evaluation or measurement to ask new questions and then create a historical view of those reports in order to understand trends. This would be an impossible task if only aggregates were kept; only new questions could be asked and there would be no background to put measures into perspective.

In addition, libraries can more easily understand behavior between different demographic groups in the library. Aggregates mask underlying patterns which might be able to be brought out if the same measures can be collected for different groups. Since the library serves sets of population, there will be situations where each group uses the library for different information needs and in different ways. For a simplified example, assume one population group uses reference and another uses electronic resources; then aggregating over all library users will result in very bland and unusable results – half of the time, people use reference and half of the time, people use electronic resources. Looking at the groups separately would bring these differences to light, which then produces much more actionable results.

4.2 Online Analytical Processing

Online Analytical Processing, or OLAP, is a method by which an analyst can explore a data set by examining a series of traditional-style reports through an interface that makes it easy to navigate through different variables, aggregation types, timeframes and other dimensions . Conceptually, OLAP is a decision support system that navigates through the results of a large set of previously-run database queries. When setting up the OLAP system, the developer determines dimensions of variables that might be interesting (time, demographic, metadata from the bibliographic surrogate) and extracts subsets of the data from the data warehouse to create a large number of traditional tabular reports. These are created and updated regularly, so that when someone is exploring these reports, there isn't processing time for each query to be run against the data warehouse (Chaudhuri & Dayal, 1997).

This OLAP tool creates an easy-to-use interface for librarians and researchers to use in exploring the data for interesting patterns by selecting variables for each dimension of the report, choosing timeframes, and selecting aggregation measures. After looking at a table, the explorer can then make a dimension like time or subject more or less granular. At any point in the exploration, a view of the data can be saved and turned into a regular report. In addition, the drill-down feature allows someone to access the underlying records that make up any cell in a report, which makes it much easier to understand the data behind the aggregations. This easy exploration allows a much larger audience the chance to explore the data warehouse than would be possible if each exploration required a new SQL query. Implementing an OLAP-style interface will remove a significant barricade that traditionally stands between librarians and the data about their users and services – they no longer need to write database queries or wait days for a request to be processed.

Joe Zucca has been instrumental in creating the Penn Library Data Farm (http://metrics.library.upenn.edu/prototype/datafarm/). This project is a data warehouse with an interface that allows librarians, researchers, and anyone interested the ability to explore a cleaned version of this library use data. While not a true OLAP, this data farm project represents one of the most advanced bibliomining-style projects in existence. The fact that Zucca has made this accessible to all is to be commended; it provides an example of what libraries and, more importantly, library system vendors, should move toward in order to enable a culture of exploration and evaluation.

4.3 Visualization

Another way of exploring data from the bibliomining data warehouse is through visualization. Visualization can allow one to see patterns quickly that are not apparent from just reviewing the numbers. As an example, I was analyzing survey data about the desirability and likelihood of various futures related to medical informatics. One question related to the future scenario of there being no print library, and the mean scores for both measures were slightly above average. To visualize these results, I started by plotting them in two dimensions, which allowed me to see that they were spread across the likelihood/desirability spectrum. Color-coding the dots brought out a pattern: faculty members felt than an all-electronic library was desirable and likely, while librarians felt that this outcome was neither desirable nor likely. Visualizing these responses turned an "above-average" score with little meaning into a rich finding that could then serve as inspiration for follow-up exploration.

Bibliometricians have been using visualization since the mid-1990's in order to explore the relationships between people, works, subjects, and other metadata (White & McCain, 1997). Many bibliometric and informetric methods focus on the connection between two works due to a citation or author, and over a large enough data set these connections allow for the visualization of clusters in these virtual spaces (Börner, Chen, and Boyack, 2003). The bibliomining data warehouse contains these links, so these visualizations are still possible; however, new types of visualizations can be explored when introducing links between works from users. The bibliomining data warehouse will contain the non-aggregated data needed in order to make these types of visualizations a reality.

4.4 Data Mining

The goal of data mining is to explore the dataset for patterns that are novel and useful. Data mining can be directed, where there is a particular goal or topic area in mind, or undirected, where the goal is to find something interesting. Data mining includes several of the techniques already discussed as well as other tools from statistics and artificial intelligence such as neural networks, regression, clustering, rule generation, and classification. Data mining is the process of taking a cleaned data set, generating new variables from existing ones (such as creating a yes/no flag from a numeric variable), splitting the data into model building sets and test sets, applying techniques to the model building sets to discover patterns, using the test sets to ensure the patterns are more generalizable, and then confirming these patterns with someone who knows the domain (Berry & Linoff, 2004). These patterns are then seeds for more thorough explorations, which may result in new visualizations, new reports, and even new aggregate measures which can become part of the regularly collected measures.

In order to perform data mining, non-aggregated data is needed. Even if a library does not have the tools or expertise to mine their own data, they can still empower researchers to explore the data if the maintain a non-aggregated version of data from their systems. By using the demographic surrogate concept, the library will protect the identity of their patrons while still enabling others to find patterns attached to the demographics. While some information is lost in the replacement of individual patrons with demographic surrogates, it is hypothesized that patterns related to the demographic groups can still be discovered. These demographic-based patterns are the patterns useful in allowing librarians the ability to better customize services for user groups. The involvement of the library is important; in order to create patterns that are useful, the librarians must be involved in designing the taxonomy for the demographic surrogates.

One subarea of data mining that would be useful in digital library evaluation is Web usage mining. This branch of data mining starts with the transaction log from a Web server, which allows the digital library service to track what is being selected by patrons. The challenge in transaction log mining is the data cleaning, as transaction logs are dirty and very challenging to work with. This concept was explored by Srivasta et al. (2000) and was applied directly to digital libraries by Bollen et al. (2003).

Text mining is an area related to bibliomining, but is not directly supported by the bibliomining data warehouse. In text mining, the context of the works is explored for patterns. The bibliomining data warehouse, as defined here, would not support text mining as the works are represented by bibliographic surrogates. Someone doing a text mining project, however, would be able to take their textual data and append fields from the bibliomining data warehouse. This could greatly enhance text mining projects, as it would tap into information contained in the works, metadata about the works, and information about the users of the work.

Kostoff et al. (2001) found that combining text mining and bibliometrics allowed for a much greater understanding of the author community than either method used separately. They started with four papers and extracted all of the citing papers. They then used text mining to look for thematic clusters in the text of this body of works. One key finding was the works that cited the original pieces but were from a different discipline could be removed from the body of works from the appropriate discipline. The disadvantage of this process was the amount of manual processing time required to

analyze only four original works and their first-generation citations. Regardless, this work demonstrates the potential of combining concepts from bibliometrics and data mining to greatly improve understanding of a situation.

4.5 Bringing Together Researchers and Practitioners through the Open Efforts project

One unanswered question is how a library can perform some of these complex analyses without the time, tools, or expertise. Researchers have the techniques, but do not have the data needed for analysis. The bibliomining data warehouse can serve as a point of virtual collaboration between researcher and practitioners. The concept is that, through collaborative groups, researchers and practitioners will develop a standard or schema for the data warehouse. The data will be taken from the library systems, de-identified, and passed to the data warehouse. Researchers can then access the data warehouse and view the data from their varied perspectives. As researchers develop models, measures and tools for examining the data, they will place those tools into a management information system attached to the data warehouse. The practitioners can then access this management information system and use the tools to explore the data that originally came from their library. The result is that researchers get the data from multiple services that they need to create generalized results and practitioners can see different views of only the data from their library.

Inspired by the Open Source movement in software design, Nicholson and Lankes (2005) have termed this the "Open Efforts" project. They are developing this project for digital reference, calling it the Digital Reference Electronic Warehouse (DREW). The goal of DREW is to develop an XML schema to allow digital reference transactions from different services and in different communication forms to live together in one space, to allow researchers to access these archives and explore them using a variety of methods, capture the results of this research into a management information system, and then allow the reference services to view their own archives through the tools created by the researchers.

The DREW project also has the potential to be a true "bibliomining" data warehouse. There are several ways in which reference transactions can be linked through bibliometric concepts. Many digital reference services allow those answering questions to refer to other transactions in their knowledge base, which is similar to the role a citation plays in authorship. In addition, many digital reference transactions contain links and citations to other works. If two transactions each cite the same third item, then these two transactions have something in common. This is akin to bibliographic coupling from bibliometrics (White and McCain, 1989), and therefore similar methods of analysis and visualization used to explore bibliographic coupling could be applied to DREW. It is expected that other bibliometric techniques will be useful in exploring digital reference as scholarly communication.

The potential for the Open Efforts project is significant. Researchers and practitioners have had a difficult time working together in the past. One of the primary reasons that the goal of practitioners is to understand their own library services and the goal of many researchers is to create generalizable statements. This results in a challenge between developing a study high in internal validity to benefit the library and high in external validity to benefit the researcher (McClure, 1989). The data-based collaboration proposed here allows both sides to get what they need from the research. In addition, where there is data available, researchers normally not interested in a discipline might be enticed to explore a new domain; it is expected that creating these data warehouses will therefore greatly increase the number of researchers exploring library data.

5. Placing Bibliomining in Context

Many data mining companies give the impression that these pattern-seeking tools will reveal everything needed to know to make smart decisions. This is not true in the corporate environment, and is certainly not true for bibliomining. A data warehouse will contain large amounts of information about the digital library, but it must be supplemented with other types of information to lead to a holistic understanding of the service. The process of users seeking and services providing information is complex, and to understand it properly requires multiple perspectives and measures.

In order to aid library decision-makers and library/information science researchers in understanding how bibliomining fits into a larger understanding of resources and users, two conceptual frameworks have been developed. The need for two frameworks comes from the fact that library managers and researchers have different goals when it comes to understanding digital libraries. Library decision-makers are more focused on an understanding of their own digital library, while researchers want to look to generalize their findings about one library to a larger population.

Bibliomining can help both of these groups, but in different ways. For decision-makers, it can aid them in a more in-depth understanding through details of the utilization of their services. For researchers, creation of data warehouses for multiple libraries can help them explore a larger breadth of institutions with the same tools previously used with only one library. Since decision-makers and library science researchers benefit from bibliomining in different ways, there are two different conceptual frameworks to support their work.

5.1. Conceptual Framework for Decision-Makers

Bibliomining is useful in understanding one portion of digital library services, but library decision-makers must realize that there are more perspectives and topics not available through the bibliomining data warehouse needed to understand the library from a holistic view. The bibliomining data warehouse does not capture how a collection meets standards or matches the mission of the library. These things require data to be collected about the library system in context, and bibliomining only captures information about the manipulation of the system. The bibliomining data warehouse also doesn't capture information about the actual use of the digital library resources. It only records what a user selected, but decision-makers can not tell from that if the item was relevant to the user's query. In order to learn about relevance, the user must be consulted, which will not be part of the bibliomining data warehouse. In addition, Saracevic and Kantor (1997) presented the idea that there is a difference between something being relevant and something being valuable in a framework for library evaluation. This perspective is also not present in the bibliomining data warehouse.

Therefore, bibliomining provides information only about how users accessed the existing system. It does not provide information about how relevant the information was for the user, nor does it aid in understanding where the system did not meet the needs of the user, either through not having the appropriate information or having a difficult system to use. Bibliomining is a way to measure the internal (librarian/researcher) view of the use of that system. Bibliometrics can extend this somewhat, if one assumes that a citation indicates value to that author.

A conceptual framework, fully developed in Nicholson (2004b) and summarized here, places bibliomining in context with other forms of measurement needed to collect information from these other topics and perspectives and demonstrates how that measurement framework is then useful in cumulative evaluation. Bibliomining is one tool useful in the measurement aspect of this relationship. It can be used to aid in the quantification of aspects of the service. However, there needs to be a person involved, either directly or embodied through criteria, in order to evaluate the service. Therefore, bibliomining is useful in measurement but is not sufficient for evaluation.

Table 1 presents the matrix for holistic measurement for library services (based on Nicholson, 2004b) that summarize the different areas of measurement. The two variables in this measurement framework are the perspective of measurement and the topic of measurement. The perspective of measurement differentiates between who is asked to create the measurements: the internal, or system-based, perspective comes from the librarian or researcher making observations about the digital library; and the external, or user-based, perspective comes from asking the user about aspects of the digital library. The topic dimension is divided between measurements taken about the content of the library system and the manipulation of that system. This differentiates questions about what is contained within the digital library from questions about the access and utilization of those materials.

Table 1: Holistic Measurement Matrix for Library Services

	Topic	
Perspective	**Content of Library System**	**Use of Library System**
Internal (Librarians, Library and Information Scientists)	Measurements based on policies, procedures, standards, simulated users (**Bibliomining using library content and metadata**)	Measurements based on data-based artifacts representing interactions with library interface and materials (**Bibliomining using usage data and other artifacts of user behavior**)
External (User)	Measurements based on relevance and aboutness of results, usability of system	Measurements based on value of system, knowledge states of users, and user citations to materials (**Bibliomining using bibliometric data and other artifacts of creation**)

Bibliomining can support three of the cells in this matrix. Focusing on the library content and the metadata surrounding artifacts of behavior provides information about the Internal view of the Use of the Library System itself. Introducing citations made by library users in scholarly works (through bibliometrics) or other outlets (such as blogs, Web publications, or through paper-collection services like TurnItIn.com) can give the library some indication of what

was valuable. However, this External View of the Use of the Library System is only one facet in the many-faceted challenge of relevance and value.

Library managers and evaluators must also take other measurements from all quadrants of this matrix as the first step in gaining a holistic understanding of their library. After measurements are collected from these different perspectives and on these different topics, then those measurements are evaluated by users, librarians, and decision-makers in order to create a holistic understanding of the digital library services.

5.2. Conceptual Framework for Library and Information Scientists

As compared to library decision-makers who need to focus on a particular library system and would most benefit from a holistic knowledge of that particular system, library scientists are looking to increase the understanding of librarianship on a broader scale by working toward generalizable statements. In order to understand the context for bibliomining in this research, we turn to archeology. Archeology is an appropriate model because the process archeologists follow of collecting artifacts to make statements about people who lived in an area is akin to the process used to create the bibliomining data warehouse. When users wander through a digital library, they leave behind data-based artifacts of their visit. These artifacts may be a record in a Web log, a proxy server login, or a query used in a digital library search. During the data warehousing phase, these artifacts are collected from different places associated with the library system and reassembled in order to gain a more complete understanding of the users who "lived' in that digital space. This concept is developed as "Internet Archeology" by Nicholson (2005).

In many library studies, these artifacts of use are collected, cleaned, counted, and put on display through conference presentations and articles with a nod toward how these statistics compare to those of other libraries. This is similar to archeology up until the 1960's: researchers went to a site, gathered artifacts, displayed them to others and then attempted to make connections between the artifacts from this and other expeditions. However, in the last 50 years, archeologists have extended their conceptual framework for research through "new" archeology and post-processual archeology. New archeology is focused on increasing the knowledge base instead of simply collecting more artifacts in dig after dig (Johnson, 1999). This can be related to the current state of digital library evaluation; rather than asking specific questions that might advance knowledge of digital library users, many researchers go on one virtual "dig" after another, collecting more and more measures without building toward knowledge (Nicholson, 2005). Post-processual archeology brought in the importance of realizing that there was more to the situation than could be told by artifacts; issues like social context and community influences needed to be considered (Johnson, 1999).

The resulting framework of research is more complex but can result in the creation of more generalizable statements of knowledge. Archeologists start by examining readily available artifacts for patterns. They create generalizations from these patterns and develop hypotheses. Studies are then created to test each hypothesis, and the archeologists revisit the site with a sampling methodology, gather new artifacts, talk with people in the area, and test the hypothesis. Finally, these hypotheses are tested in other settings to further scientific knowledge. This pathway is known as the Hypothetico-Deductive-Inductive method (South, 1977).

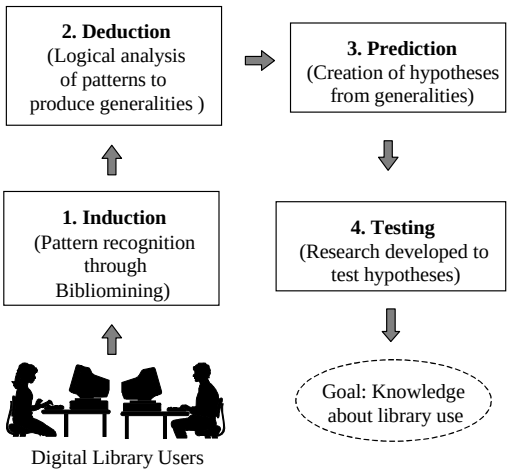

Figure 6: Archeology-inspired Framework for Digital Library research

Figure 6 shows an application of this framework to digital library research and places data mining in context with the rest of the research process. Just as in the previous framework, bibliomining is an important step in the process, but it is only one step in a more complete process. First, the artifacts of digital library use are collected, and bibliomining is used to determine patterns. These patterns can then lead to generalities, which can inspire hypotheses. Then, studies are designed to test these hypotheses; these studies may be quantitative or qualitative (or both), and may require talking to digital library users or collecting new data. The results from these studies can then lead researchers down the path of new knowledge.

5.3. Understanding both Frameworks

In both of these frameworks, bibliomining is not the end of the exploration process. It is one tool to be used in combination with other methods of measurement and evaluation, such as LIBQUAL, E-metrics, cost-benefit analyses, surveys, focus groups, or other qualitative explorations. Using only bibliomining to understand a digital library can result in biased or incomplete results. While the information provided by bibliomining is useful, it needs to be supplemented by more user-based approaches to provide a more complete picture of the library system. These frameworks help library decision-makers and library researchers in putting bibliomining in its proper context.

Understanding these frameworks is critical when library decision-makers and researchers work together on a bibliomining project. As each party is interested in different results from the bibliomining process, communication is needed throughout the project in order to ensure the needs of both are met. The Open Efforts project proposed above will provide the data for each group to begin their explorations; however, the differing needs of practitioners and researchers may lead each in a different direction after exploring the bibliomining data warehouse. The hope is that tools created by each for exploration can be made available so all can benefit from new ways of thinking about the data.

6. A Research Agenda to Advance Bibliomining

There have been a number of researchers working on bibliomining projects over the last few years; however, there has not been a consolidated effort to move the field forward (Nicholson, 2004a). This has resulted in different projects doing the same preliminary tasks, sometimes in very different ways. The goal of this research agenda is to identify problems that are common to different bibliomining projects. As these common problems are systematically addressed through research, new projects will have a stronger research framework from which to begin.

6.1. Data Collection

The first step in any bibliomining project is collecting the data. Even this represents a challenge to librarians with a multitude of systems that support digital library services. Zucca documented this process for one academic library, and had many struggles in connecting Web log data to user information through a proxy server connection (2003). The success of this project demonstrates that it is possible; however, the challenges faced suggest that systematic research is needed in order to allow others to begin the bibliomining process more effectively.

There are typical configurations of systems that support digital library services. For example, for services associated with a traditional library, there may be: an Integrated Library System (ILS), which contains metadata about works in the library, circulation data, user demographics, and acquisition information; the Web-based front-end to the digital library that serves as the hub between different types of digital library services and resources which may also include some type of authentication system; a system to support interlibrary loan; and a system to support digital reference services. In addition, external systems such as citation databases or census data might hold data needed for the bibliomining process.

Currently, a library decision-maker or researcher wishing to engage in bibliomining must determine how to collect this data and match it between systems. This is a time-consuming and frustrating process and may certainly stop bibliomining projects in the early stages due to the cost and time involved. There are several lines of research that could address these problems.

The first step is for current projects to document their collection and matching process. There are only a few major vendor choices for each system used by library services, and, therefore, each new project can inform us as to the difficulties of that particular connection.

As we examine the ways different systems connect, this can inspire research to develop standards for data. There are several standards in development that will aid in the data collection and measurement problems – Project Counter, which is a standard designed for vendors to reporting usage of online services (Project Counter, n.d.), and NISO Z39.7-200x, which consists of metrics and statistics for libraries (NISO, 2004). The problem with both of these standards is that they focus on aggregate-level data and not the underlying records. For the bibliomining data warehouse to be successful, non-

aggregated data must be kept. These aggregate standards are useful in that they help libraries to adopt a data-driven mindset and therefore can be used as the starting points from which aggregate-level data standards can emerge.

After standards for data fields are proposed, the next step is to integrate these into the digital library systems. This will involve cooperation with the system creators. One valuable result would be an easily exportable data warehouse built in to the library systems. This warehouse could be matched with other warehouses in the library through common fields. This would allow the digital library service to choose different systems for each aspect of their offering and then combine the data from them into one bibliomining data resource. This concept of the digital library-wide data warehouse raises concerns about user privacy.

6.2 User Privacy

As presented earlier, the bibliomining data warehouse can provide the method for keeping information about the materials used in the library without maintaining specific information about the users of the library. There are some research questions posed by this issue. For example, one research question this raises is the effect this anonymization has on the power of the data mining tools to discover patterns. A typical method of analyzing this type of use data is to look at how a specific user accesses resources over time; however, that is not directly possible with this data warehouse. The interesting question is to see if similar patterns could be teased out using the more general demographic variables and how traditional data mining algorithms can be adjusted to function appropriately in this environment. There is an active line of related research regarding privacy-protecting data mining (Limpaa, 2003) that would be an appropriate home for this line of research.

Another question about data privacy comes from digital reference transactions. While it is easy to replace personal fields with demographic variables, the challenge comes in the personal information that users can embed in the question. Many users will include information in the text of the question that does not belong in an anonymized data warehouse. Text mining and natural language processing research are needed to develop algorithms to clean these records. This de-identification is similar to a problem faced in medical informatics: in order to appropriately use medical records in research, the personal information must be removed from the facts of the case. This is usually done manually, although researchers are working toward methods of automatic de-identification of medical records. As these methods are developed in medical informatics, we will apply them to the de-identification of library records.

6.3. Variable, Metric, and Model Generation

While researchers have developed metrics for library statistics (Project Counter, n.d.; McClure, Lankes, Gross, & Choltco-Devlin, 2002), they have primarily focused on fields from one data source. Once the bibliomining warehouse has been constructed, the possibilities grow for the discovery interesting variables for mining and metrics for evaluation, just as Kostoff et. al. (2001) discovered. This line of research would start in the data mining process, looking for relationships between individual variables that allow for deeper understanding. Through the patterns discovered with data mining, new metrics and measures can be proposed. For example, one challenge in using Interlibrary Loan data for collection development is that of the “super user”, i.e. heavy use of the service for a single project. If a model could be developed to predict and automatically identify super-users from their behavior, then what is left from ILL data may be more useful in collection development. These models could then be applied to e-resource use to separate out one-time high-demand needs from needs that represent the general user base. As multiple digital libraries employ these measures, it will allow researchers to compare and contrast different services and advance the understanding of the field.

6.4. Integration of Management Information System and Data Mining Tools

As researchers discover algorithms that are useful for bibliomining applications, a fruitful next step is to integrate these algorithms into the systems that drive digital libraries. The approach taken by the Information Institute of Syracuse for the Digital Reference Electronic Warehouse (DREW) is to attach a data warehousing component to the QABuilder digital reference system (Information Institute of Syracuse, n.d.). Then, different reporting, exploration, and mining tools are integrated into the digital reference system so that any library implementing the QABuilder system also has a data warehouse and a management information system. In addition, the manager has the ability to explore the data through OLAP-like features and export the specifications for any view of the data through XML to other services using the same system. Additionally, digital reference researchers can send out an XML specification for a report and use this to collect the same data from multiple systems for a research project. Finally, the system is modular, so that new analysis tools and data mining algorithms can be plugged into the data warehouse for expandability.

This combination of a built-in data warehouse, interactive reporting module, standards for report description, and modular design will make it much easier for library decision-makers to get involved with bibliomining. Therefore, a

fruitful line of research is to work toward developing these integrated modules for other systems that support digital libraries.

6.5 Multi-System Data Warehouses and Knowledge Bases

The final area of this research agenda is the creation of services that span many digital libraries. If the earlier stages in this research agenda are completed, the moving from data warehouses that span the services of a single library to data warehouse that combine data from many libraries is a formidable yet achievable goal. One area where this type of multi-system data warehouse would be useful is for library consortia. These groups of libraries join up to negotiate for a better rate on electronic journals, but currently are at a competitive disadvantage to the producers of these materials. Through the capacity to understand how resources are being used and what types of users are involved, these consortial groups can make much better decisions.

Another type of multi-library project where this concept is used is those that bring together different digital collections through a common front-end, such as the National Science Digital Library (NSDL) project. The goal of the NSDL project is to bring together digital library collections through one infrastructure. This project is in the building phase for integration, and it is proving to be a fruitful area for research. This concept of joining together digital library sources and services while still maintaining identity for those participating is in the exploratory phases.

The advantage of multi-library data warehouses addresses one of the problems with these methods. Data mining requires large amounts of data. Large corporations with hundreds of thousands of users and comparatively few items can gain considerable information from user-based data mining. Libraries have a higher item/user ratio; however, the power law from bibliometrics suggests that relatively few items from the collection will receive most of the use.

One solution to collecting more data with the hope of discovering patterns in the noise is to join data warehouses with libraries that have similar user groups and similar collections. If the libraries agree upon the demographic surrogates or develop a crosswalk algorithm that will map one demographic to another, they may be able to discover patterns out of the larger data sets that did not appear before. If there is a core set of items used by different demographic groups, then with combined data warehouses, the patterns will be easier to discover in the localized noisy data. After discovering patterns, librarians will need to ensure that these patterns apply to their own library before making decisions based upon them.

There are two methods for combining utilization and collection metadata between different systems. The first is to standardize a series of metrics; however, different people measure metrics in different ways using differing underlying assumptions. This has been a problem in the measurement and evaluation of the NSDL; different services count a “visit” in different ways resulting in data that is not easily comparable or combinable across services. The bibliomining approach to this problem would be to create a single warehouse for all services and capture information from the Web server logs from different services. A standard would need to be created for the form that the logs take, but once all of the logs live in the same data space, then it would be easier to collect measures the same way across all services. In addition, this would allow the NSDL to better understand how the collection of services is working as a group instead of many individual services.

Another method is to create a standard for record-level data, where one access of a research work is one record in the dataset. Then, this low-level data is collected from different services into a multi-service data warehouse. This data warehouse then is much more useful than just the aggregates; patterns from across different services can be discovered to advance our knowledge of digital libraries. This last method is the approach taken by the DREW project: create an archival schema in order to develop a data warehouse of anonymized digital reference transactions from different library services to benefit library and information scientists.

This second approach is one that could also grow out of current standards such as MARC and COUNTER. As the library profession develops and accepts these standards, the vendors of library systems also integrate them. A standard similar to COUNTER but for item-level usage could be incorporated into library service vendor’s systems, thus allowing the library to run a program that generates a data warehouse on a regular basis. A similar concept could be built into a content management system, where key information from Web server logs is extracted in and placed into a data warehouse on a regular basis. These warehouses, assuming that they follow a standard, could then be easily combined from different libraries. In addition, it would set the stage for the Open Efforts project discussed earlier, where researchers can then create models, measures, and tools based on the standards for data. These tools could then be either integrated directly into the systems provided by library vendors or live in a separate tool that connects into the library data warehouse.

7. Conclusion: Moving beyond Evaluation to Understanding

The final and most long-lasting area of research that bibliomining can inspire is improving understanding of digital libraries at a generalized, and perhaps even conceptual, level. These data warehouses will combine resources traditionally

unavailable in this combined form to researchers. As in the archeology framework presented earlier, this is the point where bibliomining can inspire new research questions. For example, what connections can be made between patron demographics, and bibliometric-based social networks of authors? How much influence do the works written and cited by faculty at an institution have on the patterns of student use of library services? How do usage patterns differ between departments or demographic groups, and what can the library do to better personalize and enhance existing services? These might involve qualitative methods to support the quantitative data, or may include gathering other types of data from the other quadrants of library measurement to enhance the bibliomining data. Exploring these new research questions may lead to insights through this science of librarianship.

In addition, these large multi-system data warehouses can allow visualizations of a knowledge space through understanding connections between the resources used or commonalities in reference transactions. Taking methods currently used for the visualization of bibliometric data, expanding them to include connections between works, and adding animation to demonstrate how this data changes over time can allow the understanding of how knowledge spaces evolve.

Due to the artifact-based nature of digital library services, library decision-makers and researchers have the ability to understand the information-seeking behavior of patrons previously available only in a tightly-controlled research environment. These data-based artifacts can be collected, anonymized, and combined from different systems and services into data warehouses that provide decision-makers with the depth and researchers with the breadth of information needed to better understand digital libraries. Using bibliomining as a core part of conceptual frameworks allows researchers and decision-makers to work together in developing research questions with both short-term interest and long-term impact.

8. References

American Civil Liberties Union. (n.d.). *USA PATRIOT act*. Retrieved May 18, 2004 from http://www.aclu.org/SafeandFree/SafeandFree.cfm?ID=12126&c=207

Barabási, A. (2003). *Linked*. New York: Penguin Group.

Banerjee, K. (1998). Is data mining right for your library? *Computers in Libraries, 18*(10), 28-31.

Berry, J., and Linoff, G. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (2nd edition)*. Indianapolis, IN: Wiley.

Bollen, J., Luce, R., Vemulapalli, S., and Xu, W. (2003). Usage analysis for the indentification of research trends in digital libraries. *D-Lib Magazine* 9(5). Available online at http://www.dlib.org/dlib/may03/bollen/05bollen.html.

Borgman, C. L. & Furner, J. (2002). Scholarly communication and bibliometrics. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 36, pp. 3-72). Medford, NJ: Information Today.

Borgman, C. L. (Ed.) (1990). *Scholarly Communication and Bibliometrics*, Newbury Park, CA: Sage Publications, Inc.

Börner, K., Chen, C.M., & Boyack, K.W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology 37*, 179-255.

Buckland, M. (2003). Five grand challenges for library research. *Library Trends, 51*(4), 675-686.

Chaudhuri, S. and Dayal, U. (1997). An overview of data warehousing and OLAP technology," *SIGMOD Record 26*(1), 65-74.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management 35*(3),401-420.

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

Eirinaki, M. & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Transactions on Internet Technology 3*(1), 1-27.

Geyer-Schulz,A. Hahsler, M., Neumann, A., and Thede, A. (2003). An integration strategy for distributed recommender services in legacy library systems. In M. Schader, W. Gaul, and M. Vichi, editors, *Between Data Science and Applied Data Analysis, Proceedings*

of the 26th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Mannheim, July 22-24, 2002, Studies in Classification, Data Analysis, and Knowledge Organization, pages 412-420. Springer-Verlag, July 2003.

Information Institute of Syracuse (2004). *QABuilder.* Retrieved May 21, 2004 from http://vrd.org/qabuilder.shtml

Johnson, M. (1999). *Archeological Theory: An Introduction.* Oxford: Blackwell.

Kao, S. Chang, H., and Lin, C. (2003). Decision support for the academic library acquisition budget allocation via circulation database mining. *Information Processing &Management 39*(1), 133-148.

Kostoff, R., del Río, J., Humenik, J., García, E., and Ramírez, A. (2001). Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling. *Journal of the American Society for Information Science and Technology 52*(13), 1148-1156.

Limpaa, H. (2003). *Privacy-preserving Data Mining*. Retrieved May 18, 2004 from http://www.tcs.hut.fi/~helger/crypto/link/data_mining

McClure, C., Lankes, R. D., Gross, M., & Choltco-Devlin, B. (2002). *Statistics, Measures, and Quality Standards for Assessing Digital Library Services: Guidelines and Procedures.* Syracuse, NY: ERIC Clearinghouse on Information & Technology. Retrieved May 21, 2004 from http://quartz.syr.edu/quality/

McClure, C. (1989). Increasing the usefulness of research for library managers: Propositions, issues, and strategies. *Library Trends, 38*(2), 280-294.

Michail, A. (1999). Data mining library reuse patterns in user-selected applications. *14th IEEE International Conference on Automated Software Engineering*, Washington, DC: IEEE Computer Society, 24-33.

Murphy, D.E. (2003, April 7). Some librarians use shredder to show opposition to new F.B.I. powers. *New York Times*, pp. A12.

Nicholson, S. & Stanton, J. (2003). Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. In H. Nemati and C. Barko (Eds.), *Organizational data mining: Leveraging enterprise data resources for optimal performance* (pp.247-262). Hershey, PA: Idea Group Publishing, 2003.

Nicholson, S. (2003). The bibliomining process: Data warehousing and data mining for library decision-making. *Information Technology and Libraries, 22* (4), 146-151.

Nicholson, S. (2004a). Bibliomining bibliography. *The Bibliomining Information Center*. Retrieved March 1, 2004, from http://bibliomining.org

Nicholson, S. (2004b). A conceptual framework for the holistic measurement and cumulative evaluation of library services. *Journal of Documentation 60*(2), 162-182.

Nicholson, S. (2005). A framework for Internet archeology: Discovering use patterns in digital library and Web–based information resources. *First Monday 10*(2). Retrieved March 23, 2005 from http://www.firstmonday.org/issues/issue10_2/nicholson/index.html

NISO. (2004). *Z39.7 library statistics – E-metrics data dictionary*. Retrieved May 18, 2004 from http://www.niso.org/emetrics/
Project Counter. (n.d.) *Counter – Counting Online Usage of Networked Electronic Resources*, Retrieved May 18, 2004 from http://www.projectcounter.org

Sandstrom, P.E. (2001). Scholarly communication as a socioecological system. *Scientometrics 51*(3), 573-605.

Saracevic, T. & Kantor, P. (1997). Studying the value of library and information services: Part 1, Establishing a theoretical framework. *Journal of the American Society for Information Science*. 48(6), 527-542.

Srivastava, J., Cooley, R., Deshpande, M., Tan, P.T. (2000). Web usage mining: Discovery and applications of usage patterns from Web data. *SIGKDD Explorations(1)*2. 12-23.

South, S. (1977). Method and Theory in Historical Archeology. New York: Academic Press.

White, H.D. & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science 1972-1995. *Journal of the American Society of Information Science, 49*(4), 327-355.

White, H.D. & McCain, K.W. (1989). Bibliometrics. In M.E. Williams (Ed.) *Annual Review of Information Science and Technology 24*. Medford, NJ: Information Today. 99-168.

Wilkinson, D., Thelwall, M., & Li, X. (2003). Exploiting hyperlinks to study academic Web use. *Social Science Computer Review, 21*(3), 340-351.

Zucca, J. (2003). Traces in the clickstream: Early work on a management information repository at the University of Pennsylvania. *Information Technology and Libraries, 22*(4), 175-179.