Preprint - Nicholson, S. (in press; expected Fall 2006). Approaching librarianship from the data: Using Bibliomining for evidence-based librarianship. *Library Hi-Tech* 24(3).

Approaching Librarianship from the Data: Using Bibliomining for Evidence-Based Librarianship

Scott Nicholson, Syracuse University School of Information Studies, srnichol@syr.edu

Abstract

Given the current lack of systematic studies in librarianship, it can be difficult to do proper traditional Evidence-Based Librarianship   This article starts by deconstructing Evidence-Based Librarianship down to the representations of the users involved. The bibliomining process, or the combination of data warehousing, data mining, and bibliometrics, is used as a framework to build a different path to EBL. Bibliomining-based Evidence-Based Librarianship is not appropriate for all topics; however, when the artifacts of library use can be gathered and explored, this method can provide a different path to reach the goals of EBL.

Introduction

Traditional Evidence-Based Librarianship starts with prior published systematic explorations of library phenomena. Ideally, these studies are controlled randomized trials that demonstrate the effectiveness of a library service or resource on a group of users. This concept comes from Evidence-Based Medicine, where studies focused on the same intervention can be gathered from explorations around the world. These studies are then examined and combined in order to create evidence. This evidence is then used in combination with local evidence for decision-making (Sackett et al, 1996).

There are, however, some difficulties in applying this model to the science of librarianship. The most significant one is that there are not many randomized controlled trials that demonstrate the change brought about by a library service. Some definitions of EBM allow the use of the best available evidence, regardless of type (Sackett et al, 1996). This more reasonable definition is one commonly applied to allow Evidence-Based Librarianship (EBL); however, there are few types of library services that have the quantity of published research projects that allow the combination of results to produce meaningful and useful evidence.

Over time and with guidance from the EBL community, better evidence can be produced. The EBL community can continue to produce guidelines for library researchers as to how to produce research appropriate for the EBL approach. Through shared resources such as the new EBL journal, *Evidence-Based Librarianship and Information Practice*, and professional organizations, the EBL community will continue to move forward in guiding and collecting appropriate research useful in traditional EBL. At this time, however, the body of research needed for reliable EBL is not strong.

The purpose of this work is to present a different path to Evidence-Based Librarianship. Instead of starting from research published by others, the proposal here is seated in bibliomining, or the combination of data mining, data warehousing, and bibliometrics for the measurement and evaluation of library services. Through standards for data collection, libraries can work together to build shared data warehouses to power decision-making. While the tools used will be different from traditional EBL, the results can be more powerful and flexible while still being based in EBL concepts.

Deconstructing Evidence-Based Librarianship

Before presenting the bibliomining model of EBL, I will first deconstruct traditional Evidence-Based Librarianship. The purpose of this is to conceptualize EBL down to the underlying data, and then rebuild it using a bibliomining frame of reference.

Evidence-Based Librarianship starts with a topic. This topic will most likely be a library service or collection, but might also be a specific user community or library setting. After identifying the topic, the librarian identifies local evidence and local restrictions related to that topic. This information guides the choices in the next step, which is finding other research on the same topic. Starting with an awareness of the local setting allows the librarian to search more effectively for research that might be applicable in the situation.

After determining the topic and restrictions on the types of research to collect, the librarian searches for related studies. The “gold standard” for research is randomized controlled trials that show the impact of a library service, but these are rare in library research. Any generalizable research that has been done properly is acceptable as some level of evidence. Studies employing controls to reduce the effect of local bias are more desirable than studies focused on the activities of only one specific library.

After research has been collected, it is analyzed. Ideally, the librarian combines the results of a number of studies that have been performed in similar fashion in order to produce evidence. This is much easier in the medical setting, where the intervention (a treatment or drug of some type) can be standardized. It is more challenging in a library setting, where, due to their nature, libraries provide “treatments” that are relevant to a local population. The treatment given in one library to resolve an information need may be different than the treatment given in another library due to differences in population, available resources, and local policies. This provides a challenge in trying to combine research from different studies, each of which discusses a local library service that was relevant to a different population and based on a different local collection.

The results of these studies are aggregates of measures; occasionally, raw data are available, but usually only the aggregates can be accessed. Each measure reflects some aspect of a library user, collection or service. If these measures are not collected in the same way, then it is problematic to combine them. Only in cases with standardized measures, such as LIBQUAL or libraries using the same version of an automation system, can these measures be safely combined from different studies. Any other time, combing measures from different studies is a challenge that can lead to misleading or incorrect results.

User Surrogates

These aggregate measures represent, in many cases, some aspect of a group of users. When EBL techniques are applied, the advantage is that the librarian accesses these “user surrogates” from different library settings. These surrogates have usually been aggregated in some way, so those doing EBL cannot see the underlying surrogates; they can access only the aggregations. If these aggregations are created using different statistical methods, combining them may result in useless results.

In addition, these aggregates hide underlying subgroups and patterns. For example, if half of the people surveyed loved a service and half hated a service, then a mean or median of these data would result in the service receiving an average rating. If the study provides only these aggregates, those trying to apply those aggregates to another setting will have data that is not truly representative of the end-users. This could be misleading.

The problem is compounded, as mentioned above, when these aggregates are combined with other aggregates, which may or may not have been created using the same statistical method. Finding patterns within the user groups or identifying underlying similarities in types of users becomes much more difficult, if not impossible. The evidence used in traditional EBL consists of these aggregates as reported in research studies.

Ideally, those writing these research study reports will also make the underlying data available. In reality, this does not happen often. Some researchers are possessive of the data they spent the time and effort to collect and clean. Others are not confident enough in their own explorations and fear others disproving their findings. Privacy issues may prevent the data from being released. At the end of the writing process, many researchers have moved on to other projects and do not have the time to clean, document, and release an older dataset.

Thus, at the end of this deconstruction, the resources available for EBL are aggregates as reported in research studies. There also may be raw datasets available. These aggregates and datasets, when created using different techniques, can be difficult or impossible to combine to create the evidence needed for Evidence-Based Librarianship. In reality, there are relatively few published studies for a particular library phenomenon; this lack of evidence is currently the most problematic aspect of applying traditional EBL methods.

Instead of starting with the research reports, the proposal made in this article is to start with the data. The concepts behind Evidence-Based Librarianship are useful, but the current state of library research makes it challenging to put together trustworthy evidence. What if librarians could start with underlying data collected in similar ways and aggregate the data as appropriate for their own needs and settings? For many library services, it is possible to create this type of data warehouse through the bibliomining process.

The Bibliomining Process

Bibliomining is the combination of data warehousing, data mining, and bibliometrics used to understand library services. The bibliomining process starts with the collection of data from different sources such as the library's automation system, patron demographic sources, Web server logs, and interlibrary loan systems. This data are collected into a data warehouse that provides a place for a copy of data from these various systems formatted in a way to facilitate analysis. The data from different sources are connected through shared fields such as Call Number or Patron ID, and after the match is made, any personally identifiable information is destroyed to protect the privacy of the patron. The advantage of this structure is that demographic and categorical variables representing a user can be associated with the library sources and services involved without infringing upon the privacy rights and expectations of the patron (Nicholson and Stanton, 2003). This data warehouse provides a resource for evidence upon which explorations and studies can be built.

In the bibliomining process, questions are posed about the users, services or resources. Based upon the question, subsets of the data warehouse are collected and data shortcomings are identified. This information guides the researcher as to what other evidence must be collected in order to explore this issue. The researcher gathers additional data and integrates it into the data warehouse. Techniques from statistics, data mining, and bibliometrics can be used to discover patterns in the data. These patterns then provide evidence for decision-making and many times offer new seeds for exploration (Nicholson and Stanton, 2003).

Moving beyond the scope of a single library, another developing concept associated with bibliomining is that groups of libraries can share their data warehouses, once they have been cleaned. This allows one library to examine the evidence collected by other libraries. These multi-system data warehouses can be invaluable in asking "what if" questions, as there may be other libraries that have implemented the options under consideration. Since libraries are regularly faced with decisions regarding new and different services, accessing usage data from libraries with similar populations can provide a library with the evidence needed to make good decisions.

Creating these multi-library data warehouses can be very challenging. If two libraries use the same automation system, the same Web server, and the same systems for other services such as ILL, then creating the data warehouse is an easier problem. Most libraries, however, have created their own patchwork of systems to support their services. Attempting to bring together the underlying data from different library systems can be difficult or impossible. A common method employed by others combining different datasets is to write some type of crosswalk program that converts data from one structure into a different structure. This type of crosswalk works best when there is some type of standard for transaction-level usage as everyone can work to put their own data into a standard format.

The Need for Transaction-Level Standards

Bibliomining exploration has been typically applied within the setting of a single library(for example, see Zucca, 2003). Standards such as MARC and Dublin Core provide a standard for describing library resources, but standards for item-level usage and the use of other library services either do not exist or are not widely implemented. Because of this, it is challenging to share data about this type of usage. The COUNTER project (COUNTER, 2005) provides standards for the usage of e-journals, but their standard is only at the monthly aggregate level. For bibliomining purposes as described here, data is needed at the individual transaction level.

Library systems take advantages of standards such as the MARC record. Cooperative cataloging, for example, changed the workflow of libraries and allowed many libraries to share resources. Services such as Interlibrary Loan are much more efficient when participating libraries allow searching using a shared protocol such as Z39.50. These advancements suggest that similar standards for library services might allow more powerful shared networks to be developed.

These standards do not currently exist. When a library creates its own data warehouse, there currently is little guidance and few standards. This makes it very unlikely that it will be easy for libraries to share their data warehouses. Even if the data covers the same type of library user or service, different (or no) standards for the data may make it impossible to connect the two data warehouses. Without the ability to share data, the ease of working on Evidence-Based Librarianship through the bibliomining process is diminished.

Digital Reference is an example of one type of library service moving toward standards for multi-library data warehouse. One goal of the Digital Reference Electronic Warehouse (DREW) project is to create a standard for digital reference transactions. There is currently no implemented standard for digital reference transactions, and while it would be useful to have a fielded database of reference transactions, it is a formidable task. The first goal of the DREW project is to create this standard in conjunction with major digital reference service providers in order to create a common schema for describing a transaction. This will then allow a shared data warehouse of digital reference transactions to be created (Nicholson and Lankes, In press). This DREW project, if successful, will serve as a model for other library services to create standards and shared data warehouses.

Using Shared Data Warehouses for Evidence-Based Librarianship

These multi-library data warehouses will contain surrogates for library services and use. The studies gathered for traditional EBL report aggregated surrogates from some type of sample of a population. Accessing a data warehouse, however, allows access to the non-aggregated surrogates. For a digital library service, if the collection is systematic, the entire population is available (as compared to only a sample from the population). The result is that, for certain types of questions, the bibliomining approach of building a shared data warehouse will provide richer and more complete data than what can be extracted from a set of studies.

There are several advantages to this approach over traditional EBL. When available through automatic capturing methods, the data points can be from a more complete set of users. It removes the bias introduced through aggregation in different studies; many times, authors select an aggregation method that will supply them with the best evidence for making their argument. Admittedly, any aggregation method introduces bias, but allowing the librarian to apply bias appropriate for the situation. Having access to the raw data allows the researcher to create the best aggregates for their own situation.

Most importantly, it allows the researcher to ask questions that connect different fields. For example, if a published study does not connect circulation figures to the days of the week, then that study is not useful to someone attempting to use traditional EBL to make better scheduling decisions. If that data is captured in the data warehouse, then as long as the day of the week is maintained, the researcher can make that connection. Another researcher may want to connect the same data to the month of circulation. Being able to explore the data through bibliomining tools allows the researcher to find evidence related to their need.

Another advantage to this approach is the potential speed of the analysis. If all of the data needed are collected, then appropriate evidence can be gathered from the data warehouse in a short period of time. Many times, library managers and administrators need to make a decision quickly, and a data warehouse will enable them to use some type of evidence in making that decision. Tools like Online Analytical Processing (OLAP), created for corporate managers to explore their company's data warehouse, give library managers a method of exploring their data without needing to know a database query language. Some library data warehousing projects, such as DREW (Nicholson and Lankes, in press) and the Normative Data Project (Molyneux, 2005), are building OLAP features into their interfaces.

There are a number of concerns with this approach. The most significant concern is the need to protect the privacy of the individuals through the collection of subsets of personal data. Another significant concern, discussed earlier, is the lack of standards to make it easy to connect different data warehouses. The bibliomining-based EBL is only appropriate for a subset of the decision-making needs of a library (Nicholson, 2004). If there is no record kept of the use of a library service, then this data warehousing approach is impossible. Many studies are based on surveys and interviews, and the sort of data collected through these approaches is usually not available in a data warehouse; therefore, the decision-making need may dictate the choice of method for Evidence-Based Librarianship.

Even in cases where the primary data source comes from other studies (such as survey or interview data), it may be that supplemental data can be extracted from the data warehouse in order to enhance the evidence. For example, while interviews will come only from a small sample of a population, the data warehouse can give the population-level view that is typically missing from qualitative studies. Librarians employing Evidence-Based Librarianship need to consider both previously published studies as well as data collected from the operation of library services in order to make strong decisions.

Future Paths for Evidence-Based Librarianship

The current shortcoming of traditional Evidence-Based Librarianship is the lack of appropriate research articles executed in similar ways that could be combined to make evidence for decision-making. The suggestions provided in this article develop a different form of EBL that is based on data currently gathered in libraries.

Why does this body of literature, strong in the medical field, not exist in librarianship? In the medical field, there are practicing doctors and medical researchers. Both groups contribute to the body of research; if doctors are too occupied doing the practice of medicine to write, research articles are still produced at an acceptable rate. In libraries, however, the ratio of librarians to library scientists/researchers is much different. While library scientists write articles about library services, many fewer librarians contribute significantly to the body of literature. Without a larger number of librarians contributing to the research, the body of research will not be significantly large enough to allow traditional EBL.

One challenge for practicing librarians is finding the time to collect the data. The bibliomining process assists librarians in doing research in a more timely fashion. This has the capacity to increase the number of research projects and the number of publications to power traditional EBL. The method of EBL proposed in this article, therefore, can not only power EBL in the short term but also can encourage the development of the body of literature to allow traditional EBL to blossom. In the long run, traditional EBL and bibliomining-based EBL will complement each other to allow library managers and administrators the evidence needed to make strong decisions.

Resources Cited

COUNTER. (2005). The COUNTER code of practice: Journals and Databases: Release 2. *COUNTER – Counting Online Usage of NeTworked Electronic Resources*. Retrieved August 14, 2005 from http://www.projectcounter.org/code_practice.html

Molyneux, B. (2005). NDP goes live. *SIRSI OneSource 1*(7). Retrieved August 14, 2005 from http://www.imakenews.com/sirsi/e_article000426388.cfm

Nicholson, S. (2004). A conceptual framework for the holistic measurement and cumulative evaluation of library services. *Journal of Documentation 60*(2). 164-182.

Nicholson, S., & Lankes, R. D. (In press). The Digital Reference Electronic Warehouse (DREW) project: Creating the infrastructure for digital reference research through a multi-disciplinary knowledge base. *Reference and User Services Quarterly 46*(2).

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

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996) Evidence based medicine: what it is and what it isn't. *British Medical Journal 312* (7023), 71-72. Retrieved July 14, 2005 from http://www.cebm.net/ebm_is_isnt.asp

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. 175-178.