The Challenges of Nonstandardized Vendor Usage Data in a Statewide Metasearch Environment: The Library of Texas Experience

By: William E. Moen, Fatih Oguz, and Charles R. McClure

Moen, W.E., Oguz, F., & McClure, C.R. (2004). The challenges of nonstandardized vendor usage data in a statewide metasearch environment: The Library of Texas experience. *The Library Quarterly, 74*(4), 403-422.

Abstract:
This article describes a research and development project the objective of which was to develop a means to produce standardized statewide usage data made available from the Texas State Library and Archives Commission. Using a range of data collection and evaluation methods, the project staff determined that there were significant problems in producing statewide standardized and comparable database usage statistics. This article provides background information on key issues related to these problems, offers a number of techniques that might be adapted for use in other situations, indicates the opportunities presented by metasearch applications for recording usage data, and makes recommendations for future research and work to obtain more accurate and standardized database usage statistics.

Article:
INTRODUCTION
Recent years have seen the deployment of what can be called the first generation of metasearch applications; these applications offer single search interfaces and have the potential to reduce the barriers to searching structured and full-text databases [1]. In particular, metasearch applications may enable broader use of costly commercial databases licensed by organizations, consortia, and states to serve their user communities. These applications come at a time when many organizations, consortia, and states struggle to make sense of vendor-supplied usage data of licensed resources. Efforts by librarians and online database vendors are leading to agreements for standardization of a range of statistics describing network use, users, and usage; these statistics are typically referred to as “E-metrics” [2].

Many libraries, consortia, and states are dealing with constrained funding for existing services and resources, while, at the same time, users are increasingly relying on networked information services and resources to discover, identify, select, and access information in response to their needs. Libraries spend millions of dollars annually for electronic resources. For example, the National Center for Education Statistics reports for fiscal year (FY) 2001 that public libraries spent nearly $300 million on electronic access and materials in electronic format [3]. The Association for College and Research Libraries 2002 statistics reported a similar amount for academic libraries’ expenditures for electronic serials, electronic indexes, and electronic full-text

periodicals [4]. The need for reliable and comparable data that reflect possible changes of user behaviors in use of library resources (traditional and networked) has never been greater. Yet, libraries have difficulties in preparing usage statistics because of lack of availability of standardized usage data. In particular, the nonstandard and oftentimes differing data from the online database vendors must be massaged and transformed to create comparable cross-vendor statistics that can be used for decision making.

This article examines such a situation in the context of a statewide database licensing program in Texas. The Texas State Library and Archives Commission (TSLAC) and the TexShare Program have provided licensed database resources to academic and public libraries for a number of years. A TSLAC initiative, the Library of Texas, recently deployed a metasearch application for searching Texas library catalogs and the licensed database resources. Under contract with the TSLAC, the authors have analyzed longitudinal usage data for the licensed database resources, identified issues in dealing with nonstandardized usage data, and developed a log analysis tool for usage data within the context of the metasearch application.

BACKGROUND ON TexShare AND THE DATABASE LICENSING PROGRAM

TexShare is the name of a statewide consortium of Texas libraries that was established in 1988 to improve library service to Texans. TexShare is administered by the TSLAC and focuses on the efficient sharing of library holdings, with an emphasis on electronic information resources and traditional collections of books and journals. The TexShare mission statement states:

> TexShare maximizes the effectiveness of library expenditures by enabling libraries to
>
> - Share staff expertise
> - Share library resources in print and electronic formats, including books, journals, technical reports, and databases
> - Pursue joint purchasing agreements for information services, and
> - Encourage cooperative development and deployment of information resources and technologies.
>
> TexShare programs contribute to the intellectual productivity of Texans at the participating institutions by emphasizing access to rather than ownership of documents and other information sources.[5]

Originally comprising academic libraries, TexShare membership expanded over the years and currently includes academic libraries, public libraries, and libraries of clinical medicine. TexShare enables participating libraries a means to offer a broader range of materials and services than any single library can provide. It provides a range of services in support of its mission, including interlibrary loan services; database licensing program; a courier service that affords two-day delivery among libraries statewide; grants for digitizing specialized collections; the TexShare Card, which permits borrowing from participating libraries; and training programs for TexShare library staff, to enable them to serve their customers better. The TexShare database licensing program is the focus of this article. Through this program, TexShare offers two sets of databases. It licenses a set of databases available to all TexShare members (i.e., the TexShare Core Databases). In addition, TexShare member libraries are eligible to subscribe to additional

TexSelect databases at negotiated subscription fees that are usually less than member libraries would pay if they subscribed individually to these databases.

Since the first set of TexShare databases were made available in the early 1990s, the number, type, and coverage of the licensed database resources have changed due to funding opportunities and other factors. The Electronic Information Working Group, comprising representatives of TexShare libraries, reviews and recommends electronic information products for the TexShare program and provides input from member libraries for the database licensing program. In 2000, the TSLAC submitted a successful proposal to the Texas Telecommunications Infrastructure Fund (TIF) Board for a large grant that included new funding for database resources through TexShare. In June 2001, approximately 60 databases became available for TexShare libraries because of the TIF Board funding. "The TIF grant provided the opportunity to build a premier, consolidated database program" [5].

THE LIBRARY OF TEXAS: DATABASE AND RESOURCE DISCOVERY SERVICES

The successful grant submitted by TSLAC to the TIF Board proposed the development of the Library of Texas (LOT): "The Library of Texas is a library resource sharing partnership of the Texas State Library and Archives Commission, and Texas academic and public libraries. The LOT strengthens Texas libraries by enhancing the content, services, and resources available to local libraries and their users" [6]. The LOT is modeled as a statewide virtual library that provides a set of services to extend the reach and range of Texans to library and other information resources [7, 8]. The LOT has four components: a statewide resource discovery service (i.e., the metasearch application), a wide selection of TexShare databases, indexing and preserving electronic government documents, and training librarians on electronic resources. The LOT continues to expand TexShare Core and TexSelect databases to TexShare member libraries. For example, in 1999, TexShare database expenditures were approximately $2 million. In the initial year of TIF grant funding for the databases, that amount increased to nearly $10 million. More specifics on the details of these databases offered through the LOT will be discussed below.

The other LOT component of interest in this article is the Resource Discovery Service (LOT RDS). This service provides a standards-based common search and retrieval interface to assist users in discovering information from a variety of information sources, including library catalogs, indexing and abstracting databases, full-text databases, and others. The LOT RDS is a metasearch application through which a user can use a single search interface to submit searches against one or more online resources and have the metasearch application deal with the specifics of connecting to the target resources, represent the query in appropriate syntax and semantics for each target resource, and then compile the results from the individual search targets to present to the user. The LOT RDS addresses the problem described by Judy Luther [1]:

> If our users make it to the library's web site at all, chances are they are confronted with library terminology they don't understand and a long list of databases they have to decipher and choose among. The result? Libraries are losing potential users. Librarians license valuable and costly full-text databases that we know contain the information researchers are seeking. But in a three-click world, each vendor's database remains a separate silo of information that our users don't find. Even if patrons are familiar with

searching the OPAC, that won't help them retrieve articles. Library services that require training or require the user to come to the library undermine the advantages of licensing electronic content.

To address this problem, as of March 2004, the LOT RDS provided access to the catalogs of sixty-three public libraries, twenty-six academic libraries, and forty TexShare databases.

The Texas Center for Digital Knowledge at the University of Texas received a contract from the TSLAC to support the planning and implementation for this metasearch application through the Z Texas Implementation Component of the Library of Texas (ZLOT) Project [9]. The ZLOT Project addressed three aspects of the LOT RDS:

- Virtual catalog: Texans will be able to search across multiple library catalogs from a single interface (thus creating the sense of a virtual catalog) to identify library resources without regard to geographical location of either the searcher or the resources.
- Integrating search and retrieval interface: From a single interface, Texans will be able to search diverse resources easily, thus integrating access to library catalogs, the state licensed online databases, and other resources.
- Document delivery service framework: Once Texans have discovered and identified library resources, the LOT will provide resource delivery services, for example, interlibrary loan for materials or electronic access for digital resources.

Staff of the ZLOT Project produced functional requirements and technical specifications for the application [10]. The initial contract included work activities to develop an evaluation plan for the LOT, and, in a subsequent contract, ZLOT Project staff began implementing parts of the plan. One evaluation activity was to analyze the TexShare database usage during the period when funding for the databases was part of the LOT initiative. Another evaluation activity was to develop mechanisms for analyzing usage of the LOT RDS. The remainder of this article focuses on these two activities.

A LONGITUDINAL ANALYSIS OF TexShare DATABASE USE

ZLOT Project staff undertook a longitudinal analysis of TexShare database use in order to produce reliable usage data for TexShare databases and identify potential areas for automating the collection, analysis, and reporting of TexShare database usage. The analysis was conducted for the period June 2001–August 2003 (part of FY 2001, FY 2002, and FY 2003). During this period, TIF Board funding enabled TexShare and the LOT initiative to increase the number and variety of TexShare Core Databases. The initial set of TexShare databases through the LOT initiative became available in June 2001. The TSLAC has been producing usage reports on TexShare databases for a number of year. The intent of the ZLOT effort was (1) to reanalyze the vendor-supplied usage data to corroborate the TSLAC reports and (2) to understand the opportunities and challenges of using computing tools and procedures to reduce the level of effort in preparing such reports and identify recommendations for automating the procedures where possible.

Table 1 provides a summary of vendors and number of resources included in the study. This summary indicates that a single vendor may provide a licensed resource encompassing one or

more separate databases. This is important to know exactly what the usage data reflect. Do the data reflect use of discrete databases or the use of a licensed resource, or some confounding of the two? Next, in the case of TexShare databases, the number and type of licensed resources change over time. As the table shows, in FY 2002 there were sixty-five discrete databases available; in FY 2003, because of funding reductions, only forty-seven databases were licensed. This has implications for interpreting trend information on database usage. Usage must be interpreted in the context of the number of database resources available for reporting periods. Simple aggregated counts of usage over time, without regard to the number of database resources available, are not sufficient.

Table 1 **Summary of TexShare Licensed Resources and Databases**

	Licensed Resources			**Databases Available**		
	FY 2001	**FY 2002**	**FY 2003**	**FY 2001**	**FY 2002**	**FY 2003**
Big Chalk	1	1		1	1	
Ebsco	22	23	21	25	27	20
Gale	12	12	10	12	12	10
Grolier	4	4		4	4	
Handbook of Texas	1	1	1	1	1	1
Netlibrary	1	1	1	1	1	1
OCLC	11	10	10	11	10	10
Proquest	5	5	2	5	5	2
R. R. Bowker	2	2		2	2	
Teton Data System	1	1	1	1	1	1
Tdnet		1	1		1	1
Total	60	61	48	63	65	47

The analysis of TexShare database usage encountered a number of challenges at the outset: vendors may have multiple databases per licensed resource, a changing number of databases per reporting period, and different types and formats of vendor-supplied usage data; comparable usage data were difficult to produce given different definitions of data elements such as sessions, searches, and so on, used by individual vendors; vendors may not store historical usage data to compare with extant data stored by the library. Staff for the ZLOT Project developed and reused procedures for systematically managing the analysis to address some of these challenges. For example, project staff used the procedures originally developed by the TSLAC for transforming the vendor-supplied data into the reporting elements needed by TSLAC. This provided project staff with a way of replicating the reported statistics as developed by TSLAC and to identify and report discrepancies if they occurred between the ZLOT and TSLAC analyses of the same vendor-supplied data. Through this approach, the project staff discovered the challenge mentioned above. In some cases, project staff could not retrieve historical usage data to reanalyze and confirm previous analyses.

REPORTING REQUIREMENTS AND VENDOR-SUPPLIED DATA

One of the most pressing challenges in producing useful statistics for decision making from vendor-supplied data is facing the challenges of nonstandard, ambiguous, and differing data supplied by the vendors. If a library works with only one vendor, this problem is more manageable. In the case of the TexShare databases, multiple vendors with multiple databases provide the TSLAC with usage data. Even within a single vendor, transaction reports for different databases may show variability in what data elements are available. Managing and accurately analyzing the vendor-supplied data requires customized procedures for massaging and transforming the supplied data into required reporting form. TSLAC wanted to report database use for the following three measures of use for each library type: number of sessions, number of searches, and number of documents (i.e., full-text downloads). In some cases, the vendors do not provide appropriate data that can be transformed to support these required measures. For example, table 2 summarizes the availability of vendor-supplied data that can be used to generate data for the three measures for libraries of different types. Further, the data that are provided typically must be analyzed by staff and made compatible to result in data that are accurate and rely on the same definition for the required measures.

Table 2 Availability of Vendor-Supplied Data for TSLAC-Required Measures

Vendor	Sessions	Searches	Documents	Library Type
Big Chalk	Yes	Yes	Yes	Yes
Ebsco	Yes	Yes	Yes	Yes
Gale	Yes	Yes	Yes	Yes
Grolier	Yes	No	Yes	No
Netlibrary	No	No	Yes	Yes
OCLC	Yes	Yes	Yes	Yes*
Proquest	Yes	Yes	Yes	No
R. R. Bowker	Yes	Yes	Yes	Yes
Teton Data System	Yes	Yes	Yes	No

* Data provided for public libraries only.

Some vendors provide usage data structured into elements than can easily be transformed to the required TSLAC reporting measures. For example, one vendor provides the following data elements for its databases: logins, searches, abstracts, and full-text articles. These map easily to the TSLAC measures: Logins = Sessions; Searches = Searches; Full-text Articles = Documents. Other vendor-supplied data require much more creative transformations to result in the TSLAC measures. For example, one vendor supplies the following for some of its databases: total visitors, total page views, total hits, total bytes transferred, average visitors per day, average page views per day, average hits per day, average bytes transferred per day, average page views per visitor, average hits per visitor, average bytes per visitor, and average length of sessions. Transforming these data into usable and comparable measures required by TSLAC is a challenge, and of more concern, the data resulting from the transformations may be less reliable than if the vendors supplied standardized reports with clearly defined data elements for usage data across their databases. Efforts under way through the National Information Standards Organization (NISO) on E-metrics standards [11], by the International Coalition of Library

Consortia (ICOLC) [12], and with the COUNTER Project [13] suggest that the various stakeholders recognize many of the issues related to statistics for networked information resources use and utilization; these groups are moving toward standardization and implementation of data reporting that may alleviate some of the issues described briefly above. Due, however, to changing database structures, new technology applications, and different needs by purchasers of the databases, issues related to standardizing measures may be difficult to resolve.

THE MEANING OF THE NUMBERS

Even with the challenges listed previously, the TSLAC is required to compile usage statistics that indicate the use of the licensed TexShare databases. In the longitudinal analysis conducted by the ZLOT Project, staff compiled usage statistics for the individual vendors for sessions, searches, and documents (where the data were available). In addition, project staff compiled aggregate statistics for the three measures for the total activities across all vendors and their databases. Since the TexShare consortium includes libraries of several types, it is useful to compile statistics that indicate relative use of the database resources by library type. Table 3 contains the aggregate usage statistics for the three measures.

Table 3 **Aggregated Statistics for FY 2002 and FY 2003**

	FY 2002			FY 2003		
	Academic	**Public**	**Undifferentiated**	**Sessions**	**Public**	**Undifferentiated**
Sessions	4,896,326	922,041	111,251	6,125,215	1,549,281	269,143
Searches	13,398,506	2,922,277	15,099*	16,932,824	3,317,900	179,005*
Documents	8,864,893	1,441,872	373,371	9,483,005	1,382,223	1,421,540

* One vendor does not provide search statistics.

One of the first issues in reporting aggregate data by library type occurs because some vendors do not report usage data by library type. In table 3, the Undifferentiated column shows the data for the measures where it is not possible to separate academic and public library transactions. A second issue emerges because one or more vendors do not provide data that can be transformed into one of the three required measures. This shows most clearly in the Undifferentiated column, where in both FY 2002 and FY 2003 the number of sessions is much larger than the number of searches, which results in a nonintuitive result. While project staff can explain this situation, the results indicate the softness of the aggregate statistics.

The longitudinal analysis suggested that there is ongoing use of the TexShare databases. Reported on an annual basis, for FY 2003, there were over 6 million academic library sessions and 1.5 million public library sessions, almost 17 million academic library searches and some 3.3 million public library searches, and over 9.4 million documents examined in academic libraries and some 1.3 million examined in public libraries. Fiscal year 2003 showed increases over usage in FY 2002. Users conducted some 500,000 sessions per month and some 1,000,000 (and sometimes more) searches each month.

Interpreting the numbers and understanding their meaning remain challenges. The data suggest that TexShare databases received wide use during FY 2002 and FY 2003 and that this use

appears to be increasing over time. A determination of whether this use is "significant," "high," or "low," however, is difficult to make. A number of approaches can be considered, however, to better understand these data:

- There are approximately 1,300 academic and public libraries (including branches) in Texas, so the annual use data can be normalized by comparing the average number of sessions (or downloads) per year by library (either for all 1,300 libraries or by academic or public library).
- There are approximately 21.3 million residents in Texas, so the annual use data could also be normalized by comparing the average number of all sessions (or downloads) per year per person.
- The total cost of the TexShare databases for each fiscal year could be used to determine the average cost per session, or download, (1) for all libraries or by academic or public library and (2) on a per capita basis for residents of Texas.
- Assuming the ability to provide a "legal service population" for each academic and public library participating in TexShare, the usage data can also be normalized on a per capita basis for each library—which could indicate possible variations in the use (and therefore, the cost) for each library.

As an example, with a total of some 20.3 million searches conducted in FY 2003 and, for example, a total cost of the TexShare databases of $10 million, the average cost per search was about $.49. Or, with 20.3 million searches conducted in FY 2003 and with some 21.3 million residents in Texas, there was almost one search per person conducted.

Another way in which the TSLAC illustrates the value of the TexShare databases is by producing data on cost avoidance for academic and public libraries through statewide licensing of TexShare Core Databases. This allows the TSLAC to indicate the cost savings through statewide licensing; yet, there may be other approaches to describe the efficiencies of statewide licensing of databases. For example, consider the following questions: If the average cost per search is $.49, is this "good" or "bad"? If, on average there is one search per resident per year, is that "good" or "bad?" Ultimately, TSLAC and the TSLAC governing board will need to place a value judgment on what constitutes "good" or "bad" usage. The judgment can be informed by ongoing longitudinal data and trends in those data, benchmarking the data compared to other similar states (for those few states where there is comparable and accurate data), correlating usage statistics with data coming from focus groups and individual interviews (among other techniques) that explore satisfaction and use, and comparing actual usage with target goals and objectives and expectations for desired levels of use of these valuable database resources.

The numbers by themselves do not stand on their own but need to be interpreted, linked to other usage data, and put in context of costs, data needs, and reporting requirements. While it is possible to compile the number of searches, sessions, or documents downloaded, the numbers do not indicate, for example, if these activities satisfied a user's information need, if the user found the information helpful, or if the information obtained was accurate. Additional data collection efforts, including interviews, focus groups, and surveys, could, along with the numbers, help to have a better assessment of use and utilization. The quantitative usage data provided, however,

are indicators of use but are not necessarily indicators of satisfaction or other related factors related to quality and impact.

THE OPPORTUNITY FOR USAGE STATISTICS WITH METASEARCH APPLICATIONS

The efforts of NISO, Project COUNTER, and ICOLC may lead over time to improved and standardized reporting by vendors of usage data. The advent of metasearch applications, however, presents an opportunity and a series of challenges for libraries and consortia to capture usage data in form and substance that can serve their reporting and decision-making requirements.

A metasearch application, such as the LOT Resource Discovery Service, offers the users a single search interface to multiple resources. Users can submit a search to one or more of the target resources (typically catalogs and databases), and the application sends the query concurrently to the selected target resources. The application displays search results in various groupings. Users, with the appropriate access permission, can access all licensed content in the databases. Unlike general purpose Web search engines (or Web metasearch engines), the metasearch applications provide access to what has been called the invisible or hidden Web [14]. Typically, the metasearch applications target library catalogs, full-text databases, abstracting and indexing databases, and other online database resources [1]. Metasearch applications have the potential to reduce the barriers to access by reducing the need for users to learn different native search interfaces of the individual target resources.

Figure 1 shows the various access mechanisms to the TexShare databases available to users without a metasearch application. Since the TexShare databases are licensed resources with access restricted to specific groups of users (e.g., patrons of public libraries, members of an academic library's community), the various access mechanisms include some type of access control (e.g., proxy server, IP address, user name/password login). In this approach, users search one database at a time (or several databases if such multiple database searching is provided by an individual vendor's search interface).

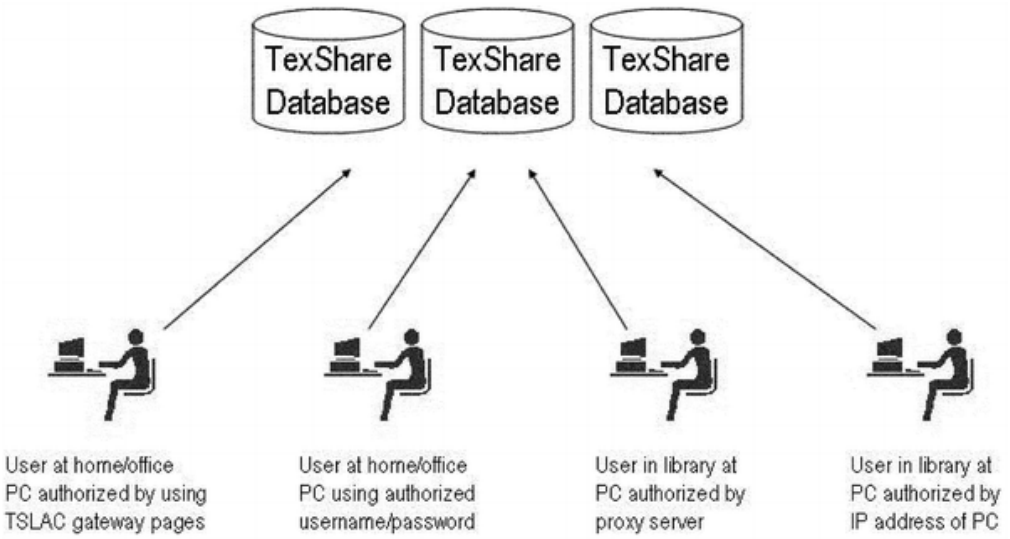

Figure 1: Access paths to TextShare databases

Figure 2 shows the LOT RDS as an intermediate application between the users and the target resources. The LOT RDS provides access control to licensed resources via IP address or user name/password login. From the search interface, authorized users can select TexShare databases as well as library catalogs and other resources to which their searches will be sent. Access paths to the TexShare depicted in figure 1 are still available in addition to the LOT RDS path.

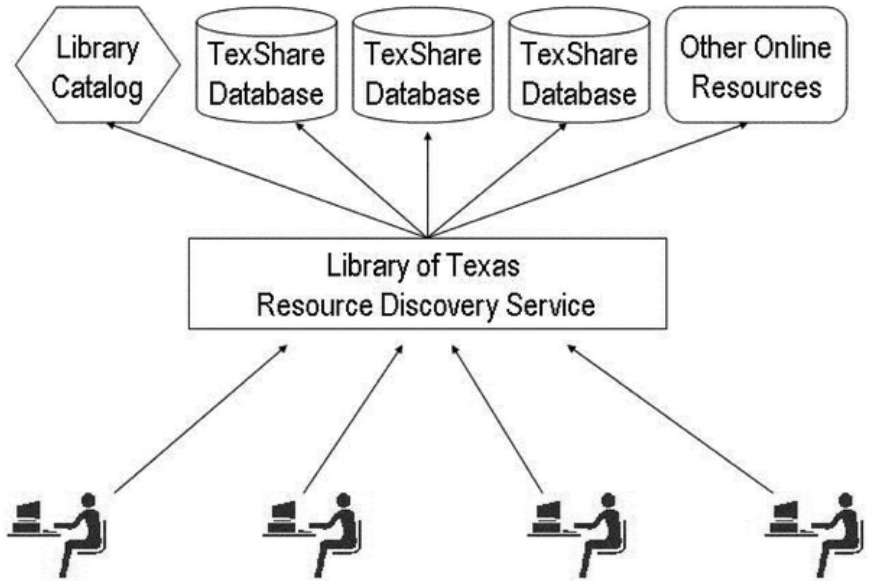

Figure 2: Metasearch access to TexShare and other databases

A metasearch application now becomes a point at which user transactions can be recorded. Further, libraries may have the ability to prescribe specific data to be collected in the metasearch transaction logs when working with the application vendor. These library-specified data can be used to develop usage reports required by the library. In the case of the LOT RDS, TSLAC and ZLOT Project staff developed a list of requirements for a transaction log, and Index Data, the metasearch vendor developing the LOT RDS application, configured the application to record the required data elements. For example, TSLAC and project staff identified reporting requirements, including counts of sessions and searches by users associated with individual Texas libraries; counts of usage of various target resources by users; counts of errors; and mechanisms to create reports for various periods of time (e.g., monthly). More specifically, TSLAC required the following reports, which informed the requirements for the transaction log file:

- Monthly by organization name
  - Total sessions
  - Total searches
    - Total searches per target resource
  - Total full-text downloads
  - Total link outs to target native interfaces per target resource
- Monthly by target resource
  - Total searches

Total successful searches

Total search errors

- Monthly by organization for each target resources

  Total searches.

To ensure the data were available, TSLAC and ZLOT Project staff identified a list of data that would be recorded for each transaction in the LOT RDS log file:

- IP address of user computer
- Log level:

  Error

  Z3950

  Verbose

  Redirect
- Session ID
- Time/date stamp
- User/system action

  Z3950

  Search

  Full

  Merge

  GetIt

  Verbose

  Favorite define

  Favorite save

  GetIt

  History

  History clear

  ILL

  ILLsubmit

  Marcdownload

  Marcview

  Outlink

  Sesclear
- Error indication and message
- Error number
- Home library of user
- Home library organization ID
- Query number

- Number of results
- Search and search term
- Search target
- CCL
- Redirection
- Referer
- Target resource ID
- Search target ZURL
- Subject category search initiated from
- Simple/advanced search page used
- Search time

With a well-known structure for the log file that records data needed for usage reports, metasearch applications provide a new transparency to usage data not currently available from the database vendors. Local analysis tools can be developed for analyzing and generating reports in the form required by the library.

The ZLOT Project staff developed a LOT RDS Log Analyzer tool using open-source software (e.g., MySQL, PhP) that would parse the RDS log file, summarize the data, and produce the reports required by TSLAC. Figure 3 illustrates the Web-based interface to the LOT RDS Log Analyzer. Various reports and in various formats (e.g., Tab Delimited and Comma Separated Value [CSV]) are available. Figure 4 shows a summary report for sessions by month for one library. (The LOT RDS became publicly available in March 2004; transactions before that month reflect testing of the system.) PhP, a common scripting language used for Web-based interaction with databases, generates tables and graphs to summarize and present the data. Downloadable reports in CSV and tab-delimited formats allow further analysis and generation of additional reports and analyses.

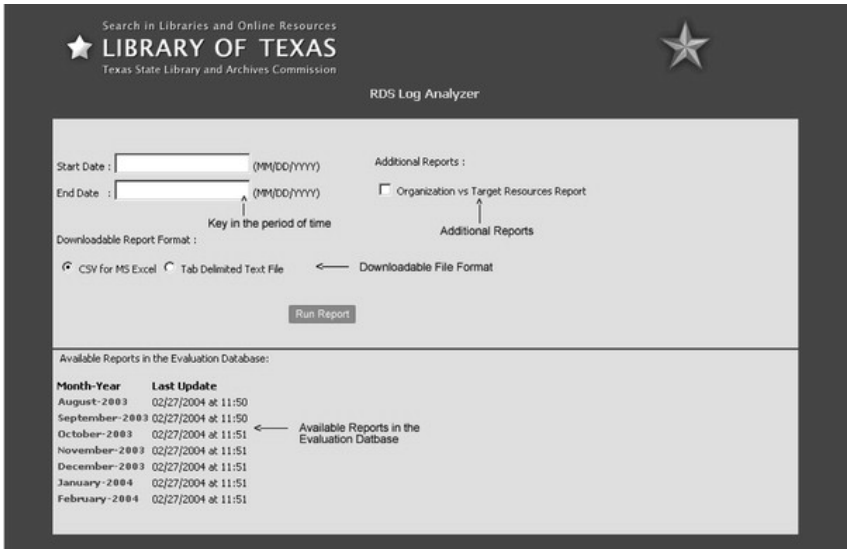

Figure 3: LOT RDS Long Analyzer Interface

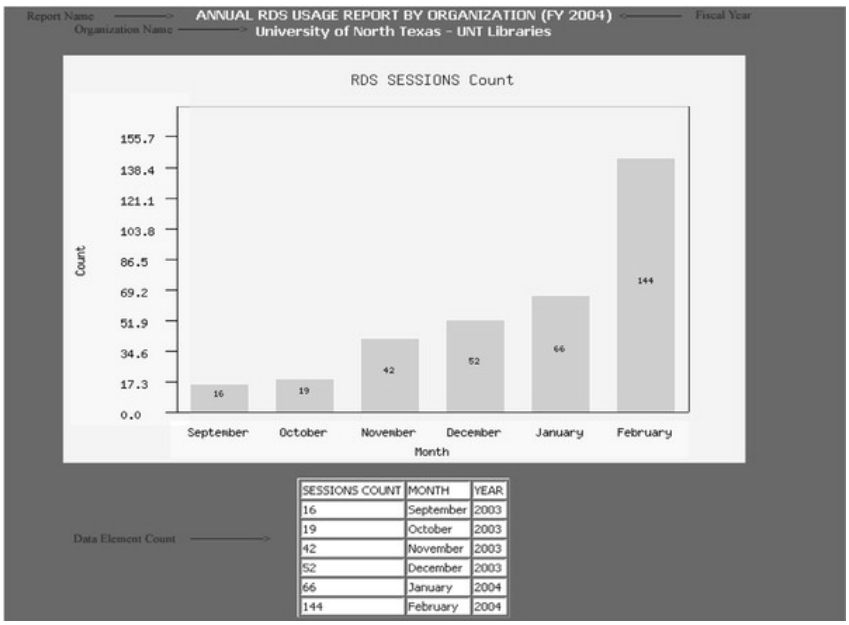

Figure 4: Sample report from LOT RDS Analyzer

As we have seen in the above list, data recorded in the LOT RDS log file goes beyond sessions, searches, and document downloads. TSLAC and ZLOT Project staff identified data elements for the log file to support generation of the required reports but identified data that could help understand users' behaviors and use of the metasearch application's functionality. For example, the LOT RDS allows a user to define groups of favorite search targets that persist across sessions. In a subsequent session, the user can select the saved group, and when the search is submitted, it is sent to each of the search targets listed in the group. By recording when users define and save groups, it is possible to understand the extent to which users use this functionality. Similarly, analysis of the log file can indicate the extent of use of other functions provided by the metasearch application. This analysis is important and necessary at this early stage of metasearch applications to understand user interaction with this innovative technology.

Metasearch applications do not mean complete independence from the reports provided by the database vendors. Through the LOT RDS application, users can link directly to a database's native interface. Subsequent transactions on the native interface are not recorded in the LOT RDS log file since the user is interacting directly with the vendor's database. And, because Texas users can access the TexShare databases through mechanisms outside of the LOT RDS, vendor data will still be necessary to develop complete usage reports on the various resources. Yet, this example shows the advantages of being able to prescribe the data that should be recorded for database transactions in the metadata search application. Library specification of metasearch log files can result in more control by the library and ease the subsequent generation of reliable usage reports.

MOVING FORWARD

Wonsik Shim and Charles R. McClure [15] point to the need for libraries and vendors to work together in resolving the problems with vendors' current usage data reports. Efforts by NISO, ICOLC, and Project COUNTER are necessary to develop community agreements (i.e., standards) that reflect the requirements of libraries and the capability of the vendors. Agreements

related to definitions, procedures, and data reporting, however, have to be implemented by all concerned parties to be effective. Libraries may have some market power to influence vendors' conformant adoption of standards given the extent of expenditures by the academic and public library communities. Yet, only concerted effort by a significant number of libraries will likely pressure the vendors to comply.

As is evident from the analysis of TexShare usage data, the current situation makes it very difficult for even motivated librarians to exploit the available data. The experience with the longitudinal analysis of TexShare database usage clearly indicates that the challenges faced by libraries and consortia in dealing with heterogeneous usage data require an inordinate level of effort to produce meaningful data that are useful for a range of decision-making activities. Further, the different types of transformations used to process the heterogeneous data to produce comparable and meaningful statistics may yield less than reliable and valid data. Finally, ingesting the heterogeneous data streams, transforming those data, and producing aggregate summary reports are not easily susceptible to automation. Any efforts at automating these processes may be undermined by arbitrary changes that vendors choose to make to existing usage data reports (whether in the format of the reports, the data elements report, etc.).

To produce the data in the longitudinal analysis, and with an acceptable level of accuracy, required a level of effort not anticipated. Overall, the effort demonstrated: the difficulties in obtaining accurate descriptive data from individual vendors; some key differences among vendors in their definitions (or lack of definitions) of key terms such as sessions, downloads, and so on; the problems encountered by TSLAC in maintaining accurate data and collating those data from the various vendors; the problems encountered when retrieving vendor data from earlier months and having those data differ from the data as originally retrieved from the vendor site; and the difficulties in developing and integrating standardized reporting tools across different vendor data to summarize usage across the TexShare databases. As part of the overall LOT evaluation plan implementation, the ZLOT Project staff has worked on developing an evaluation database to hold selected data that can be used to meet TSLAC reporting requirements. One of the reasons for undertaking this longitudinal analysis of TexShare database usage was to work closely with the vendor-supplied usage data to determine ways to automatically ingest summary statistical data into the evaluation database. However, the experience from this current analysis suggests that TSLAC will likely need to continue producing usage statistics manually until there are better standards and interoperability with the data being produced by the vendors. In a number of instances, when developing the summary tables reported here, data from vendors had to be individually reviewed and sometimes corrected before they could be summarized accurately.

Project staff believe that usage data help libraries understand how their resources and services are used (or not), and these data are a critical component of ongoing evaluation efforts. Licensed database and other digital resources reflect major investments by libraries and consortia. Usage data for these valuable and expensive resources are necessary for library managers to secure funding for the resources; allocate appropriate resources regarding selecting and deselecting databases; and understand patterns of use. Further, as illustrated in the discussion above about the LOT RDS transaction log, analyses of those data can provide a nonintrusive method for

identifying user behaviors that can lead to design and functionality improvement of a service (such as a database) or application (such as metasearching).

Whether the databases are licensed by an individual library or consortia, the nonstandardized data streams and other usage data (e.g., from a metasearch application) need to be managed. Ongoing evaluation efforts require a data management approach that addresses issues such as overall responsibility for data management and specific responsibility for collecting the data, designing and maintaining a management information system, entering the data in a management information system, verifying data with vendors, analysis of data, and reporting data; the budget available to support data management of online usage statistics; the audiences that the usage statistics will be reported to and analyses that remain internal to the organization; and the necessary skills of staff to collect, analyze, and report these data. The extent to which data management activities can be automated is open to question—as shown in this article. The LOT RDS Log Analyzer, however, shows what can be done in automated processing of log files and generating required reports when the usage data are well structured and appropriate to the evaluation and reporting needs of the library.

As libraries, consortia, and statewide digital libraries continue to grow and expand, they will continue to rely on licensed and other types of databases resources. Database usage statistics are essential for ongoing evaluation of the information and access services provided by individual libraries and consortia. Usage statistics can be used to justify or refine database selection. Ongoing collection, analysis, and reporting of usage statistics are essential components for collection development, as well as overall library services planning. Further, basic usage statistics can be used to develop a range of performance measures and indicators for outcomes assessment. Shim and McClure state the issue clearly as a result of work they completed for the Association of Research Libraries E-Metrics Project: “Librarians need reliable and accurate statistics that will allow them to make good resource allocation decisions (e.g., cost-benefit analysis, contract negotiation, justification of expenditures), meet user needs (e.g., identifying barriers of access, understanding user behaviors), and develop strategic plans (e.g., user education, peer comparison) for the development and operation of electronic services and resources. Strategies suggested in this article, combined with other data collection approaches, are currently the best overall means to obtain meaningful descriptive information about database users” [15, p. 501].

The findings presented in this article also suggest the need for ongoing research in this area. The project summarized here occurred over a two-year period with the support of the TSLAC. Related work on developing assessment tools for statewide databases and statewide digital libraries is also under way in the state of Florida [16]. There continues to be a significant need to develop more accurate and descriptive usage statistics and measures, automated means to collect such data (given the amount of data under consideration), and better education of librarians on why the data are needed and how to use them in decision making. The E-Metrics Instructional System, now under development with funding from the U.S. Federal Institute of Museum and Library Services, will be a freely accessible online instructional system to learn about E-metrics and get help collecting and using E-metrics data [2].

Given the current difficulty in working with nonstandardized data from the vendors, libraries may forgo efforts in compiling and analyzing the data. Good decision making will suffer if reliable and valid information is not available. One might suggest that the vendors may benefit from the current situation in that libraries find it too difficult to compile and analyze the data and so licensed database resources are chosen not on the basis of adequate information but on the basis of other factors such as politics and individual personalities. With such large expenditures at stake regarding the purchase and administration of statewide databases, accurate and timely usage data are essential.

Yet, as in the case of Texas, funding for licensed databases is not guaranteed. Vendors' revenues may suffer as a result of libraries not being able to make the case to funding agencies that these licensed resources are not only vital but are also cost-effective and highly used and bring benefits to users. Making such cases requires reliable and valid data that are easily used by libraries to generate a range of reports to support the claims. The NISO, ICOLC, and Project COUNTER initiatives, as well as metadata applications that offer new opportunities for tracking usage, must move forward. Libraries, database vendors, and system providers all have a stake to move from nonstandardized usage data to the development, implementation, and use of community agreements that define data elements, agree on data collection procedures, and report usage data similarly.

REFERENCES

1. Luther, Judy. "Trumping Google? Metasearching's Promise." *Library Journal* 128 (October 1, 2003). Available at http://www.libraryjournal.com/article/CA322627.
2. Bertot, John Carlo; McClure, Charles R.; Davis, Denis M.; and Ryan, Joe. "Capture Usage with E-Metrics." *Library Journal* (May 1, 2004). Available at http://www.libraryjournal.com/article/CA411564?display=FeaturesNews=Features=1987=151.
3. National Center for Education Statistics. "Public Libraries in the United States: Fiscal Year 2001." Report. March 2004. Available at http://nces.ed.gov/pubs2003/2003399.pdf.
4. Association of College and Research Libraries. "ACRL 2002 Academic Library Trends and Statistics." Report. Available at http://www.ala.org/ala/acrlbucket/statisticssummaries/2002stats/2002statisticalsumm.htm.
5. Texas State Library and Archives Commission. "About TexShare." Available at http://www.texshare.org/generalinfo/about/.
6. Overview of the Library of Texas. Web site. Available at http://www.tsl.state.tx.us/lot/overviewlib.html.
7. Moen, William E., and Murray, Kathleen R. "A Service-Based Approach for Virtual Libraries Designing." *Texas Library Journal* 78 (Fall 2002): 96–100.
8. Moen, William E., and Murray, Kathleen R. "Designing and Demonstrating a Resource Discovery Service for the Library of Texas." *Texas Library Journal* 78 (Fall 2002): 101–6.
9. Texas Center for Digital Knowledge. "Z Texas Implementation Component of the Library of Texas Project." Web site. Available at http://www.unt.edu/zlot/.
10. Moen, William E., and Murray, Kathleen R. "Functional Requirements for the Library of Texas Resource Discovery Service: ZLOT Project Deliverable C." Texas Center for Digital Knowledge, University of North Texas. June 2002. Available at http://www.unt.edu/zlot/fr_index.htm.

11. National Information Standards Organization. “Information Services and Use: Metrics and Statistics for Libraries and Information Providers: Data Dictionary (Draft Standard).” NISO report Z39.7-200X. 2004. Available at http://www.niso.org/emetrics/.
12. International Coalition of Library Consortia. “Guidelines for Statistical Measures of Usage of Web-Based Information Resources.” December 2001. Available at http://www.library.yale.edu/consortia/2001webstats.htm.
13. Counter Online Metrics. “COUNTER: Counting Online Usage of Networked Electronic Resources.” Web site. Available at http://www.projectcounter.org/index.html.
14. Murray, Kathleen R., and Moen, William E. “The Deep Web: Resource Discovery in the Library of Texas.” *Texas Library Journal* 80, no. 1 (Spring 2004): 16–19, 22–24.
15. Shim, Wonsik, and McClure, Charles R. “Improving Database Vendors’ Usage Statistics Reporting through Collaboration between Libraries and Vendors.” *College & Research Libraries* 63 (November 2002): 499–514.
16. McClure, Charles R.; Bertot, John Carlo; and others. “Measures and Statistics to Assess the Florida Electronic Library.” Report. Information Institute, Division of Library and Information Services, Florida State University, 2004.

The authors want to thank Beverley Shirley, Russlene Waukechon, David Hardy, and Kevin Marsh at the Texas State Library and Archives Commission for their assistance in the work on the longitudinal analysis of TexShare data and the work on the LOT RDS log analysis too. Funding for this work was provided by the Texas State Library and Archives Commission.