5-2004

Determining the Data Needs for Decision Making in Public Libraries

Thena Slape Jones
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a thesis written by Thena Slape Jones entitled "Determining the Data Needs for Decision Making in Public Libraries." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Information Sciences.

Peiling Wang, Major Professor

We have read this thesis and recommend its acceptance:

Kendra S. Albright, N. Douglas Raber

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a thesis written by Thena Slape Jones entitled "Determining the Data Needs for Decision Making in Public Libraries." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Information Sciences.

Peiling Wang
Major Professor

We have read this thesis
and recommend its acceptance:

Kendra S. Albright

N. Douglas Raber

Accepted for the Council:

Anne Mayhew
Vice Provost and Dean of Graduate Studies

(Original signatures are on file with official student records.)

DETERMINING THE DATA NEEDS FOR DECISION MAKING IN PUBLIC LIBRARIES

A Thesis
Presented for the
Master of Science Degree
The University of Tennessee, Knoxville

Thena Slape Jones
May 2004

DEDICATION

To my father-in-law, the late Gordon W. Jones, M.D., who said every woman needs a hobby;

and

To my husband, Kevin, who has cheerfully put up with mine.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Dr. Peiling Wang, for her guidance in all aspects of this thesis and for her generous hospitality. She suggested alternatives when things did not go as planned, and pushed me for improved clarity and organization. I would like to thank the other members of my committee, Dr. Kendra Albright and Dr. Douglas Raber, who emphasized the business and management aspects of the study. I would like to thank the subject library and its administrators who participated whole-heartedly in the project; with special appreciation to the system administrator, network services manager, and adult services coordinator for answering many, many questions.

I would like to thank those who supported my pursuit of this Master's degree. My employer, immediate supervisors, and co-workers eased scheduling conflicts each semester. My fellow students provided online company as we worked on papers and projects at all hours. Most importantly, I would like to thank my husband and three children who always provided encouragement, enthusiasm, forbearance, or patience—in different measures—through the entire course of study.

ABSTRACT

Library decision makers evaluate community needs and library capabilities in order to select the appropriate services offered by their particular institution. Evaluations of the programs and services may indicate that some are ineffective or inefficient, or that formerly popular services are no longer needed. The internal and external conditions used for decision making change. Monitoring these conditions and evaluations allows the library to make new decisions that maintain its relevance to the community.

Administrators must have ready access to appropriate data that will give them the information they need for library decision making. Today's computer-based libraries accumulate electronic data in their integrated library systems (ILS) and other operational databases; however, these systems do not provide tools for examining the data to reveal trends and patterns, nor do they have any means of integrating important information from other programs and files where the data are stored in incompatible formats. These restrictions are overcome by use of a data warehouse and a set of analytical software tools, forming a decision support system. The data warehouse must be tailored to specific needs and users to succeed. Libraries that wish to pursue decision support can begin by performing a needs analysis to determine the most important use of the proposed warehouse and to identify the data elements needed to support this use.

The purpose of this study is to complete the needs analysis phase for a data warehouse for a certain public library that is interested in using its electronic data for data mining and other analytical processes. This study is applied research. Data on users' needs were collected through two rounds of face-to-face interviews. Participants were selected purposively. The phase one interviews were semi-structured, designed to discover the uses of the data warehouse, and then what data were required for those uses. The phase two interviews were structured, and presented selected data elements from the ILS to interviewees who were asked to evaluate how they would use each element in decision making.

Analysis of these interviews showed that the library needs data from sources that vary in physical format, in summary levels, and in data definitions. The library should construct data marts, carefully designed for future integration into a data warehouse. The only data source that is ready for a data mart is the bibliographic database of the integrated library system. Entities and relationships from the ILS are identified for a circulation data mart. The entities and their attributes are described.

A second data mart is suggested for integrating vendor reports for the online databases. Vendor reports vary widely in how they define their variables and in the summary levels of their statistics. Unified data definitions need to be created for the variables of importance so that online database usage may be compared with other data on use of library resources, reflected in the circulation data mart.

Administrators need data to address a number of other decision situations. These decisions require data from other library sources that are not optimized for data warehousing, or that are external to the library. Suggestions are made for future development of data marts using these sources.

The study concludes by recommending that libraries wishing to undertake similar studies begin with a pre-assessment of the entire institution, its data sources, and its management structure, conducted by a consultant. The needs assessment itself should include a focus group session in addition to the interviews.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

STATEMENT OF THE PROBLEM

Introduction

Public libraries offer many different programs and services to their communities. Library decision makers examine community and library conditions, such as population characteristics, funding levels, availability of physical resources, and staff capabilities, in order to select the appropriate services offered by their particular institution. Their decisions may include selecting a service response, such as Information Literacy, Local History and Genealogy, or Cultural Awareness, during a formal planning process (Nelson 2001). Or the decisions may center on collection development policy (Evans and Zarnosky 2000). Informed decisions must be made on every facet of library service, from selection of vendors to operating hours to marketing.

These decisions will need to be revisited. The internal and external conditions used for decision making change. Evaluations of the programs and services may indicate that some are ineffective or inefficient, or that formerly popular services are no longer needed. Monitoring these conditions and evaluations allows the library to make new decisions that maintain its relevance to the community.

Administrators must have ready access to appropriate data that will give them the information they need for library decision making. In the corporate and government environments, increased automation has resulted in an explosion of stored data. This electronic data, carefully selected, prepared, and combined, may be processed by analytical software to present results to improve decision making. This is accomplished through a decision support system, which consists of a data warehouse and the set of applications that perform the analysis (Berry and Linoff 2000). The role of the data warehouse is to bring the various data elements together in a manner that makes them available for the analytical tools such as online analytical processing or data mining. These programs comb through data in the warehouse and find hidden associations or reveal trends.

Today's computer-based libraries also accumulate electronic data, but are not using this information effectively for decision making. Though their data stores and information needs are much smaller than those of corporations, libraries are increasingly interested in analyzing their data in a similar manner (Guenther 2000). Public libraries collect most of their electronic data by way of acquisitions, cataloging, circulation, and patron records, generally combined in the integrated library system (ILS). The ILS and other computer-based programs would seem to make data readily available for decision support. These systems do

not, however, provide tools for further analysis—examining the data in more complex ways to reveal trends and patterns upon which meaningful decisions can be made. Nor do they have any means of integrating important information from other programs and files, where the data are stored in incompatible formats.

Harnessing data for decision support requires a large commitment in terms of time, personnel, and money, which are generally in short supply in public libraries. The data warehouse is the foundation of the decision support system. Developing the warehouse involves determining the decision situations it will support, identifying the necessary data, creating logical and physical data models, determining data cleaning and mapping requirements, creating the warehouse architecture, importing the data, and maintaining the warehouse (Kimball 1996; Kimball et al. 1998). Most businesses, needing an enterprise-level data warehouse, will either rely on in-house systems analysts and IT departments, or hire outside consultants. Projects of this size are beyond the scope of most public libraries, yet many may wish to create a small data warehouse either to address a particular decision task or to use as a proof of concept before creating a larger system.

The data warehouse must be tailored to specific needs and users to succeed. Without this focus, it will have no relation to the information needs of its intended users, and will be among the large percentage of data warehouse implementations that are considered failures due to lack of use. A library that wishes to pursue decision support can begin by performing the needs analysis to discover and prioritize potential uses for the data warehouse and to determine the location of the data elements needed to support these uses.

Publications on creating data warehouses and data mining environments are intended to guide enterprise implementations, which are expected to increase company profitability. They are expensive undertakings requiring project teams and specialists for each stage of the warehousing project. Such a large scale commitment is often not possible for a library wishing to create a small data warehouse or a proof of concept warehouse project. More importantly, libraries do not have a profit motive. They should be aware that introducing a profit-oriented business process to their decision making could create the expectation, or requirement, of further business approaches to library administration in the future that are antithetical to the purposes of a public library. Libraries may nevertheless use decision support to gain a better knowledge of the people they serve, or examine library functions for efficiency and effectiveness.

Statement of the Problem

Libraries have a growing collection of data that they gather in the course of operations, and are increasingly aware that the data could be exported to a data warehouse and examined with analytical software tools to make them more meaningful for decision making. Success of a data warehouse for decision support rests in part on tailoring the warehouse to the needs of the users and their particular decision situations.

There are no published guides for determining the users' data needs in small organizations. Libraries that wish to examine data warehousing must adapt the techniques intended for creating an enterprise warehouse and devise their own implementation strategy based on their intended use, their staff capabilities, and their ability to employ staff with the expertise to create and maintain the warehouse. Libraries may begin by pursuing the needs analysis phase to determine what aspects of library decision making would benefit most from decision support.

Purpose of the Study

The purpose of this study is to complete the needs analysis for a data warehouse for a particular public library. This library is interested in using its electronic data for data mining and other analytical processes.

The objectives of this research are:

1. to understand data needs for better decision making in public libraries;
2. to prioritize these data needs;
3. to suggest schemes and methods to reconcile data from original diverse sources for building a data warehouse to support these needs.

In order to meet these objectives, the following research questions are posed:

1. What are the decision situations encountered by library administrators?
2. What data do public library administrators need to support their decision making?
3. How may the necessary data be obtained and organized for decision support?

This study examines one library's data requirements for a decision support system. It demonstrates a means for accomplishing the first stage, i.e., the needs analysis phase, of a data warehouse for decision support. Then a model is developed to show relationships among the data elements selected for the data warehouse.

Decision Support Systems and the Data Warehouse

A decision support system includes all technical resources that are used in the collection, analysis, and presentation of data in ways that make them meaningful for decision

making. It is different from an online transaction processing system, which is designed for rapid and secure recording of transactions. Decision support systems have two primary components: the data warehouse, and the set of analytical software tools.

The data warehouse is used to collect the data identified as necessary for decision making. The warehouse architecture is specifically designed for querying. Data elements in the warehouse are described and standardized if they derive from different sources. Querying is performed by the second component, the set of analytical tools. Three of these analytical tools are querying and reporting, online analytical processing, and data mining.

Querying and reporting is the simplest tool. It is used in many operational databases to produce pre-defined reports. These reports may be somewhat customizable, but the design of the program limits the ways that data may be selected, manipulated, and compared. When the program data are exported to a data warehouse, querying and reporting can overcome the limitations of operational reports and provide results that are tailored to each users' information needs.

Online analytical processing (OLAP) allows analysis of data along different entities, called dimensions. A multidimensional analysis of circulation records may examine records by patron zip code, patron age, library branch used, fiction genres or non-fiction call numbers, and time of checkout. A graphical interface gives users easier data selection and manipulation.

Data mining is not a single application, but rather a set of analytical processes that allow the discovery of patterns in data. Some processes are directed, meaning that the records are processed along a specified variable. Classification, estimation, and prediction are directed data mining processes. Other processes are undirected, in which the program is asked to discover the relationships among the records. Affinity grouping and clustering are examples of undirected data mining.

Visualization tools are included in many of the software applications used for querying and reporting, OLAP, and data mining. These tools allow the query results to be presented graphically, in a format–such as a chart, graph, or map–suited to the query and the user. Sometimes visualization is considered a data mining tool, when it is used to create a graphical presentation of complicated data relationships.

A data warehouse, with data from a variety of operational systems and other data sources, is very complex. It is difficult to create and to maintain. To reduce the complexity and expense a library may create a more limited data warehouse, known as a data mart. The data mart is usually based on a single operational database. More data marts may be created over time. If they are carefully designed to a common set of data naming and definition standards, they will eventually form an integrated data warehouse. Part of the needs analysis is determining the best library function to use for an initial data warehouse or data mart.

In a public library setting, these functions may include budgeting, acquisitions, circulation, collection development, and personnel, among others. The initial data warehouse for a public library is developed from a prioritized list of user needs related to these functions, and on availability of quality data from a single source. Some libraries use an integrated library system (ILS) in which each function is a module that accesses data from a single bibliographic database, blurring the line between them and allowing the creation of a data warehouse that can answer a greater number of decision situations.

Methodology

This study is applied research. A study of data needs for library decision making must be contextual rather than general. Even if the focus were limited to public libraries of similar size, the context of each institution is unique in governance, community, mission and goals, management style, physical layout and constraints, software limitations and capabilities, and many more variables. Therefore, this study focuses on the needs of one particular library.

Data on users' needs were collected through two rounds of face-to-face interviews. Participants were selected purposively, as being all those members of the library administration who implement the long-range planning decisions for the entire library system under the direction of the Board of Trustees, or who plan and evaluate programs and services in their particular departments that will fulfill the library's objectives. Examination of the administrative structure and discussions with the two appointed library contacts indicated that nine persons from the top two tiers of the administrative structure met these criteria.

The phase one interviews were semi-structured. The content analysis scheme was developed during analysis, based on patterns of responses. The phase two interviews were structured, with open-ended questions at appropriate intervals. The interview questions were used to create a decision support matrix, which was organized using the schemes developed in phase one. The organization of the data in this manner clarified users' data needs and suggested the incremental stages of warehouse development.

Elements of the Problem

The study was designed with the following known difficulties that users have in articulating their information needs: (Berry and Linoff; Kimball et al. 1998; McClure et al. 1989):

1. They may not recall their problems at the time of interviewing, even when asked to recall a specific incident in which they did not have needed data, or the data required considerable time and effort to collect and analyze.

2. They may not express their information needs accurately, in a manner analogous to reference desk transactions, where the patron often does not ask for what he needs but what he thinks will lead him to the information he needs.

3. They may be unrealistic in their expectations regarding availability of data sources and capabilities of decision support systems.

Delimitations of the Study

The library system

The subject of this study is a medium-sized East Coast regional public library system with seven branches and a bookmobile, funded by four different city or county governments. It serves a population of about 234,000 people (U.S. Bureau of the Census 2001). According to FY 2001–2002 Annual Statistics, it circulated 5,476,752 items to 161,619 library patrons with a budget of $7,353,203.

The four constituent governmental units (known here as the City, and Counties A, B, and C) were fairly equal in population after World War II, with 1950 census populations ranging from 10,148 to 12,158. The city was the commercial center for these and other nearby counties, which were all strictly rural. Huge growth in a major metropolitan area to the north, coupled with spiraling housing prices in that area, caused a surge in population in County B in the 1960s and County A in the following decade, particularly with the completion of a major interstate highway through the area. County C, physically separated from the city and other counties, remains rural and undeveloped (tables 1 and 2).

The City had supported a library for its residents for several decades before the regional system was established by the state in 1969, as a model for areas lacking library services. The participant jurisdictions began funding the system in 1971.

The library system is governed by a seven-member board of trustees; two each from the City and Counties A and B, one from County C. They are appointed by the localities' governing bodies, and often one trustee from the larger localities will be a county supervisor or city councilman. The board sets policy for the system and hires the director.

The director implements policy as set forth by the Board of Trustees. As the organizational chart in figure 1 indicates, the library uses a hierarchical management system with the director and deputy director at the top of the hierarchy. Personal style leads the two present incumbents to work as a close team, though the director is the library's voice to trustees, funders, and other library supporters, while the deputy director communicates policy to library staff. Seven managers and coordinators report to the deputy director. They

Table 1.
Population of the Library Service Area: 1950–2000

	City	County A	County B	County C	Regional Total	State
Population counts						
1950	12,158	11,920	11,902	10,148	46,128	
1960	13,639	13,819	16,876	11,042	55,376	
1970	14,450	16,424	24,587	12,142	67,603	
1980	15,322	34,435	40,470	14,041	104,268	
1990	19,027	57,403	61,236	15,480	153,146	
2000	19,279	90,395	92,446	16,718	218,838	
Percent change						
1950–1960	12%	16%	42%	9%	20%	20%
1960–1970	6%	19%	46%	10%	22%	17%
1970–1980	6%	110%	65%	16%	54%	15%
1980–1990	24%	67%	51%	10%	47%	16%
1990–2000	1%	57%	51%	8%	43%	16%
1950–2000	59%	658%	677%	65%	374%	117%
1980–2000	26%	163%	128%	19%	110%	34%

Table 2.
Area and Population Density, 2000

	Area (sq. mi.)	Population	Density
City	11	19,279	1,833
County A	401	90,395	226
County B	270	92,446	342
County C	229	16,718	73

Sources for both tables: U.S. Census Bureau, State and County Quick Facts, http://quickfacts.census.gov/qfd/
State Statistical Abstract

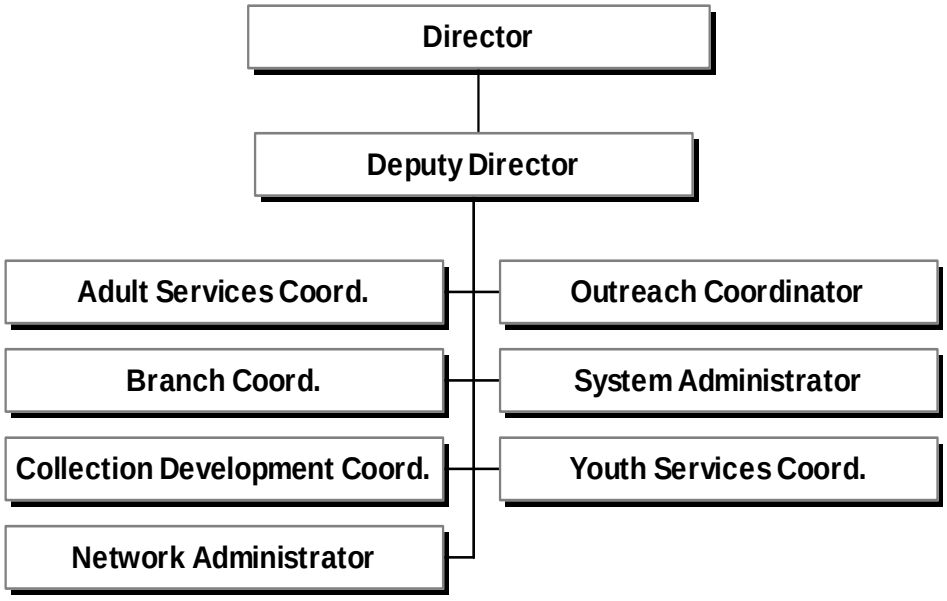

Figure 1. Organization chart

are frequently asked to examine data and provide reports and recommendations on the aspects of the library that they manage. Final decisions are made by the director and approved by the Trustees, based on information provided by the coordinators.

The library has seven branches to serve its population area. The largest branch is the headquarters in the City. This branch, renovated in 1991, supports the entire regional system. It also contains the system's two special collections, in law and in local history and genealogy. Because initial growth in counties A and B was contiguous to the city, this branch serves many residents of those counties as their preferred branch; about 70% of the circulation at this branch in 2001 was to non-city residents. Part of the regional funding formula is centered on this statistic.

Branch B, in County B, was opened in 1992. It contains 23,000 sq. ft. with capacity for 100,000 volumes; it is the second-largest facility in the system.

Counties A and B have similar population counts, but County A is physically larger. It has two branches, A1 and A2. Branch A1 opened in 1994 with 16,000 sq. ft. and capacity for 50,000 volumes. It has been at capacity since 2000, and must move books constantly to the system's off-site storage facility to make room for new titles.

Branch A2, opened in 1998, is much smaller. It replaced a two-room facility that had been at capacity for several years.

The rural nature of County C requires three small branches to serve its dispersed population. All three branches have moved into new facilities in the last two years. These branches are run by part time paraprofessionals who answer to the branch coordinator. The nearest branch is about a 35 minute drive from headquarters, and the farthest is just over an hour from headquarters. The library is committed to providing equal services to all branches, but the remoteness of County C from the remainder of the system has sometimes made this difficult. The library's adoption of a new integrated library system in August 2002 provided a user-friendly Web interface for renewing checked out items and placing hold requests. The impact has been felt system-wide, but there are indications that it is particularly important in this county. Administrators believe that statistics for County C can now provide a clearer profile of its residents and their library preferences.

Technology in the library system

The library's first automated library system was Dynix. Implementation of the program modules began 1991, with circulation, cataloging, acquisitions, and public access catalogs in place within a year. Implementations of the serials module, DialPAC (dial-up telnet access to the catalog) and other modules were completed by 1994. By this date, also, the reference and juvenile desks in the City and Counties A and B had access to CD-ROM applications and staff access to FirstSearch. The Five-Year Plan published in 1993 predicted rising technology costs for the future not only because of the cost of acquiring and establishing computerized services, but also due to the additional expense of upgrading equipment and implementing a LAN/WAN system. This system was established in 1996 for the three major branches, with the smaller branches brought into the WAN in 1997. By June 1997, the library had a World Wide Web page available on the Internet.

In 1993, the library hoped to offer public dial-in access to its CD-ROM databases in five years. By 1998, however, more database subscriptions were moving from CD-ROM to Internet access, and by 2000 most of the library's online database subscriptions were available to library patrons from any online computer.

Dynix, the library's original integrated system, was replaced with Horizon in August 2002. Patron response has been enthusiastic. Employee response has been mixed. Horizon has many improvements over the previous system, but also lacks many processes and reports that staff members and administrators alike used to increase efficient service. Some of the complaints have been resolved over time, as staff learn more of Horizon's capabilities or adjust their own work processes.

Factors impacting library services

Several factors are having an increasing impact on the library and its services. The first is the rapidly growing population of the service area, one of the fastest growing regions in the state. Not only does increased population impact buildings, materials, and services because of sheer numbers, but the population growth has also caused shifts in the

characteristics of this population. This growth and change in demographics will continue for many years to come; it is an overarching tenet in all aspects of public planning.

Equally important, though of shorter duration, is the effect of the economic downturn on library planning. The state has seriously cut all public funding, including libraries, with more cuts promised in the near future. Local taxes have been held fairly constant, with the net result that actual funding for all public programs is decreasing since the tax dollars must be stretched further due to population growth and inflation. It is in this milieu that the library seeks to evaluate its programs and services, shifting services to meet new needs, continuing successful and popular programs, all while maintaining its historic high quality in an increasingly cost-conscious atmosphere.

The library has created and continuously updates a five-year plan. It seeks to remain updated on population growth, new areas of commercial and residential development, changes in the demographic characteristics of the population, and changes in curriculum and educational support needs.

Influences toward decision support

In the spring of 2002 the library's administrators attended a workshop on data mining conducted by Dr. Scott Nicholson, assistant professor in the School for Information Studies, Syracuse University. Dr. Nicholson opened with an overview of the potential benefits of data mining and OLAP techniques, as shown by the PowerPoint handout kept by one of the participants. The group then brainstormed ideas for applying these techniques to this specific library and concluded by drawing a series of dimensional models that reflected some of these library needs. The administrators came away with considerable enthusiasm for investigating data mining and decision support systems. Within a year, the library underwent two major events that emphasized the fact that the library has considerable data, but is unable to access them in meaningful ways.

The first event was the library's migration from Dynix to Horizon as its integrated library system. Horizon and Dynix have different reporting capabilities. Horizon has some welcome new report features, but omits others which some of the administrators had relied on to evaluate and maintain library services. The problems are of two types: (1) inability to use fields from the major tables (patron, bibliographic, and item) that exist and are used, but that Horizon has not included in its patron and item report generators, and (2) Horizon's summarization of circulation statistics far earlier and at a higher level of aggregation than that used in Dynix. It was expected that administrators would cite these problems during this study as areas in which a data warehouse could collect system information for creating reports beyond the capabilities of the new ILS.

The second event was the threat of one of the member constituencies of the regional library system to withdraw and set up its own library system. This county was attempting to keep public services at their then current level, though for a rapidly increasing population,

while maintaining or even cutting the real estate tax rate. Some county supervisors believed that the county could run its own library system at less cost than its participation in the regional system, and still offer comparable quality of services. Both the county and the library required statistics to support their arguments, ranging from branch use patterns of county residents to the impact on start-up costs if the county could not retain all materials and furnishings in the two current county branch buildings. Over the next few months, library administrators put aside other tasks to concentrate on gathering statistics, many of them not available in standard library reports. The impasse has since been resolved with the county remaining in the system, but it was again expected that administrators would recall these informational problems, looking for improved access to data with tools to better manipulate the numbers.

Limitations of the Study

This is an intrinsic study of the decision making needs of a specific public library. The goal is to examine this subject in depth for this institution, rather than to generalize to a broader population (Stake 1995). The decision making needs for this library are intertwined with its particular environment, so that generalization to other public libraries is not possible (Swisher and McClure 1984).

Creating a data warehouse for decision support requires technical planning and implementations that are not part of this study, including analysis, cleaning and extraction of the data, and determination of the specific analysis needs of the users for acquisition of specific analytical and visualization tools.

Definitions

data mart. A small data warehouse (*which see*), usually limited to a single subject area. Data marts may stand alone, or may form an integrated data warehouse by using standardized data definitions across all marts.

data mining. A number of different statistical techniques that are used on large data stores for prediction or discovery. Data mining techniques may be directed, also called verification driven (the user tells the software what to look for), or undirected, called discovery driven (the user asks the software to examine the data and find actionable patterns or clusters hitherto unknown). Data mining includes clustering, classifying, estimating, predicting, and affinity grouping. Market basket analysis is one type of affinity grouping data mining technique.

data warehouse. A software repository for data collected from one or multiple operational databases or external data sources, gathered in a manner to coordinate them for analysis with an analytical software package. The warehouse is the storage component of a decision

support system. The warehouse may vary in size and complexity, and may be created in a spreadsheet, relational database, or multidimensional database.

decision support system. Also called DSS. (1) An information system designed to collect and analyze data needed for decision making. The data are collected in the data warehouse, where they are available for analysis. There are many types of analytical tools; several will be needed to serve the various users of the data warehouse. The analytical tools are generally broken down into several categories, including querying and reporting, online analytical processing, and data mining *which see*. (2) A type of information system that is characterized by the ability to apply rules to the data or perform simulations needed for strategic decision making. *See also* management information system.

entity relationship modeling. The logical architecture of most relational databases, which are configured for online transactional processing, *which see*. Data is organized into tables in a manner that reduces or eliminates redundancy, in order to optimize transactions and reduce errors. Operational databases are usually built on the entity relationship model, which is not efficient for querying.

integrated library system. Also called ILS. A computer-based system that integrates several library operations. The bibliographic database is usually the central component of the ILS, and has separate user interfaces for cataloging and for searching. The circulation module contains the patron database and automates circulation functions. Other modules that may be integrated in the system are acquisitions, ILL, and serials control.

management information system. Also called MIS. (1) An analysis tool that operates on the data either in an operational database or in a data warehouse. An MIS organizes and generates timely information for management oversight at the management and operational levels of an organization. Reporting includes detailed transaction lists, summary reports, and exception reports or alerts that notify managers when particular parameters have been exceeded. Some integrated library systems provide basic, pre-defined MIS reports in their report generators. (2) An integrated system that uses data from operational and external sources. It may include statistical components and algorithms for more complex forecasting and what-if planning used in strategic decision making, blurring the distinctions between an MIS and a DSS *which see*.

needs analysis. The process of determining the decision situations and data needs of the users of a decision support system and examining the availability of data sources.

operational database. The database that contains the data for an operational, or online transactional processing, system. The database is designed to enable rapid and secure transactions, rather than analysis.

online analytical processing. Also called OLAP. Where querying and reporting activities return static reports from a database, OLAP allows visual navigation of the values, or dimensions, in the warehouse and allows manipulation of the data for different views.

online transactional processing. Also called OLTP. Computer based processing of transactions. Operational systems are OLTP applications, and may include purchasing, ordering, creating and following work orders, cataloging collections, budgeting, circulating library materials, etc.

querying and reporting. Retrieving, sorting, formatting and viewing specified data from a database using SQL *which see*. There are several programs and MIS applications that enable the user to generate an SQL query using natural language queries or graphical interfaces.

SQL. "Structured Query Language," the standard language for querying relational databases.

CHAPTER 2

REVIEW OF THE LITERATURE

Introduction

This chapter is divided into three sections. The first section, on decision support systems in libraries, reviews some of the ways that public and academic libraries use decision support systems either for the entire information system, or to answer specific problems. The second section examines the factors in decision support planning that result in a successful data warehouse. The final section, on needs assessment, reviews methods for accomplishing the needs analysis and goals prioritization in order to establish the scope of the data warehouse.

Decision Support Systems in Libraries

Data collected for decision support may be used for information needs at three levels of decision making (Dowlin and Magrath 1982; Shelly, Cashman, and Vermaat 2001). In (1) strategic planning, the executive management defines an organization's goals and objectives, identifies the resources needed to fulfill them and sets policies to guide implementation. Finally it evaluates performance of the organization to determine corrective action, while also evaluating the external environment to determine necessary changes to the goals and objectives. In (2) management control, administrators make tactical decisions that fulfill organizational objectives. They plan programs or services, allocate resources, instruct supervisors in implementing activities, and evaluate the results for modification or correction. In (3) operational control, supervisors make day-to-day decisions on the activities or tasks that carry out tactical decisions made by administrators, and they evaluate performance for modification or correction.

With the advent of automated library systems, libraries have anticipated using their computer-based data for more efficient management at all three levels. Most of them have relied on the preprogrammed reports generated by their automated software. Researchers and librarians comment that though the ILS collects a great deal of library data, the pre-programmed reports allow only limited views of the available information, with no way to analyze the data for patterns and trends (Atkins 1996; Nicholson and Stanton 2003). Some libraries have looked for ways to make these data collections more useful by extracting the ILS data for use in a spreadsheet, database, or statistics program, either alone or in combination with data from other sources both internal and external to the library.

The library and information science literature offers some general discussions of how libraries may use their data for decision making. Nicholson and Stanton provide an overview of sources of data for decision support and the kinds of analysis that may be performed on the resulting data collections (2003). The most important data source is the ILS, which provides data that describe each item or title in the system, and then provides data that reflect the use of each item. ILS acquisitions and cataloging modules provide information on the materials owned by the library. Acquisitions data includes cost of the item and the source of purchase, order and delivery dates, and other information associated with acquiring materials. Once acquired, each item is described in its cataloging record as to title, author, format, publication, call number, subject headings and genres, etc. The circulation component of the ILS contains the patron records, checkout processes, and circulation statistics. The patron records in a public library setting generally do not contain much valuable demographic information and may be augmented by using the ZIP codes to associate the records with demographic data from the census. The checkout process associates data from item and bibliographic records with data from patron records, and generates circulation statistics. The combination of patron and item records is the primary source of patron use information in the library. Once the patron data has been stripped of personally identifiable information, the associated records can be stored in the data warehouse to be mined for information on usage trends. Items that are used in-house are not captured in standard circulation statistics. Some circulation systems allow enumeration of in-house uses. This differs from regular circulation in that the use is not associated with a particular patron. In-house use counts are kept separately from circulation statistics.

Other sources of library data include reference statistics gathered from the reference desk logs, and OPAC search logs from the OPAC server.

If the data from these systems were exported to a data warehouse they would be available for the kinds of reports required in each library setting. As Schulman points out, however, implementation of data warehousing, data mining, and OLAP techniques allows analysis that goes beyond reports (1998). Data mining can suggest patrons' needs by examining how they use the library. The same data can provide evidence on past behaviors to predict trends. This information can be used to tailor programs and refine collection development decisions. When the ILS data are combined in the warehouse with data from other library services, the library can gather usage information in even greater depth. Large data stores in a data warehouse can be mined for improved functions and better customer service by using a number of DSS tools to find new ways to look at the data (Nicholson and Stanton 2003).

Guenther addresses the difficulties of combining data from disparate sources in a data warehouse (2000). She divides data sources into two types–structured (relational databases) and unstructured (spreadsheets, word processing documents, everything else). Structured data may come from different programs with proprietary formats. Sources must be tied together with metadata, or definitions of the data, in order to standardize the field definitions of the different data types in the data warehouse. A library also needs to

determine the appropriate repository for the data. Spreadsheets may be appropriate for short term collecting and reporting, but relational databases structure the data for searching and analysis. A library may collect all of its inventories and reports to create an overall structure of fields and relationships for the database. Data may then be collected first in spreadsheets for short-term analysis, and then dumped quarterly into the database for pattern and trend analysis.

Nicholson identifies some analysis tools that may be used on library data (2003). The ILS probably provides basic management information system (MIS) capabilities, but the reports are limited and inflexible. An MIS built on top of the data warehouse allows the library to create the reports it needs, using ILS data combined in meaningful ways that are not possible with the ILS report generator, or combining the ILS data with external data the library feels is necessary for management information. If the ILS does not provide alerts, an MIS may be set up to signal appropriate managers when pre-defined limits on particular variables are exceeded. OLAP tools allow easy querying of the data and present reports that may be manipulated along different dimensions to view the data from different perspectives. The data warehouse is constructed to power the OLAP tools, running many queries before the tools are even invoked, so that response to user requests is immediate. Data mining uses algorithms to discover valid, usable patterns in large amounts of data.

Libraries will rarely reach the level of complicated data mining systems used by large corporations, but the same principles of data manipulation for improved service can be used for a library decision support system. Schulman suggests that such a system can track behavior patterns by barcode, both online and in the building, to build a history of resource use (1998). This history can be used to predict future use of library services by evaluating both past use and usage affinities. Public libraries could, for example, predict reserve list duration for best sellers, or find what traits are shared by the patrons who use the online databases. The accumulation of library information can show changes in patron behavior very rapidly if examined by computer; the task is too large and too slow to do so manually. By revealing new patterns, libraries can modify programs and collection development.

Nicholson and Stanton divide the potential uses of analytical tools into three categories: improving library services, organizational decision-making within the library, and external reporting and justification (2003). In order to improve library services, administrators may analyze how patrons use the library, looking for patterns that can be used to customize library service. Search tracking in the OPAC will allow the library to see what patrons are seeking, whether they find it, and how they use the system. Patterns in item use can help predict future purchasing, while examination of reference desk reports and ILL transactions can indicate weak areas in the collection.

In order to improve organizational decision-making in the library, administrators can monitor staff performance and decision-making. Evaluating the circulation of new items can indicate problems in the acquisitions or cataloging processes. The evaluation may uncover problems in selecting items for purchase, or assigning subject headings, call numbers, and

other finding aids during cataloging. ILL librarians may examine lenders to see if group participation aids in ILL fulfillment, or if particular lenders are faster or slower in providing loaned materials.

In order to provide information for external reporting and justification, libraries may use any of the information retrieved for other purposes. Clustering patrons by the types of materials they use can create insights into communities of library users, and demonstrate to funding agencies how their constituents use the library. Collecting data on materials use in different formats can help not only in collection development, but in budget justification.

The literature provides examples of the ways that public and academic libraries have gone beyond the bounds of their ILS to use data for management and decision making. Two public libraries have built management information or decision support systems that are completely integrated with library automation. Pike's Peak Library District began development of a decision support system in 1979 (Dowlin and Magrath 1982). It was implemented in stages, beginning with general automation of the library, and then proceeding with integration and management support. The management support stage provided data to administrators that could be aggregated or disaggregated as needed, in a framework that was comprehensible to library staff and lay persons. As of 1982 the library was preparing to add the decision support component that would integrate non-automated data on the external environment to provide support for strategic planning.

The management information system designed and built by McClure for the DeKalb County [Ga.] Public Library utilized the county's IBM mainframe computer (Mancini 1996; McClure et al. 1989). Substantial preparation for this general use system entailed interviews with decision makers, examination of the current data collections and their reports, and examination and selection of software. Data sources are NOTIS (the library's ILS), branch library activity reports, the county payroll and personnel files, and external sources such as demographics from the 1990 census. The system can combine information such as patron visits and circulation to help in deciding on branch hours and staffing. It can integrate external statistics to provide a fresh look at the changing county demographics and project future changes, thus giving new directions to library planning. It answers "what if" questions, such as the least painful way to accommodate drastic budget cuts. It helps allocate materials among the branches. This system combines daily record keeping functions with the analysis and prediction functions. It required five years to create, and must be continuously monitored and adapted.

Clark gave details on how the DeKalb MIS has contributed to library strategic planning (1993). The library system used its MIS to examine the community and establish five goals for the library. To fulfill those goals, the library adopted 14 roles, each of which is carried out by one or more branches. Examples of roles are business information center, services to senior citizens, job information center, or services to children. Each branch devotes 60% of its budget to materials for its role. Objectives quantify how each role will be executed.

These public libraries use MIS and DSS tools that are designed into the library's total automation system. Analytical tools are more often used as adjuncts to existing installations. Such implementations generally begin with a single analysis question, or related questions that may be answered from select data sources. Decision support can develop incrementally to eventually cover multiple analysis areas, as demonstrated by projects undertaken at the libraries at De Montfort University and Wesleyan University.

De Montfort University library in Leicester, UK has experimented with several decision support techniques under grants from the British Library (Adams 1995a). An initial system used library acquisitions records, national statistics, and data on library patrons to aid in resource allocation, service planning, and income analysis. A second project experimented with ways to produce performance measures in accordance with a methodology advanced by King Associates (Adams 1995b). The third system design linked several separate systems into a total decision support system. Internal and external data were operated on with several tools to create profiles of groups of library patrons. This has had wide implications for study during times of rapid change in educational methods and library services.

Wesleyan University Library (Middletown, CT) has instituted a number of statistical analysis projects, each designed to answer a particular management problem (Cheng, Bischof, and Nathanson 2002). Collection usage statistics yield information on patron use of journals, e-journals, databases, and general circulating materials. These data provide information in different areas of library management. The budget for materials acquisitions may be re-allocated by examining circulation statistics for newly acquired materials by budget categories, to see what percentage of the new materials in each category never circulated. Serials acquisition policy will be influenced by data collected on preferred journal formats as the library moves toward a policy of acquiring each journal only online or in print. To determine patron preferences in journal format, circulation and in-house use statistics for print editions are compared with access data for their corresponding e-journal formats. In examining database use, the library realized that many patrons bookmark their favorite databases rather than accessing them from the home page. The library realized it would need to develop additional means of information patrons of database changes besides posting notices on the home page. Reference service statistics are being used to evaluate new reference tools. When reference statistics dropped, the library received a grant to organize a consortium with other colleges to provide 24-hour live digital reference. Once the grant period ends, statistics will indicate whether the consortium approach works for live digital reference. Reference services statistics now combine data from the reference desk, live digital reference, phone, and e-mail, and taken together will indicate the cost of reference services and the cost effectiveness of each type of service.

Decision support systems, data warehouses, and statistical analysis projects frequently address a specific problem. This includes prototypes, trials, and models, which may initially analyze a single situation, but with the intent of developing a system that will also address other needs. Budget allocation, including serials selection and acquisitions efficiency, is the

most frequently named area of analysis. Other systems examine such areas as ILL transactions or reference services.

Ottensmann, Gnat, and Gleason report on a one-time statistical study by the Indianapolis-Marion County Library in which circulation rates were used in combination with the U.S. Census to determine how strongly demographic characteristics are related to circulation patterns at each of its twenty-one branches (1995).

Three libraries have proposed or tested decision support systems, using data from the ILS, that are intended for use in a variety of decision making situations. The first two of these projects use only data from the ILS, while the third incorporates data from other sources. The first project, at the George A. Smathers Library, University of Florida, proposed the development of a data warehouse to support data mining techniques for management decision making. The goal was to demonstrate a method of building a decision support system using inexpensive software for the different needs of each library unit (Su and Needamangala 2000). The first phase of the project was to create the warehouse, the user interface, and query tools. The second phase, to be undertaken later, will be implementation of the data mining tools. The pilot project for the first phase was run on a PC with Microsoft Access for the warehouse, using data on 20,000 monograph titles. The data came from several NOTIS files that could not be examined together in the ILS. Database usability was demonstrated by running SQL queries related to acquisitions efficiency. Queries established the circulation rates of materials purchased under different plans, and the processing costs of materials that did not circulate. The purpose of the project was to develop a system that was flexible and inexpensive, and that could be used in different units of the library for their own decision processes.

In the second project, Gleeson and Ottensmann hypothesized an "ideal data set" derived from circulation and cataloging records that would allow examination of the library's collection by age by year by category (1993, 94). The categories could be budget categories, age level collections, adult fiction or non-fiction, or other desired groupings. Information derived from the data could be used in budget allocations, personnel scheduling, or facility location and design. CLSI, the integrated system in use at both the Indianapolis-Marion County Public Library and the Monroe County (IN) Public Library, contained all the necessary data, but the report generator could not create some of the desired reports such as those requiring frequency distributions, or evaluation by age of items. A decision support system was developed and given a trial run at the Indianapolis-Marion County Public Library on a limited data set based on Quattro Pro (Ottensmann and Gleeson 1993). The library had a complex budgeting system for print materials, and librarians tested different allocation models using circulation information. The librarians learned to refine the models further and use the system for developing alternative budget proposals. With this system the library also had better predictive information on circulation and stock, which allowed forecasts for personnel and collection development, and even modifications to the weeding policy.

In the third project, the Kansas State University Libraries developed a prototype decision support system based on a data warehouse integrating data from multiple sources with ILS data. The eventual goal is evaluation of books, periodicals, and electronic services using OLAP and data mining. (Bleyberg et al. 1999; Cole, Somers, and Emery 2001). The prototype was limited to serials, with data drawn from the integrated library system, human resources, the student information system, inter-library loan, an online database, and an online document delivery service.

Decision support systems may also be designed specifically to address a single management question or decision situation. Serials reduction in academic libraries is addressed a number of times in the literature. The Sterling C. Evans Library at Texas A&M University set up a program to examine data from NOTIS in order to effectuate a $464,000 reduction in serials subscriptions (Atkins 1996). The NOTIS data were combined with a usage study for the initial analysis, using an SAS program. The analysis of the data enabled the library to work with teaching faculty to reduce the number of titles by 1,100. It was so successful that the library now runs the same SAS report annually (without the usage study) to keep abreast of serials costs. Another SAS program was written using the same data to provide library information to the North American National Title Count in 1993. The SAS program required four days to create, but data generation requires less than two hours. The information derived from this study gave the library insight into its strengths and weaknesses, and the library also runs this program annually to tracks its growth.

At Wichita State University, serials cancellations also brought about attempts to capture journal use electronically via barcode scanning (Dadashzadeh, Payne, and Williams 1996). This usage capture coupled with other serials information derived from NOTIS led to the development of the Periodicals Analysis Database (PAD) Project. This decision support system has been used over three serials reduction cycles to maintain the strength of the collections while holding down costs. The information is maintained in a Paradox database. It has proven its worth by making collection development librarians more aware of cross-disciplinary titles while making faculty more aware of the difficulties of maintaining subscriptions as prices increase. The library has also discovered the fiscal trade-offs of document delivery. The collection of all this information is quick and reliable. Future implementations call for the addition of a "what if" simulation component that could streamline calculations based on further considerations.

The University Libraries at Virginia Tech (Virginia Polytechnic Institute and State University) also constructed a decision support system to better manage serials, with the potential for use in serial subscription cancellations (Metz and Cosgriff 2000). The system eventually incorporated results from a faculty survey, reshelving data, ILL data from ILLiad, data from CARL Uncover Reveal, ISI Journal Citations Reports, and general information from the serials vendor. The data were collected in an Excel spreadsheet, seemingly in a single spreadsheet rather than into connected spreadsheet modules. The spreadsheet was large enough that it was split into two sections for different needs. The reduced data set was used to identify and price the most important titles. The complete database is still available

for collection development decisions. Future efforts will include a redesign of the data for improved flexibility, and finding ways to keep the information current when the underlying data sources vary in their costs and update cycles.

Data from sources other than the ILS may also be analyzed alone for decision support. The Winona State University [MN] Library examined reference desk reports to determine optimal staffing of the reference desk (Dennison 1999). Reference desk logs showed the number of questions asked, and their degree of difficulty, in each hour. Reports for one year were statistically analyzed, and showed that the desk was unnecessarily double-staffed a few hours each week. Other times were shown to be understaffed, and reference desk staffing could be adjusted to meet use. At the University of Tennessee, Knoxville, OCLC ILL Management Statistics were downloaded each month and imported into an Access database file (Hammons 1999). Reports were generated for ILL management, including indicators on net borrowers and lenders, and average turnaround times for lending libraries. A collection development report was added later to provide pertinent information to the collection development librarian.

Some of these decision support systems are concerned with data from single sources, or from similar sources such as multiple files from a single ILS, while others are concerned with more complex data assemblies. The system-wide DSS and MIS in two public libraries took several years to complete, and need continual adjustment (Dowlin and Magrath 1982; McClure et al. 1989). Decision support systems designed to answer multiple questions from a unified data warehouse require careful consideration of data sources, and present challenges in acquiring and cleaning the data for the warehouse (Bleyberg et al.; Su and Needamangala). There are only two libraries that reported on multiple, completed decision making projects: Wesleyan University which does not appear to totally integrate the data used for all projects, and De Montfort University whose projects are funded by the British Library (Adams 1995a; Cheng, Bischof, and Nathanson 2002). These facts suggest the complexity and expense in creating a data warehouse for decision support. Libraries undertaking similar projects will want to investigate recommended procedures for creating a successful data warehouse.

Factors in Creating a Successful Data Warehouse

Creating a data warehouse and its attendant decision support tools is a complicated and expensive undertaking. It is estimated that at least 50% of data warehousing projects—perhaps more—result in failure, though many second attempts are successful (Mukherjee and D'Souza 2003).

There is considerable discussion in the literature over what causes a DSS project to fail. Mukherjee and D'Souza discuss factors in both failed and successful implementations, based on their own experience and on reports on these topics in the literature. They found that the organizational factors in failed implementations included "the lack of an executive

sponsor, the lack of business objectives, the lack of user involvement, and organizational politics." (2003, 83).

When data warehouses do not match user needs or business goals, they ultimately fail. User involvement was also an important factor in successful implementations. During the planning stage, user needs should be established and consensus reached on warehouse goals. This is called setting the scope of the project. Because there are usually competing goals and needs within a business, establishing the scope is usually difficult, which is why it is crucial to have agreement on the ranking of the business needs and selection of the scope. To omit this agreement is to risk failure (Kudyba and Hoptroff 2001). It is usually at this stage that a company finds it needs to review its goals and objectives, and define success factors for the IS project. (Joshi and Curtis 1999).

When even limited projects do not have a specific audience to begin with they must eventually be retooled. For example, the PAD project at Wichita State University turned out information that was useful only for library administrators, but was useless for collection developers. Their needs, along with some faculty criticism, brought about the development of a collection development component that made this system more useful (Dadashzadeh, Payne, and Williams 1996).

McClure's report on the design phase of the MIS for the DeKalb County Public Library states that the project was initially confounded by "competing objectives for an MIS, the difficulty of determining information needs, the value of user involvement, and the need for a supportive organizational climate." The focus of the project was changed due to information found during focus group interviews and the audit of current information gathering methods (McClure et al. 1989, 194).

The critical points in the development of a warehouse or DSS have been developed by observation. There are a few studies, however, that examine actual causes of data warehousing failure. Wixom and Watson (2001) created a data warehousing success model, and tested it using survey responses from 111 organizations. They found that user participation was an important factor in planning the data warehouse, where their input kept the project focused on their needs. The authors theorized that since many large projects are now broken down into creation of smaller data marts, the users in each department have an opportunity to voice their needs in planning, while having to participate in only their own part of the overall project.

Payton and Zahay (2003) present a case study of a regional health care payer and its data warehousing implementation. The company had created a corporate warehouse, then realized it was not being utilized by the marketing division. The researchers conducted focus group sessions with marketing and IT employees to find out why. Among the results, two facts emerged: (1) standard analyses performed by marketing applications were not supported by the data in the warehouse, and (2) there had been no provision for historical data–a

prime prerequisite of CRM, which examines customer history. The needs of the users had not been considered in constructing the corporate data warehouse.

Needs Assessment and Planning for a Successful Decision Support System

None of the library and information science literature provides details on needs assessment and planning for a data warehouse; however, two research papers cite methods published by Kimball (Bleyberg et al. 1999; Bonifati et al. 2001). Even though the techniques described in these sources are intended for large companies that are often geographically dispersed, emphasis is still placed on focus group sessions and in-person interviews with actual users as the most appropriate methods for obtaining their perceptions of their data needs (Kimball 1996; Kimball et al. 1998). Once interviews are complete the responses are compiled. Then they must be prioritized, and consensus reached on initial decision needs that will be addressed by the decision support system.

Data sources are assessed at the same time the business requirements are being collected. This step is designed to provide a sense of the feasibility of undertaking some of the requirements collected from the interviews. This occurs concurrently with user requirements assessment, so that the two lines of investigation can be more readily converged into a data model.

Italian researchers proposed a method for requirements gathering that they based on Kimball's techniques (Bonifati et al. 2001). This method uses three stages for business needs analysis. The first examines the needs of the users, the second examines the design of the data in the potential operational database, and the third combines the two where they fit appropriately. The user requirements are collected in several focus group interviews, using the GQM paradigm (Goal/Question/Metric) designed at the University of Maryland. The interviews produce dense descriptions of the user needs, which the researcher and the users together compress into several needs statements that are ranked by the users. The researcher creates idealized star schemas, or dimension descriptions, from these statements before going on to examine the data. The operational database structure is examined and used to create another set of dimension descriptions. This process involves an automated graph analysis technique that produces many candidate star schemas. In the third stage, the idealized and the actual schemas are compared, and the best fits are selected for potential implementation. The researcher selects the final statement for implementation based on the earlier rankings by the users.

The prototype decision support system at the Kansas State University followed Kimball's data integration techniques to assure that their small prototype system would integrate with future data warehouse implementations (Bleyberg et al.). This implementation began with the data and then moved to determining the decision situations that the prototype would support. Before creating the tables for the data warehouse, however, the programmers met extensively with the librarians who would use the system. It was this

collaboration that also provided the standardized data definitions that were intended to unify future warehouse implementations (Cole, Somers, and Emery 2001).

Personnel considerations are important in planning and implementation of decision support and data warehouses. This factor is not addressed directly by the resources on data warehouse planning, which assume that teams of business and technology professionals are in charge of such projects. In library settings this is not usually the case, and three sources commented on how personnel choices affect the success of decision support projects. All three situations emphasize the importance of employing experienced personnel for data warehousing projects who can accomplish the necessary tasks efficiently.

At Kansas State University, the prototype planning and development was originally a loose collaboration among three departments, using the services of student interns (Cole, Somers, and Emery 2001). This arrangement was not productive. The University has since committed to development of the data warehouse with full time employees, under the direction of the library.

When the Texas Tech library initiated statistical analysis of its serial data, it hired a statistician to write the required program, which cost about $1,000 (Atkins 1996). The library knew it had qualified personnel, but realized they were unlikely to have time to devote to a new project. The programmer completed the project in a week. Each year's work requires less time and expense to complete.

Wesleyan University library discovered that data management quickly becomes an issue as the volume of collected data increases (Cheng, Bischof, and Nathanson 2002). Furthermore, data collection is useless without any knowledge of how to select and analyze the data, and interpret the results. Wesleyan hired a librarian with a background in statistical analysis to manage its projects.

CHAPTER 3

RESEARCH DESIGN

Introduction

The purpose of this study is threefold: (1) to understand data needs for better decision making in public libraries; (2) to prioritize these data needs; and (3) to suggest schemes and methods to reconcile data from original diverse sources for building a data warehouse to support these needs. Interviews with nine administrators in the two top levels of the organizational structure were conducted in two phases. The first phase interviews explored the types of decisions made in public libraries and the data elements needed to support those decisions. The second phase interviews examined select data elements and explored how they could be used in decision situations.

Rationale

A data warehouse by nature is application-specific. It is tailored to a small group of well-defined users with decision related needs. Its success depends on understanding its users' needs and the types of decisions the information will support. This applied research is a user study specifically oriented to a data warehouse for an individual library. The users of the data warehouse are identifiable as the members of the top two levels of library administrators. Their data needs as they perceive them will form the basis of the data model.

Participants and Selection

Participants of the study were nine persons from the top two administrative tiers of the library system. They were selected purposively, as being all those members of the library administration who implement the long-range planning decisions for the entire library system, under the direction of the Board of Trustees, or who plan and evaluate programs and services in their particular departments that will fulfill the library's objectives. Examination of the administrative structure and discussions with the two appointed library contacts identified the nine persons who met these criteria. These are the director and deputy director at the top level of administration, and the branch coordinator, outreach services coordinator, adult services coordinator, youth services coordinator, collection development librarian, system administrator/technical services coordinator, and network manager at the second level of administration.

Data Collection

Data collection for this study was the interview method. Two rounds of interviews were conducted with all administrators to establish their *decision situations* and their *data needs* for those situations. All interviews were conducted in the administrators' offices or in the staff conference room and were tape recorded and transcribed. A summary of the interview was e-mailed to each administrator for approval or corrections, whereupon the tape recordings were destroyed. Specific information on data collection for each round of interviews will be found below in the descriptions of each interview phase.

Instrument Development for Phase One Interviews

The phase one interviews were semi-structured, designed to elicit information on what kinds of decisions the administrators make, and their own perceived data needs for making those decisions. The first three questions of the Phase One Interview Guideline (Appendix A) allowed the administrator to identify herself, state her position and her immediate supervisor, and list her official duties and examples of special assignments. Questions 4 and 5 prompted administrators to recall instances in which they had trouble obtaining needed data, and then to state where they eventually obtained the data, if at all. Questions 6 and 7 offered administrators the opportunity to expand on their perceived data needs, while keeping their suggestions related to practical use.

Pilot Study Phase One

A pilot study was conducted prior to the interviews. Participants, from the second and third tiers of administration, were other than those who participated in the actual interviews. Necessary adjustments were made to the interview guideline based upon responses in the pilot study.

Data Collection Phase One

Participants received a copy of the Phase One Interview Guideline in advance of the interviews to give them time to consider their answers.

The interviews were conducted with the nine administrators in six sessions in February and March 2003. Each interview lasted about an hour. Of the nine interviewees, four were interviewed one on one; two administrators were interviewed together because they share knowledge of the library's online databases; three were interviewed together because two work as a team and the third had scheduling constraints that only permitted her availability at this time.

Data Analysis of the Phase One Interviews

Coding of the phase one interviews grew out of the analysis of the data. There were two ultimate analysis goals: to establish *decision situations* and *data needs*. These required different types of analysis. Decision situations are not pre-defined; they will vary with the needs of the institution and the administrator. The examples given by different administrators were coded as *decision instances* which showed similarities that allowed them to be grouped into decision situations.

Data elements are defined entities, so data needs expressed in the interviews could be analyzed without coding.

Two further concepts, *information needs* and *likely data sources*, were developed during analysis of the data. The information needs concept, used alone or in combination with incomplete data needs, contributed to the identification of likely data sources, and was not used in further analysis in the study.

Interviews were transcribed and examined for statements on decision instances, information needs, and data needs. The transcriptions of the nine interviews were read once for general meaning, then a second time to mark passages in which these concepts appeared. The marked passages in each interview were then read again to determine interview statements, each of which should have consisted of a decision instance, an information need, and specific data needs. Twenty-eight statements were identified (Appendix B), and thirteen of these contained all three concepts. Decision instances and information needs concepts that were not specifically stated could be deduced either through strong implication in the conversation, or by comparison with related statements from other administrators with similar interview statements. The interview statements were read a final time to determine likely data sources.

Decision situations

Decision situations describe the purposes for which the provided information will be weighed. The interview texts were scanned for the decision instances, which were located by looking for a phrase that answered why the administrator sought particular data or information. Decision instances were coded for decision situations.

Twenty-six of the twenty-eight interview statements contain decision instances. Two statements reflect instances in which the administrators need easier access to data for reports or for general reference, but that do not lead directly to making decisions. The twenty-six decision instances were grouped into fourteen decision situations that fall into the following five broad categories:

S1. Examine library service area. This category includes decision situations that require information on population or residential growth changes in the library's service area.

S2. Find patterns in patron use of materials and services. This category includes decision situations that require information, direct or implied, on how patrons use the library and its services. Non-identifiable patron data are important components of these decision situations.

S3. Evaluate non-circulation usages. This category includes decision situations that require information, not captured in circulation statistics, on access to and use of library materials.

S4. Frequency evaluation. This category includes decision situations that require frequency analyses of the library collection.

S5. Miscellaneous reports. This category includes operational reports normally generated by the ILS, but that are not available in Horizon.

Table 3 contains the decision situations and decision instances for the phase one interviews. Statement 16 provides an example of extracting a decision instance and decision situation from the interview text:

> *I'm interested in use patterns in terms of time frames–years, months, we're busy then, we're not busy other times, we're busy the week after Christmas–just really in terms of staffing.*

The decision instance can be expressed as examining patterns of services use for staffing decisions. This decision instance is similar to others that are coded under decision situation S2, "find patterns in patron use of materials and services."

Three interview statements were worded so that instances had to be deduced from the conversation. For example, statement 10 asks:

> *When people come to use the technology, are they using the rest of the library? . . .When they come to a program are they using other services? . . . What brings them in here? What else would they do if it were available to them?*

Comparison with other statements shows that the interviewee is concerned with targeting services by determining patterns in how people use services and collections.

Data needs

Data needs describe the particular entities and attributes that an administrator needs in order to understand a situation. Data, in the context of their relationships, provide information for decision making. None of the interview statements provided enough data detail to construct a model for a data warehouse. For example, statements 10, 14, and 15 each concern patterns in use of library services. It is not clear whether the administrators

Table 3.
Phase One Decision Situations and Decision Instances

Situation Code	Decision situation	Decision instance
S1		*Examine library service area*
S1.1	area demographics	demographics of underserved areas to determine what library is not offering them
S1.1	area demographics	planning future programs and services for growing groups that are not using library services
S1.1	area demographics	determine direction of outreach services to community at large
S1.1.1	area demographics–marketing	identify a community segment for marketing
S1.1.1	area demographics–marketing	identify a community segment for marketing
S1.2	residential development for branch planning	branch planning
S1.2	residential development for branch planning	branch planning
S1.2	residential development for branch planning	branch planning
S2		*Find patterns in patron use of materials and services*
S2.1	define patrons, target services	targeting services
S2.1	define patrons, target services	determine what services patrons use when they visit the library, target services
S2.1	define patrons, target services	evaluate usership of resources
S2.1	define patrons, target services	gain a detailed snapshot of the patron base–don't want to lose
S2.1	define patrons, target services	tailor services to those who use library; set hours, buy books to support needs, schedule programs for best times, etc.
S2.1	define patrons, target services	identify market, tailor to it as an efficient use of resources
S2.2	staffing	staffing
S2.2	staffing	staffing
S2.3	technology planning	technology planning
S2.4	effects of web marketing	evaluate web marketing
S2.5	gap analysis (who does not use the library)	look at who uses the library to find demographic groups who do not
S2.6	determine most effective physical layout	shape physical layout for most effective use

Table 3. Continued.

Situation Code	Decision situation	Decision instance
S3		*Evaluate non-circulation usages*
S3.1	iPAC search tracking to improve catalog	suggest cross references to catalog to improve search results
S3.1	iPAC search tracking to improve catalog	does the library [catalog] have a path that leads people to what they seek?
S3.2	online databases	evaluate usage of online databases
S3.3	in-house materials	evaluate in-house use
S4		*Frequency evaluations*
S4.1	turnover rate of individual collections	evaluate materials use for collection development
S5		*Miscellaneous reports*
S5.1	purchase alerts	ratio of hold requests to copies owned per title

mean circulation, in-house use, or online database use, or use of services such as public Internet computers or reference desk help, or other activities such as classes, programs, and meeting room use.

Had the data needs been expressed in detail, they would have been grouped by data source, entity and variable. Since only seventeen of the twenty-eight information statements specified data needs at any level, a second round of interviews was designed to examine data elements from a limited number of sources and discover how they could be used in library decision situations.

Information needs

Information needs describe the knowledge that an administrator expects upon the accumulation of particular data. In the second scanning, words and phrases that identified information needs were highlighted first. These phrases were usually introduced with "I need to know," "I could use," "I do not have," or "I wish I had more information on." Only one interview statement lacked an information need. Statement 9 reports:

> *In demographics, I believe that age, in particular, is important. The library has taken steps to collect birth dates [at registration] instead of grade in school; this should eventually help in targeting services.*

In this statement, "demographics" was used as a partial information need, with a single data need of "age" and a decision situation of "targeting services." Statement 13 says:

I am interested in the demographic of the user . . .. In the end this will help us tailor services to the people who come here . . .. When you register, I want to know who you are, how old you are, do you go to school, do you work out of the area."

The parallels between these two statements indicated that the administrators were speaking of the same information need, and "demographic of the patron" was assigned to statement 9.

Likely data sources

Likely data sources indicate potential sources of data for library decision making as derived from the phase one interview statements. The information needs and data needs for each interview statement identified–or at least suggested–sources of data. These sources were coded into three categories of likely data sources: the ILS, library sources other than the ILS, and external sources. Aggregated results will be found in chapter 4, table 5.

The ILS is based on the library's bibliographic database, and defines the data elements with uniform descriptions to integrate separate modules that support acquisitions, cataloging, serials, and circulation processes. Eighteen statements either named the ILS directly, or mentioned data elements or processes associated with it. In statement 13, the administrator said,

When you register here I want to know who you are, how old you are, do you go to school, do you work out of the area.

This information, taken at registration, is found in Horizon's patron records (part of the circulation module).

Library sources other than the ILS are those files and reports maintained by the library that quantify usage of library services. Two services that were specifically identified were public Internet computers and programs. Other data sources mentioned were the online database vendor reports, and web logs from the library's web server and iPAC server.

External sources are those data sources that are not owned by the library. Only two were specified–tax rolls, and the 2000 Census.

Example of phase one analysis

The following excerpt from the phase one interview of Administrator 1 is followed by the interview statements found in examining the transcript, and a demonstration of how the decision situations and data needs were categorized.

I tend to look at things more specifically than [Administrator 7] does. I know I have always wanted more information about where people live who use this library–where they come from. Where the developing areas are. That's probably something we could get from [the area development commission]. I think that where the growth is happening to help us plan library services is something we need to be more aware of.

In demographics, I think age, in particular, is important. We have taken steps to get this information better. Instead of asking for grade in school [at patron registration] we are taking birth dates. That will help us target our services and programs better.

Staffing is another thing I'm interested in, and [Administrator 7] has mentioned that. When are the peak times, and what services are they using at those times, so that we can staff those areas better.

When people come in to use the technology, are they using the rest of the library? Is there any correlation there? When they come to a program are they using other services? I don't know how we would go about getting that kind of information. What brings them in here? What else would they do if it were available to them?

One recent example of something I wish we knew concerned the Gods and Generals *programming. It's always been our intuition that more women use the library than men, and we wanted to have something that appealed to men, to bring them into the library. But that's based on gut feeling and intuition. If we knew gender and ages and what they came to the library for it would be easier to market.*

All these would help in planning, from staff to collections to branches.

Five interview statements were extracted from this interview and analyzed for decision situations (derived from decision instances), data needs, information needs, and likely data sources.

1. Where do people live who use the library. Awareness of where growth is occurring will help in planning branches. (Statement 6.)
    a. derive decision situation
        i. decision instance: branch planning
        ii. decision situation: S1.2–examine library service area–residential development for branch planning
    b. data needs: where patrons live (address or ZIP code in patron record)
    c. information need: where people live who use this library, where they come from
    d. likely data source: ILS

2. Where are the developing areas; this information is probably available from [the development commission]. Awareness of where growth is occurring will help in planning branches. (Statement 7.)
    a. derive decision situation
        i. decision instance: branch planning
        ii. decision situation: S1.2–examine library service area–residential development for branch planning
    b. data needs: not specified
    c. information need: where the developing areas are
    d. likely data source: external, probably the development commission
3. In demographics age is important to help target our services and programs better. (Statement 9.)
    a. derive decision situation
        i. decision instance: targeting services and programs
        ii. decision situation: S2.1–find patterns in patron use of materials and services–define patrons, target services
    b. data needs: age
    c. information need: demographic of the patron
    d. likely data source: ILS
4. Staffing–when are the peak times, and what services are patrons using at those times. (Statement 15.)
    a. derive decision situation
        i. decision instance: staffing
        ii. decision situation: S2.2–find patterns in patron use of materials and services–staffing
    b. data needs: time
    c. information need: peak times of service demand, and what services patrons are using at those times
    d. likely data sources: ILS, other library sources
5. When people come in to use the technology or attend programs, are they using the rest of the library? Is there any correlation there? What brings them to the library and what else would they do if it were available to them? (Statement 10.)
    a. derive decision situation

i. decision instance: determine what services patrons use when they visit the library, target services

ii. decision situation: S2.1–find patterns in patron use of materials and services–define patrons, target services

b. data needs: not specified

c. information need: when patrons visit the library for one service, what other services do they use, and what other services would they respond to

d. likely data sources: ILS, other library sources

Instrument Development for Phase Two Interviews

Although the library has other operational databases, the administrators referred only to the ILS during the first round of interviews. They alluded to external government data sources, but gave no comprehensive details on the data they wanted from those sources. They also mentioned other library data sources without specifying what kinds of usable data were in these sources, or whether the sources are presently in electronic format.

When the administrators spoke of Horizon data their statements made references to patron, circulation, and item entities. Since their comments were focused on a single, well-known and accessible source of data, the second round of interviews was designed to elicit how specific data elements from the ILS could be used to answer the administrators' decision situations. The Phase Two Interview Guideline (Appendix C) used elements selected from the patron and item tables to reflect data needs regarding patrons, item descriptions, and item circulation statistics. Circulation was expanded to include in-house use in order to capture statistics on items that are used but not checked out, in particular reference materials and periodicals. The second interview also included elements from online database vendor reports. Although these data do not originate with the ILS, they complement the in-house use statistics by indicating usage of online editions of reference materials and periodicals.

Data elements from the Horizon tables such as patron age, patron ZIP code, item record creation date, and item title were selected because they were either identified in phase one interview statements, or their potential use was at least understood in a decision situation. Additional data elements such as item collection code, item Dewey number, item branch home, and item publication date were chosen because they appeared in two informational items from the spring 2002 data mining workshop: notes made by an administrator on the slide program handout, and dimension tables sketched on easel paper during a discussion session. Data elements for the questions concerning online database use were selected from examination of a six vendor reports for April 2003.

In the interview guideline the data elements were classified into three sections:

1. patron data, which asked about usability of certain fields from Horizon's patron records
2. library materials usage, which asked about data related to
    a. the circulation of items
    b. the description of items
    c. the in-house use of items
3. online database usage (called electronic resource usage in the interview guideline), which asked about data fields related to usage of the library's online databases

An open question at the conclusion of each section allowed administrators to address additional data elements not covered in the questions. Printouts of data views for item and bibliographic records were provided to aid in recall of the many available data elements.

Pilot Study Phase Two

A pilot study was executed prior to the second round of interviews. Participants were two branch managers. Minor wording changes were made for clarity of some field descriptions.

Data Collection Phase Two

Phase two interviews were conducted one on one with each of the nine administrators during July and August of 2003. Interviews took 30–45 minutes. At the beginning of each interview, the administrator was reminded of the decision situations she had provided in the first interview. She was asked to select one or two for more in-depth consideration, and also given the option to discuss a new situation if she preferred. Administrators were not asked about data elements that did not pertain to their decision situations. Only one administrator requested the full review of every data element in the interview guideline.

The interview with one administrator failed to record on tape. The interviewer had taken notes at all sessions should such an event occur, and the content of the interview was reconstructed from the notes. The administrator was apprised of the error, and a copy of the reconstructed interview was sent to her. The interview summary was sent immediately afterward, with more detail than summaries sent to other administrators. The administrator approved the reconstructed interview and interview summary.

Data Analysis of the Phase Two Interviews

The second interview focused on data elements and asked how they could be used in decision situations. The nine administrators discussed fifteen situations in which they were required to make a decision (Appendix D). Three interviewees identified two decision situations each that they wished to discuss. Six administrators identified a single decision situation. However, when asked what additional data they required, three of these six identified data sources related to an additional decision situation.

Phase two decision situations reiterated eight of the fourteen decision situations from the phase one interviews, and introduced seven new ones (chapter 4, table 4): four related to S2 (find patterns in patron use of materials and services), two related to S4 (frequency evaluations), and one related to S5 (miscellaneous reports).

The fifteen decision situations from the second phase were organized in a table to find their common data elements. The decision situations are listed in order of decision situation codes. Five of the fifteen decision situations do not use any of the data elements from the phase two interview guideline; administrators specified other data sources for those decision situations, as noted below.

During the interviews the administrators were asked if they required data elements other than those discussed in the interview guideline. They identified eight Horizon elements and one online database vendor report element as additional data needs.

Administrators also listed fourteen other library data sources and two external data sources that they wished to see included; they did not identify the data elements needed from these data sources. The data elements from the interview, the additional data elements and the additional data sources requested are reported in chapter 4, tables 6–8. Some decision situations require few data elements, and others require many elements from a variety of sources; these are summarized in chapter 4, table 9.

Decision situations and their supporting data for the data warehouse

Decision situations that depended entirely or primarily on the data elements from the ILS and the online database vendor reports were analyzed for inclusion in the data warehouse. Additional data elements from these two sources that had been identified during the interview were incorporated with the data elements from the interview guideline for the purposes of this analysis (chapter 4, tables 10–13).

The decision situations that required data from other library sources and external sources were analyzed for future implementation in the data warehouse.

CHAPTER 4

RESULTS

Introduction

This chapter presents the results of the two rounds of interviews in three sections, corresponding to the research questions: (1) What are the decision situations encountered by library administrators? (2) What data do public library administrators need to support their decision making? (3) How may the necessary data be obtained and organized for decision support?

Decision Situations

Administrators identified twenty-one decision situations over the course of two rounds of interviews. There were fourteen decision situations in phase one. In phase two, the interviewees were reminded of their earlier interview statements, and asked to consider one or two decision situations for the second interview. They were given the option of naming another decision situation if there were new concerns. Fifteen decision situations were discussed in phase two; of these, eight had already been identified in phase one and seven were new. Decision situations from both interviews are aggregated in table 4.

During the phase one interviews, administrators tended to identify decision situations outside their sphere of responsibility. Because they confer frequently, each is aware of situations in which data is needed by other administrators. There are also situations in which information is collected by second tier managers, and then passed on to the director and deputy director for decision making; in these instances the same decision situation might be discussed by more than one person. For these reasons, the number of occurrences of each decision situation in phase one is not a factor in analysis.

Only the fifteen phase two decision situations were analyzed for decision support, as the second interview format was designed to associate them with clear data needs.

Table 4.
Summary of Phase One and Phase Two Decision Situations

Code	Decision Situation	Phase I	Phase II
S1	*Examine library service area*		
S1.1	area demographics	3	1
S1.1.1	area demographics–marketing	2	1
S1.2	residential development for branch planning	3	0
	sub-total	*8*	*2*
S2	*Find patterns in patron use of materials and services*		
S2.1	define patrons, target services	6	1
S2.2	staffing	2	1
S2.3	technology planning	1	1
S2.4	effectiveness of web marketing	1	1
S2.5	determine most effective physical layout	1	0
S2.6	gap analysis (who does not use the library)	1	0
S2.7	collection development	0	1
S2.8	youth collections	0	1
S2.9	youth programs	0	1
S2.10	web support of school assignments	0	1
	sub-total	*11*	*7*
S3	*Evaluate non-circulation patterns*		
S3.1	iPAC search tracking to improve catalog	2	1
S3.2	online databases	1	1
S3.3	in-house use of materials	1	0
	sub-total	*4*	*3*
S4	*Frequency evaluations*		
S4.1	turnover rate of individual collections	1	0
S4.2	branch collections	0	1
S4.3	evaluate library against comparable libraries	0	1
	sub-total	*1*	*2*
S5	*Miscellaneous reports*		
S5.1	purchase alerts	1	0
S5.2	discarded items reports	0	1
	sub-total	*1*	*1*
	Total	26	15

Data Needed to Support Decision Situations

Phase one interviews

In twenty of the twenty-eight interview statements (including the two that are not decision instances), data needs were indicated to some extent either by example or through a partial listing of data elements. One administrator needed to know "where people live" but did not indicate whether this would include street address and city fields, or ZIP code fields carried to five, seven, or nine digits. Data sources from within the library (besides the ILS) were identified both generally ("services") and specifically ("programs," "online database vendor reports"). External sources, when named, were guesses as to where the data might be found. Data for each of the twenty-six decision instances could be from the ILS, other library sources, external sources, or a combination of the ILS and either of the two other sources. Likely data sources, as derived from the text of the interviews, are summarized in table 5.

Phase two interviews

Interviewees were asked if they needed data other than those identified during the interview. They named eight additional attributes from Horizon, and one additional element from the online database vendor reports. They also asked for data from fourteen library data sources (other than the ILS), and from two external sources.

Table 6 demonstrates how each decision situations requires data from the ILS; the data elements are from the interview guideline and from interviewee requests. Table 7 demonstrates how each decision situation requires data from the online database vendor reports, using both the elements named in the interview guideline and in interviewee requests. Table 8 shows the additional data sources that were requested. Table 9 summarizes the data sources for the phase two decision situations. It indicates how many variables are required from the ILS, and how many data sources are required from other library sources or external sources.

Table 5.
Occurrences of Data Sources, Phase One Interviews

Data source	Occurrences
Data from ILS only	12
Data from other library sources only	4
Data from external sources only	3
Data from ILS and other library sources	4
Data from ILS and external sources	3

Table 6.
Phase Two Data Needs from the ILS

Decision Situation	Patron–age	Patron–school	Patron–9-digit ZIP code	Patron–work phone	Patron–branch	Circ–Time/date cko	Circ–date cki/overdue	Circ–last checkout	Circ–total checkouts	Circ–renewals	Item–creation date	Item–collection code	Item/Checkout–locations	Item–call number	Item–in-house use	Item–status	Item–publication date	Item–genre	Item–subjects	Item–author	item–title	Item–ISBN	Request table data
Area demographics																							
Area demographics–marketing																							
Define patrons, target services	X	X	X	X	X		X	X	X			X	X	X			X	X	X	X	X		
Staffing	X					X			X			X	X	X	X				X				
Technology planning	X					X			X				X										
Effectiveness of web marketing	X					X						X		X					X	X			
Collection development	X	X	X				X	X	X			X	X	X	X	X	X	X	X				X
Youth collections	X	X				X						X	X	X									
Youth programs																							
Web support of school assignments	X	X				X						X		X									
iPac search tracking to improve catalog																							
Online databases																							
Branch collections								X	X		X	X	X	X	X	X		X	X				
Evaluate library against comparable libraries																							
Discarded items reports												X	X	X		X			X	X	X	X	

Table 7.
Phase Two Data Needs from Online Database Reports

Decision Situation	OnlineDB–sessions	OnlineDB–duration	OnlineDB–remote/library use	OnlineDB–denied access	OnlineDB–printed	OnlineDB–IP address	OnlineDB–branch use	OnlineDB–time/day/date of use
Area demographics								
Area demographics–marketing								
Define patrons, target services	X	X	X	X	X	X	X	X
Staffing								
Technology planning								
Effectiveness of web marketing								
Collection development								
Youth collections								
Youth programs								
Web support of school assignments								
iPac search tracking to improve catalog								
Online databases	X	X	X	X	X	X	X	X
Branch collections								
Evaluate library against comparable libraries								
Discarded items reports								

Table 8.
Additional Data Sources Required for Phase Two Decision Situations

Decision Situation	Library–reference statistics	Library–program reports	Library–Ask-a-Librarian reports	Library–Questionpoint reports	Library–Dynix circulation history	Library–teacher deposit reports	Library–output measures	Library–inventory	Library–technology spreadsheets	Library–web page access data	Library–ILL reports	Library–meeting room reports	Library–requests to purchase	Library–iPAC searches	External–Census	External–Public Library Data Service
Area demographics															X	
Area demographics–marketing															X	
Define patrons, target services		X										X				
Staffing	X	X	X	X												
Technology planning								X	X	X						
Effectiveness of web marketing										X						
Collection development	X										X		X		X	
Youth collections						X	X								X	
Youth programs		X														
Web support of school assignments										X						
iPac search tracking to improve catalog														X		
Online databases																
Branch collections					X											
Evaluate library against comparable libraries																X
Discarded items reports																

Table 9.
Data Sources for Phase Two Decision Situations

	Horizon	Other library sources	External
Area demographics			1
Area demographics–marketing			1
Define patrons, target services	16	2	
Staffing	8	4	
Technology planning	4	3	
Effectiveness of web marketing	6	1	
Collection development	16	3	1
Youth collections	6	2	1
Youth programs		1	
Web support of school assignments	5	1	
iPac search tracking to improve catalog		1	
Online databases		1	
Branch collections	10	1	
Evaluate library against comparable libraries			1
Discarded items reports	8		

The numbers in the *Horizon* column indicate the number of data elements from the ILS required for each decision situation.

The numbers in the *Other Library Sources* and *External* columns indicate the number of data sources in each of these categories required for each decision situation.

Obtaining and Organizing the Data for Decision Support

Best practice indicates that a data warehouse be built in stages of smaller data marts to avoid costly failure and to accommodate limited development budgets. It is easier to create the data mart from a single operational database. Additional data elements from other sources may be incorporated in the following stages. Data from disparate sources require extra steps to examine their data definitions, determine common data fields, and map the related fields from each source to a common definition in the data warehouse.

When the Phase Two Interview Guideline was created, it was with the intention of forming the basis of a *circulation data mart* that would integrate usage of library and information materials regardless of format or access. Later examination of additional online database reports confirmed that reports vary considerably from one vendor to another, to a greater degree than originally detected. The reports must be unified as a separate project before they are ready for correlation with circulation. Once the library has several months experience in evaluating a unified online database report, it can address the problems of integrating this data with other data in the circulation data mart.

This will also be the case when data from other library sources and external data sources are included in the library data warehouse. The administrators named other sources for data besides the ILS in their interviews, but did not name the variables that they required.

These sources cannot be integrated into the first data mart, but should be considered for future data mart implementations.

The second phase interviews lead to the creation of two data marts, for *circulation* and *online database reports*

Circulation Data Mart

The circulation data mart is based on data elements drawn from Horizon. It is populated by individual item checkout events, including in-house use.

There are some differences between the data elements named in the phase two interviews, and those that appear in the data mart. The interview guideline included a question concerning renewals, but none of the administrators needed this data; therefore this element is not included in the data warehouse. Time and date of check out were combined into one question in the interviews, but they are not used identically in decision situations and were therefore separated for the data mart. Location of item and branch (or location) of checkout were combined into one question, but they are not used identically in decision situations and were separated for the data mart. Patron primary branch is a manually updated field in Horizon that is subject to errors. In its place the data mart uses the location of checkout.

Administrators requested additional data elements from Horizon besides those named in the interview guideline, and they are incorporated into the design of this data mart.

Of the fifteen decision situations, five will be addressed by a circulation data mart: define patrons/target services, staffing, collection development, youth collections, and web support of school assignments. The remaining decision situations cannot be answered from this data mart for the following reasons:

Three decision situations (area demographics, area demographics for marketing, and evaluation of the library against comparable libraries) rely on external sources.

Three decision situations (youth programs, iPAC search tracking to improve catalog, and online databases) rely entirely on other library sources. Online databases will be addressed in a separate data mart, discussed below.

For two decision situations that rely on Horizon data combined with data from other library sources, the data from the non-ILS sources is crucial (technology planning and effectiveness of web marketing). Another (branch collections) may be answered satisfactorily with ILS data only, but it requires data on all item records stored in Horizon. This data mart only contains accumulated records of items that have been checked out since the creation of the data mart.

Only one decision situation, discarded items reports, relies entirely on Horizon data; however, this decision situation also requires access to all item records and not just the records of items that have been checked out.

The remaining five decision situations that will be accommodated by the circulation data mart rely on both ILS and other library data, but may be answered satisfactorily with ILS data alone. The are described as follows:

1. Define patron, target services. This decision situation seeks to understand how groups of patrons use the library and its services. The administrator also requested data from reports on program attendance, meeting room uses, and online database vendor reports. Currently these services are not associated with patron records and establishing such an association is not anticipated. Use of item and patron records will answer this decision situation as well as possible.

2. Staffing. This decision situation also required input from library services such as reference desk logs and program reports, in order to indicate staffing needs that may not be reflected in circulation statistics. While these additional resources provide important depth to an examination of staffing needs, libraries often rely on hourly circulation statistics to judge staffing needs throughout a branch. Furthermore, at this particular library the statistics for a very staff-intensive service, signing up the public Internet computers, are collected as circulation statistics even though the signups are usually handled at the adult and youth reference desks. Examination of the subset of checkout statistics on the computer itype will reflect part of the reference and youth desk workloads.

3. Collection development. The data mart will allow the user to examine circulated materials using any of several constraints, such as date, collection code, call number, checkout location, patron age, or patron zip code. This administrator statement also called for information from other sources for a more complete picture of collection gaps, but substantial information on how the current library collection is now being used is available using this circulation data mart.

4. Youth collections. The administrator statement requested data from non-ILS sources that are on a much coarser grain that ILS circulation data, and while valuable in other situations would not integrate well for this decision situation. The basic data needs are item circulations by location by patron age group by collection code. The primary difficulty, noted by the administrator, is with the patron age. Most children's books are checked out on a parent's card, and information from the patron record will not reflect the patron who actually uses the items.

5. Web support of school assignments. The library tries to provide links to reliable information on the Internet to complement and enrich the print items available for school assignments. Examination of trends in juvenile non-fiction circulation can indicate the types of information sought, and the cycles of major school assignments

that are repeated each year, so that Web links on each subject can be refreshed and in place each year before the assignments begin.

Table 10 shows the data requirements for these decision situations. The Data Requirements column indicates the checkout information required, the patron information required, and the item information required. The data requirements are shown again in table 11 in a matrix. The two record types, patron and item, are properly known as *entities* in relational databases, and circulation is the *relationship* between them. The entities and the relationship are described in table 12, and their *attributes* (variables, or fields) are defined in table 13.

Online Database Data Mart

The vendor reports are currently delivered to the library by e-mail, fax, and through the World Wide Web. The variables that are available, and the definitions of these variables, differ from one vendor to another. The administrator who oversees the online databases manually compiles a quarterly report summarized by month for the library's 40–45 subscribed databases.

Creating a data mart for these reports will be complicated, but will result in delivery of quality data for decision making. Each vendor's full reporting capabilities must be assessed, either by visiting the web site or contacting the vendor's customer representatives. This information should indicate the lowest summary level available for reports, and full descriptions of the reporting fields.

This information must be compared among all vendors to select the common fields available across all or most reports. Report downloads or defaults should be set to provide these fields. Where there are multiple file options, those that provide comma delimited, Excel, or other formats that can be imported into a spreadsheet should be selected. E-mail delivery of the files is preferable, or at least an automated monthly reminder to visit the web site and download the file. The remaining reports that are faxed or e-mailed will have to be entered into an appropriate spreadsheet or database file by hand.

The data fields must be uniformly defined and then mapped to the data mart. Vendor reports change frequently, and the spreadsheets and data mart must be updated in order to maintain data quality and consistency.

The transaction detail available from a program like Horizon stands in contrast to the lack of detail available from most online database vendor reports. Because their structures and uses are so different, it cannot be expected that their data elements and definitions will all coincide closely. It would help administrators to know trends in use by time of day, day of week, and month of year, and whether access was remote or from within the library. Several monthly vendor reports from the first half of 2003 were examined briefly. There are some

Table 10.
Data Requirements for the Decision Situations, Circulation Data Mart

Decision situation	Data requirements
Define patrons, target services	atomic level data; each item checkout transaction with location, for data mining patron age, school, ZIP code, work phone item collection code, location, call number, publication date, genres, subjects, author, and title
Staffing	each item checkout transaction by location, by date summarized by day, by time summarized by hour patron age item collection code, call number, subjects, in-house use
Collection development	each item checkout transaction by location, by date summarized by month patron age, school, and zip code item collection code, location, call number, publication date, genres, in-house use, date of last checkout
Youth collections	each item checkout transaction by location, by date summarized by day, by time summarized by hour patron age, school item collection code, call number
Web support of school assignments	each item checkout transaction by location, by date summarized by week patron age, school item collection code, call number

Table 11.
Circulation Data Mart Entities and Attributes Used in Decision Situations

Decision Situation	Circulation–location	Circulation–date	Circulation–time	Patron–age	Patron–school	Patron–9-digit ZIP	Patron–work phone	Item–collection code	Item–location	Item–call number	Item–publication date	Item–genres	Item–subjects	Item–in-house use	Item–last checkout
Define patron, target services	X	X	X	X	X	X	X	X	X	X	X	X	X		
Staffing	X	X	X	X				X		X			X	X	
Collection development	X	X		X				X	X	X	X	X	X	X	X
Youth collections	X	X	X	X	X			X		X					
Web support of school assignments	X	X		X	X			X		X					

Table 12.
Entity Descriptions for the Circulation Data Mart

Entity/ Relationship	Definition
Patron	The patron entity describes the attributes of each individual who is registered with the library. It has a many to many relationship with the item entity (each patron may check out multiple items, and each item may be checked out by multiple patrons). It has a one to many relationship with the circulation relationship (each patron may participate in more than one circulation event).
Item	The item entity describes the attributes of each physical item copy cataloged by the library. It has a many to many relationship with the patron entity (each item may be checked out by more than one patron; each patron may check out more than one item). It has a one to many relationship with the circulation relationship (each item may be checked out in more than one circulation event).
Circulation	Circulation describes the event that relates patrons and items. It contains attributes for the time, date, and location of the checkout event. It has a many to one relationship with the patron entity (each patron may participate in more than one circulation events). It has a many to one relationship with the item entity (each item may be checked out in more than one circulation event).

Table 13.
Attribute Descriptions for the Circulation Data Mart

Attribute	Definition	Example
	Patron Entity	
Age	derived from birth_date field	52
Address	street address only	22 Sunny Side Lane
Zip code	5 digit zip code	54321
9-digit zip	use instead of address, not both	54321-2934
Record created	date patron record was created	1979-05-30
Date last used	date patron record was last used	2003-09-30
Work phone	area code plus prefix	970-228
B-status	school or age range field from b-stat field	Adult (21–64)
	Item Entity	
Barcode	unique 14 digit barcode	3 3897 00827 2856
Collection code	collection	adult CD audiobook
Item location	branch where housed	branchC
Call #	item call numbers, non-fiction	641.268
Itype	circulation policy	2-week renewable
Creation date	record creation date, ISO Standard	2001-06-12
Bib#	unique ID of associated bib record	3237729
Publication date	date of publication from bib record	1987
Title	title from bib record	Seeing a large cat
Author	from authority record	Peters, Elizabeth
ISBN	International Standard Book Number	0788798537
In-house use	sum of in-house use indications	10
Last checkout date	ISO Standard YYY-MM-DD	2000-02-04
	Circulation Relationship	
Date	ISO Standard YYYY-MM-DD	2003-10-01
Day of week	Monday, Tuesday, etc.	Wednesday
Day # of year	From 1–366	274
Week of year	ISO standard. Week from Mon to Sun; first week of year varies.	40
Month	name of month	October
Month # of year	number of month	10
Quarter	number of quarter	2003Q4
Year	year	2003
Location	name of branch where check out occurs	BranchA

vendors who provide exhaustive information, often configurable to the library's reporting needs. ProQuest and AccessScience are two of these. But the reports from the four companies that provide the eight most used databases are very sparse (table 14).

Gale, which provides five of the high-use databases, e-mails a monthly summary report. The company began reporting remote usage on its summary report after reports were examined for this study. It also offers more detailed reports that this library may either not be aware of, or has not had time to investigate and set up.

The Oxford English Dictionary and SIRS also e-mail their monthly reports.

bigchalk's monthly report, available by download through the World Wide Web, provides two report summaries for each month: remote use and library use.

Two examples will suffice to show how the vendor reports may be correlated. A brief examination of some report definitions indicate that Gale's "retrievals" may be comparable to SIRS "articles e-mailed/printed." OED provides statistics on "find word searches" and "full text searches" that could be totaled for an equivalent entry to "total searches" provided by the other four vendor reports in the table.

This data mart may be built by identifying key concepts, checking definitions, and mapping them to common terms, before proceeding to examine the more unique terms or offerings of each database.

Future development—other library sources

Eleven of the fifteen decision situations contain requests for data from library sources other than the ILS. The administrators named fourteen potential sources, at least one of which (the teacher deposit report) does not exist yet. Interlibrary Loan reports are currently only summary reports from OCLC, with no title, frequency, subject, or other data that would help in collection development.

Reference statistics, program information, and meeting room use are compiled in monthly summary reports from each branch. The monthly report is a paper form that is usually faxed to the administrative offices. At least one branch has created an Excel file and updates these statistics electronically each morning, then copies the monthly totals to the administrative assistant. The administrative assistant compiles paper and electronic reports into another Excel file. The report parameters are supposed to be uniform across the system, but statements made by one administrator during her first interview indicated that some branches devise reports to suit their own purposes.

Administrators did not identify the data they need from these sources. Before the reports can be used, administrators need to evaluate the data collected in the reports in relation to specific decision situations. This kind of scrutiny will indicate which reports, files,

Table 14.
Examples of Available Variables, Online Database Reports

Variable	SIRS	OED	bigchalk	Gale
Total accesses	X			
Total sessions	X		r/l*	X
Reference materials viewed	X			
Total searches	X		r/l	X
Full text searches	X	X		
Subject heading searches	X			
Advanced searches	X			
Topic browse searches	X			
Full text articles viewed	X			
Database features searches	X			
Sources/summaries/descriptors viewed	X			
Articles printed/e-mailed	X			
Graphics viewed	X			
Total requests		X		
Home page		X		
Welcome		X		
Entry displays		X		
Find word searches		X		
Word of the day		X		
Word list displays		X		
Static pages		X		
Retrievals			r/l	X
Total connect time			r/l	
Total views				X
Turned away				X

*r/l–bigchalk provides separate statistics for remote and library access for each of the variables

or other sources have the potential to provide useful information that could be included in a data warehouse. The logs, collection forms, or report forms may be revised to conform with decision making needs. A unified electronic reporting environment would enhance data collection for many library services.

Future development—external sources

External data from the U.S. Census or from the Public Library Data Service are already in electronic format. Four decision situations require or would be enhanced by use of census data. It is not certain that PLDS data, requested by only one administrator, may be purchased in its entirety in electronic format.

Census data, and probably the PLDS data, can be so extensive that more planning needs to be done to make certain these sources are addressed efficiently. Integrating the data with Horizon data may not be difficult, but the library needs to determine if the number of users and uses justifies the expense of planning for and integrating the data.

Future development—ILS

Four of the five decision situations coded S4 and S5 require data only from the ILS: turnover rate of individual collections, branch collections, purchase alerts, and discarded items reports. These reports are created by some ILS' programs, but not Horizon. They are examples of how management information systems can overcome the limitations of the operational system without the need for a data warehouse, unlocking data that is present but not usable. Software applications such as *EasyAsk* query generator are designed for this purpose. The library has purchased several licenses of a version of this product designed to integrate with Horizon. Administrators should investigate using *EasyAsk* for these needs.

The discarded items report can probably be addressed by exporting selected Horizon data to a spreadsheet. When items are discarded from the database, the ILS does not save any trace of them for further use. The former ILS, Dynix, saved these records to a file so that librarians could examine them for collection development. It also maintained information on whether the losses were through circulation or discarding, and provided a year-end count of collection losses as required for state reports.

During the interviews, administrators expressed other concerns with the data management used by the ILS. Some statistics are aggregated prematurely, eliminating their use for trending studies. Item checkout statistics are kept only at the item/copy level rather than at the bibliographic level, so that those numbers are lost when a copy is discarded. These and other limitations of the ILS need a separate study; solutions may range from exporting data to spreadsheets, as above, or expanding the data mart to answer some of these concerns. This study would also give the library an organized set of requirements to use in evaluating integrated library systems should it decide to investigate other systems in the future.

CHAPTER 5

DISCUSSION AND CONCLUSION

Comments

This study examined a means for determining data needs for decision support in a medium-sized public library. Decision support tools will be used with a data warehouse, which can serve two purposes: retaining historic information, usually discarded by the operational databases, for trend studies, and integrating data from disparate sources in order to provide a more detailed view of the library and its community. The data for a data warehouse is selected in order to fulfill specific information or decision making needs. The study attempted to learn those needs, and the data required to answer them, in two round of interviews with the potential users of the data warehouse. Those users were nine administrators from the top two levels of the library organization.

The first round of interviews began discovery at the decision situation (or information need) level, asking administrators to recall situations in which they needed information they knew was available but which was not easily accessible to them. They were then asked what specific data they needed for those situations.

The second round of interviews began at the data level, using data elements from the library's integrated library system and from online database vendor reports. Each interviewee was asked how specific data elements would answer her decision situations.

Twenty-one decision situations were identified over the course of the two rounds of interviews. The second round of interviews was designed so that each decision situation under discussion would be fully identified with its necessary data; however, only eight were this well defined. When administrators discussed the remaining six decision situations, they identified data sources, but not specific data entities and attributes that would be required.

The library should build a series of data marts, each designed for a specific purpose but with common data definitions to provide integration. Five of the eight decision situations can be answered by a circulation data mart that is populated by each item checkout. The ILS entities and attributes needed for those decision situations were described and defined for the proposed mart. Further considerations were outlined for development of an online database reports data mart, and for further studies leading to creation of additional data marts or expansion of any current marts.

Decision situations coded in the S1 and S2 categories demonstrate the duality of modern library practices. The S1 decision situations are clearly aimed at taking the library

into the community, either physically (by adding new branches) or by discovering populations with unmet needs. These reflect the philosophy that the library serves all of members of its service area.

The majority of the S2 decision situations are concerned with discovering who the library's patrons are, and what they want from their library, in order to target services to them. Libraries, like most tax-supported institutions, operate under financial constraints. This library does not anticipate any substantial increase in funding to support its ever mounting needs, and feels that targeting services is an efficient means of allocating resources. This course can bring a library perilously close to defining its users in advance, so that the library services and library users become a closed circle.

Libraries that use business analysis tools should consider that they may be required to run themselves as businesses. Companies use data warehouses in order to increase their competitive edge and boost profits. Public libraries are not profit driven, though they are expected to demonstrate their efficient and effective use of public monies. A library that uses profit-based customer targeting to concentrate resources on high-use patron types may signal to its governing authority that efficiency is paramount, resulting in expectation of further cost-saving measures based on a business model.

During most of the interviews, however, administrators discussing S2 decision situations repeatedly brought up the reverse use of defining the patron—to find out who does not use the library, why, and what the library can provide for them. Targeting services used by high-use patrons (eliminating or changing services that are not used, though they were designed for these patrons) may also free some resources for provision of services to other populations in the service area.

Efficiency and good customer service are valid concerns of a public library. Data warehouses and decision support allow the library to learn about its patrons. They permit an examination of the community and have the potential to increase services to underserved groups.

Implications for Future Study

The needs analysis for a data warehouse creates a focus for what could otherwise be an unwieldy data gathering process. The focus narrows questions on who will use the data warehouse, how these users expect to increase their information for decision making using the warehouse, and what data will be needed to provide the information. Business enterprises, with their greater data sources and numerous business needs, create a focus even before the needs analysis. Although libraries are less complicated institutions with smaller data sources, they would be advised to emulate the enterprise approach in this regard.

At the outset of this study it was understood that there are known difficulties in determining information needs for a data warehouse, or other decision or management information system. Potential users may not recall instances in which they had difficulty assembling necessary data or information; they may know exactly what information they need but may not express their needs accurately; or they may be unrealistic in their expectations of the final product. These difficulties would be reduced if the needs assessment should come as the second step in preparation for decision support, rather than the first step.

A library, like a business, must understand itself before launching on a data warehousing and decision support project. It should conduct a study to understand its data sources, what information it currently uses, and its own management and decision making processes. An outside consultant should be employed to conduct the study. The consultant can help the library understand its planning processes, and how information relates to that planning. This preliminary assessment will help the library know its strengths and weaknesses in its data collections and in its information gathering methods. It will enable the library to clarify which of its objectives require adjustment, clarification, discovery, or evaluation, and prioritize these objectives so that the needs assessment can proceed in a well-defined area of inquiry.

The needs assessment itself may be undertaken by interviews, as in this study, or by focus group sessions. It should be conducted by two persons: a planning or management consultant and technology consultant. Focus group sessions, followed by individual interviews, may provide the most efficient means of needs assessment. The initial focus group session (or sessions) is used to elicit information needs related to the area defined in the preliminary assessment. This provides an opportunity for potential decision support users to learn about data limitations, which were uncovered in the preliminary assessment. The session should conclude with prioritization of the information needs to be initially addressed by the data warehouse so that all are in agreement as to their importance to the planning and decision making needs of the library. Follow up interviews refine the needs of each data warehouse user, to discover individual requirements for information and reporting.

Creation of data warehouses and data support are time consuming and expensive projects. Libraries–and businesses–may be tempted to reduce costs by using current staff for planning, creating and maintaining these systems. However, all three aspects require the full-time attention of a trained specialist. The personnel costs are as necessary as the software and hardware expenses. Data warehouse development, properly funded and guided, has much a much greater potential to result in a decision support system that provides for user needs and can expand to accommodate growth and change in library decision making.

REFERENCES

REFERENCES

Adams, Roy. 1995a. Strategic information systems and libraries. *Library Management* 16, no. 1, doi:10.1108/01435129510076187, Emerald Fulltext, http://www.emeraldinsight.com/0143-5124.htm.

——. 1995b. Strategic information systems and libraries. *Library Management* 16, no. 1: 11. *Citing* King Research Ltd. Keys to Success: Performance Indicators for Public Libraries. London: HMSO. 1990.

Atkins, Stephen E. 1996. Mining automated systems for collection management. *Library Administration and Management* 10 (Winter): 16–19.

Berry, Michael J. A., and Gordon S. Linoff. 2000. *Mastering data mining: The art and science of customer relationship management*. New York: Wiley Computer Publishing.

Bleyberg, Maria Zamfir, Dongsheng Zhu, Karen Cole, Doug Bates, and Wenyan Zhan. 1999. Developing an integrated library decision support data warehouse. In *IEEE International Conference on Systems, Man, and Cybernetics*. Vol. 2. 1999, at Tokyo. IEEE Xplore, http://www.ieee.com/.

Bonifati, Angela, Fabiano Cattaneo, Stefano Ceri, Alfonso Fuggetta, and Stefano Paraboschi. 2001. Designing data marts for data warehouses. *ACM Transactions on Software Engineering and Methodology* 10, no. 4 (October): 452–83.

Cheng, Rachel, Steve Bischof, and Alan J. Nathanson. 2002. Data collection for user-oriented library services: Wesleyan University Library's experience. *OCLC Systems & Services* 18, no. 4, doi:10.1108/10650750210450130, Emerald Fulltext, http://www.emeraldinsight.com/1065-075X.htm.

Clark, Philip M. 1993. A viable MIS: DeKalb County Public Library. *The Bottom Line* 7, no. 1: 30–31.

Cole, Karen, Michael Somers, and Jill Emery. 2001. Data warehousing: Developing a support system prototype. *Serials Librarian* 40, no. 3–4: 349–53.

Dadashzadeh, Mohammad, Kathryn Payne, and John Williams. 1996. The development and implementation of the periodicals analysis database. *Serials Review* 22, no. 4 (Winter), doi:10.1016/S0098-7913(96)90071-4, http://www.sciencedirect.com/science/article/B6W63-45R6SYF-F/2/5d5245fb5247b4a7e33a27b0e1781f6d.

Dennison, Russell F. 1999. Usage-based staffing of the reference desk. *Reference & User Services Quarterly* 39, no. 2 (Winter): 158–65, Gale Group Databases InfoTrac OneFile, acc. A61755297, http://infotrac.galegroup.com/.

Dowlin, Ken, and Lynn Magrath. 1982. Beyond the numbers—a decision support system. In *Library automation as a source of management information*, ed. F. W. Lancaster. Urbana-Champaign: University of Illinois at Urbana Champaign.

Evans, G. Edward, and Margaret R. Zarnosky. 2000. *Developing library and information center collections*. 4th ed, Library and Information Science Text Series. Englewood, CO: Libraries Unlimited.

Gleeson, Michael E., and John R. Ottensmann. 1993. Using data from computerized circulation and cataloging systems for management decision making in public libraries. *Journal of the American Society for Information Science* 44, no. 2 (March), doi:10.1002/(SICI)1097-4571(199303)44:2<94::AID-ASI4>3.0.CO;2-#, http://www3.interscience.wiley.com/cgi-bin/issuetoc?ID=10049735.

Guenther, Kim. 2000. Applying data mining principles to library data collection. *Computers in Libraries* 20, no. 4 (April): 60–63, Library Literature and Information Science Full Text, acc. 200000396000, http://www.silverplatter.com/.

Hammons, James. 1999. Mining your OCLC ILL data: Using OCLC ILL management statistics with Microsoft Access and Excel (but mostly Access). *Journal of Interlibrary Loan, Document Delivery & Information Supply* 9, no. 3: 3–15.

Joshi, Kailash, and Mary Curtis. 1999. Issues in building a successful data warehouse. *Information Strategy* 15, no. 2 (Winter): 28–35.

Kimball, Ralph. 1996. *The data warehouse toolkit: Practical techniques for building dimensional data warehouses*. New York: John Wiley & Sons.

Kimball, Ralph, Laura Reeves, Margy Ross, and Warren Thornthwaite. 1998. *The data warehouse lifecycle toolkit: Expert methods for designing, developing, and deploying data warehouses*. New York: John Wiley & Sons.

Kudyba, Stephan, and Richard Hoptroff. 2001. *Data mining and business intelligence: A guide to productivity*. Hershey: Idea Group Publishing. http://www.netlibrary.com/.

Mancini, Donna D. 1996. Mining your automated system for systemwide decision making. *Library Administration and Management* 10, no. 1 (Winter): 11–15.

McClure, Charles R., Liz Hagerty-Roach, Lindsay Ruth, and Pat England. 1989. Design of a public library management information system: A status report. *Library Administration and Management* 3 (Fall): 192–98.

Metz, Paul, and John Cornelius Cosgriff. 2000. Building a comprehensive serials decision database at Virginia Tech. *College and Research Libraries* 61, no. 4 (July): 324–34, Library Literature and Information Science Full Text, acc. 200000889800, http://www.silverplatter.com/.

Mukherjee, Debasish, and Derrick D'Souza. 2003. Think phased implementation for successful data warehousing. *Information Systems Management* 20, no. 2 (Spring): 82–90.

Nelson, Sandra S. 2001. *The new planning for results: A streamlined approach*. Chicago: American Library Association.

Nicholson, Scott. 2003. The bibliomining process: Data warehousing and data mining for library decision-making. *Information Technology and Libraries* (tentative pre-print), http://www.bibliomining.com/nicholson/biblioprocess.htm (accessed November 11, 2003).

Nicholson, Scott, and Jeffrey Stanton. 2003. Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries [pre-print version]. In *Organizational data mining: Leveraging enterprise data resources for optimal performance*, ed. H. R. Nemati and C. D. Barko. Hershey, PA: Idea Group Publishing, http://www.bibliomining.com/nicholson/odmcom.html (accessed November 11, 2003).

Ottensmann, John R., and Michael E. Gleeson. 1993. Implementation and testing of a decision support system for public library materials acquisition budgeting. *Journal of the American Society for Information Science* 44, no. 2 (March), doi:10.1002/(SICI)1097-4571(199303)44:2<83::AID-ASI3>3.0.CO;2-0, http://www3.interscience.wiley.com/cgi-bin/issuetoc?ID=10049735.

Ottensmann, John R., Raymond E. Gnat, and Michael E. Gleeson. 1995. Similarities in circulation patterns among public library branches serving diverse populations. *Library Quarterly* 65, no. 1: 89–118.

Payton, Fay Cobb, and Debra Zahay. 2003. Understanding why marketing does not use the corporate data warehouse for CRM applications. *Journal of Database Marketing* 10, no. 4 (July): 315–26, Ebsco Business Source Premier, acc. 10287865, http://www.epnet.com.

Schulman, Sandy. 1998. Data mining: Life after report generators. *Information Today* 15, no. 3 (March): 52, ProQuest ABI/Inform, http://www.proquest.com/.

Shelly, Gary B., Thomas J. Cashman, and Misty E. Vermaat. 2001. *Discovering computers 2002: Concepts for a digital World*. Boston: Course Technology, Thomson Learning.

Stake, Robert E. 1995. *The art of case study research*. Thousand Oaks, Ca.: Sage Publications.

Su, Siew-Phek T., and Ashwin Needamangala. 2000. Harvesting information from a library data warehouse. *Information Technology and Libraries* 19, no. 1 (March): 17–28, Gale Group Databases InfoTrac OneFile, acc. A62741686, http://infotrac.galegroup.com/.

Swisher, Robert, and Charles R. McClure. 1984. *Research for decision making: Methods for librarians*. Chicago: American Library Association.

U.S. Bureau of the Census. 2001. Population Estimates Program. http://www.census.gov/main/www/cen2000.html (accessed January 10, 2003).

Wixom, Barbara H, and Hugh J. Watson. 2001. An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly* 25, no. 1 (March): 17–41.

APPENDICES

APPENDIX A

PHASE ONE INTERVIEW GUIDELINE

Questions for interviews were in a semi-structured format. Certain questions were asked of all participants, and their responses determined the direction of further questions.

1. What is your job title?
2. Who is your immediate supervisor?
3. What are your duties? First, give an idea of your official duties, and then some of the ad hoc assignments you have handled, or areas you participate in out of necessity or personal interest.
4. Try to recollect one or two instances in which you have needed data (or information) and have not been able to gather them in a timely manner. This might have been for a report, or in response to an unusual or infrequent request. You might have known (in general) where to find the data either inside or outside the library system, but not been able to get them easily, or might have had to manually compile data from two or more sources. Or you might have had an idea of what kind of data you needed, but were unable to them locate at all.
5. In reference to these incidents, where did you find the data that you needed? Did it take a long time to find and compile?
6. Can you think of other information you would like to have that would make you more efficient at your work, or that you think would be beneficial to the library, that would involve collecting information from a number of different sources?
7. If you had this information readily at hand, what would you do with it? How would it benefit you or the library?

APPENDIX B

PHASE ONE INTERVIEW STATEMENTS

The twenty eight interview statements are as follows:

1. Use Horizon patron information and external sources such as tax rolls to show how many people who live in a zip code use the library. Economic level information [from the external source] could indicate whether the library is serving the people who really benefit from a free public library.
2. Demographic and geographic information, both in the area and specifically on our patrons, would indicate any growing groups that do not use library services. Then the library would decide what to do about it. Desired information would include geographic location, ethnic background, language background, and income level. For many reasons, this information [except geographic] may never be available for our patrons.
3. Keeping track of area demographics will help determine activities to library patrons outside its walls and the direction of outreach services to the community at large. Example: the library may need to add Spanish to the web site.
4. Information is needed on the community that would help increase audience size for library programs; past data on the programs would aid by determining previous community response.
5. Use Horizon patron information and external sources such as tax rolls to show how many people who live in a zip code use the library. This could be used for publicity.
6. Where do people live who use the library. Awareness of where growth is occurring will help in planning branches.
7. Where are the developing areas; this information is probably available from the Development Commission. Awareness of where growth is occurring will help in planning branches.
8. Use Horizon patron information and external sources such as tax rolls to show how many people who live in a zip code use the library. Looking at library use through the community could contribute to branch planning.
9. In demographics age is important to help target services and programs better.
10. When people come in to use the technology or attend programs, are they using the rest of the library? What brings them to the library and what else would they do it if were available to them?

11. Data mining uncovers patterns you haven't even thought about. The ultimate target is usership of the resources we have.

12. Demographic and geographic information, both in the area and specifically on our patrons who attend programs, would give a detailed snapshot of the current patron base that the library reaches now and wishes to maintain. Desired information would include geographic location, ethnic background, language background, and income level. For many reasons, this information [except geographic] may never be available for our patrons.

13. Uncover the demographic of the patron to tailor services—I want to know how old you are, where you go to school, do you work out of the area.

14. Discover patterns in collections and programming to identify market and tailor to it, make sure patrons find what they need when they come here.

15. Staffing—when are the peak times, and what services are patrons using at those times.

16. What are the use patterns in terms of time frames—years, months, the library is busy one time but not another—to determine staffing.

17. There's the planning issue for 6 months to 2 years out. Data is needed to help in determining what the library will put on public access machines. A source that would help would be a report on the kind of technical information from the collection our people are going after. For example, if they are checkout out Office 2000 books then this might indicate we need to update our computers from Office 97.

18. Circulation statistics may show whether resources the library has been pushing on the web are being used or wanted, to help decide whether time and money should be spent on similar resources.

19. Discover the demographics of the patron to define non-users and consider reaching out to them.

20. Examine marketing as it relates to physical facilities and the optimum way to lay out the collections and services to get maximum usership and to be as inviting as possible to the public.

21. Find a way to implement a tracking feature in iPAC. For a certain number of days this records everything patrons search. It shows which indexes were used, and which terms turned up null sets. The library would use this to suggest cross-references in the catalog to improve search results.

22. Examine accessibility—when people come here, how do they search [the catalog]? Do they find something, or go away? Do we need to catalog better? Do we need a better path to lead them to what they want?

23. The usage statistics from online vendors are difficult to correlate to each other, or to other statistics.

24. This library relies on circulation figures for use statistics, but much item use is in-house. Recent budget cuts forced reductions in serial subscriptions, and two little-used magazines were eliminated. The coordinator has since received several phone calls from patrons who miss their favorite magazine. Statistics that are not collected cannot be included in planning.

25. Turnover rate—items in a collection vs. circulation—is a flawed but still useful tool for evaluating collections and get a sense of materials use.

26. Purchase alerts let the library know when hold requests exceed four per copy in a bibliographic record. Right now the coordinator pulls a list of all items on hold and sorts by title, yielding a list of about 12,000 holds. She scans the list for titles with high numbers of requests, and then goes to the catalog to see how many copies the library owns.

27. We need to locate numbers that have been generated, but are not convenient. Examples are number of patrons served, book stock, books added annually; even the budget, which is public information, can only be located by asking a number of people.

28. Administrator makes frequent reports for which he needs to gather information such as current configuration of the technical infrastructure, such as number of workstations, types of software, the number of telephones, how many public access computers connect to the Internet, the speed of the Internet connection. Also budget information related to how money was spent in the past for technical improvements. For every telecommunications provider, reports must delineate how much money was paid for phone service, directory service, state and local taxes. Right now he pulls paper bills together for this.

APPENDIX C

PHASE TWO INTERVIEW GUIDELINE

Preliminary Remarks

Thank you for taking the time to meet with me today. I am ready to recommend exactly what data elements will go into a data warehouse. I need to find out how you would use various data elements, in what combinations. So today we will look at some data elements in Horizon and determine if they are useful, and how they will help you make decisions. If I don't ask about something that you need, you will be given an opportunity to tell me what further sources you need.

Let's recall the [number] major items we discussed in our previous interview. At that time, you said that you need better data or better access to data for the following reasons:

[list]

Please select one or two of these that are most important to you for us to consider today.

[When one statement is chosen, repeat and/or rephrase to make certain of agreement as to meaning]. Now let's look at some Horizon data elements to see if they can help you.

I. Patron Data

There are certain data in our patron records that can be used for data mining. I would like your opinion about the usefulness of each data element:

1. About the age, how could this information help you in making decisions? [*Probe*– It would help me if you can give me some examples?]

2. About school affiliation, how could this information help you make decisions? [*Probe*–It would help me if you can give me some examples?]

3. About residential 9-digit zip codes, how could this information help you make decisions? [*Probe*–It would help me if you can give me some examples?]

4. About work phone number, how could this information help you make decisions? [*Probe*–It would help me if you can give me some examples?]

5. About the primary branch, how could this information help you make decisions? [*Probe*–It would help me if you can give me some examples?]

Beyond age, school affiliation, residential zip code, work phone number, and primary branch, are there are other data in our Horizon patron records that would help you make decisions?

II. Library materials usage

A. Circulation

Certain statistics are generated at checkout and checkin that can be used for data mining. Some of these are strictly circulation statistics, and some are associated with item records. I would like your opinion about the usefulness of these data elements or statistics:

1. About the time and date of checkout, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

2. About the date of checkin and days overdue, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

3. About the last checkout date of an item, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

4. About the total number of checkouts of an item, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

5. About the renewal information (number of renewals, number of phone renewals, and number of OPAC renewals) for an item, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

Beyond these statistics or elements drawn from circulation and item records, are there any other circulation statistics that would help you make decisions?

B. Item records

There are certain data in our item records that can be used for data mining. I would like your opinion about the usefulness of these data elements:

1. About creation date, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

2. About the collection (including age level, format, fiction or non-fiction, and fiction genre) how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

3. About the location, and the branch of checkout, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

4. About the call number, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

Beyond these elements, are there other fields in the item records that would help you make decisions? [Present a printout of an item record to aid in remembering potential data fields.]

We did not look at elements in the bibliographic records. Are there data fields in those records that would help you make decisions? [Present a printout of a MARC record to aid in remembering potential data fields.]

C. In-house use

Some materials are used in-house and this is not reflected in circulation statistics. Would this information be useful for your decision making?

If yes, how could this information help you. [Probe–It would help me if you can give me some examples?]

III. Electronic Resource Usage

There are certain data from our vendors that can be used for data mining. I would like your opinion about the usefulness of each data element:

1. About the number of sessions, or logins, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

2. About the duration of sessions, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

3. About the access from the library or remote, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

4. About denied access due to concurrency caps, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?

5. About the number of items printed, how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

6. About IP addresses used for access (used by one vendor), how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

7. About usage from each branch (possibly available from two or three vendors, though not implemented), how could this information help you in making decisions? [*Probe*–It would help me if you can give me some examples?]

Beyond this information, is there any other information regarding electronic databases usage that would help you make decisions?

Is there any other information in Horizon, in other library sources, or from outside sources, relating to [mention topic chosen for discussion] that would help you make decisions?

APPENDIX D

PHASE TWO DECISION SITUATIONS

1. Area demographics. Use census information to indicate changes in area population. If possible, examine same trends that are being watched by the public school systems.

2. Area demographics–marketing. Use general demographic characteristics of the community to identify a community segment for marketing and plan programs that meet the needs of the population.

3. Define patrons, target services. Combine patron and item circulation data to define patron groups and target services to them. Analyze uses of other library resources.

4. Staffing. Use patterns for each library service (circulation, reference desks, online reference services, program information.), at each branch, at each service desk, during the course of the day, days of the week, and weeks of the year, to determine trends in use and predict staffing needs. Optional but useful information: collection code and call numbers of circulated items for seasonal usage.

5. Technology planning. Examine electronic resources (sources: inventory and personally maintained spreadsheets) against use (source: print count reports, computer checkouts).

6. Evaluate web marketing. Correlate web page views with circulation information on the materials contained in web book lists to see if it is possible to evaluate web marketing by circulation trends.

7. Collection development. Examine patterns of use on a variety of criteria such as checkout branch, time in transit, age of patron, etc. to refine knowledge of what patrons in certain groups or areas are reading and whether those materials are readily available to them. Second need: collect and analyze collection gaps indicated by requests to purchase, ILLs, reference desk logs.

8. Youth collections. Analyze patron demographics against collection statistics to evaluate use of children's and YA collections, and see what ages are using each collection. Patron age, school; item collection code, call number, checkout date, time, and location.

9. Youth programs. Create exit survey for each participant in juvenile and YA programs: age/grade, how attendee learned about program, opinion of program.

10. Web support of school assignments. Analyze circulation particularly in juvenile non-fiction to predict timing of the major assignment cycles in each school system, to prepare web support.

11. Evaluate iPAC use for cataloging improvements. Enable search tracking to analyze use of iPAC to modify cross-references or otherwise suggest ways to lead patrons to materials they seek.

12. Online databases. Streamline the collection and collation of vendor reports. Map different field descriptions. Create automated quarterly reports.

13. Branch collection use. Examine collections at the branches by any available item description (collection code, genre, call numbers). Compare over time (such as all JE 616s on January 1 of each succeeding year).

14. Evaluate libraries against comparable libraries. Use statistics from the Public Library Data Service for comparison with other libraries.

15. Discarded item reports. Maintain records of discarded separately from Horizon. Retain author, title, ISBN, publication date, item creation date, collection code, branch location, item status (immediately prior to weeded status), date of last status update, item number, bibliographic record number.

VITA

Thena S. Jones is a reference assistant at the Central Rappahannock Regional Library, Fredericksburg, Virginia. She received the Bachelor of Arts degree in anthropology from the University of Texas at Austin in 1972. She earned the Master of Science degree in Information Sciences from the University of Tennessee, Knoxville, in May 2004. She plans to remain in public librarianship after the completion of her master's degree.