PREPRINT - Cite as:

Nicholson, S. (2006, January 15). Proof is in the Pattern. *Library Journal netConnect*, Supplement to *Library Journal*, Winter 2006. 2-6. Available online at http://www.libraryjournal.com/article/CA6298562.html

The Proof is in the Pattern: Data-Based Decision Making in Libraries

Libraries follow the corporate sector by taking advantage of data troves

ALA's *Campaign to Save America's Libraries*(2005) reports the price and demand for library materials and services continues to rise while funding is flat or falling. More and more, librarians are forced to make difficult decisions about what stays and what goes. McClure (1989) explains many librarians use ad-hoc methods to make these decisions, relying on no data or simple aggregates in determining a course of action. Others have turned to more data-driven approaches that move beyond generalized aggregations (such as running totals and overall means) to reveal underlying patterns in the data that clarify which services or materials are worth retaining.

Gathering the Data

The first challenge facing librarians in using data for decision-making is gathering the data. Most libraries are supported by myriad computer systems. The integrated library system (ILS) reduces this problem, but with interlibrary loan (ILL), Web sites or content management systems, digital reference, electronic resources, community-based information such as blogs and wikis, and whatever comes next, librarians are left with a disarray of data sources and system logs.

Others have faced this challenge. In the 1980s, the corporate sector had similar data structures; each computer system produced a different type of data and worked independently. The phenomenon is called stovepiping and creates an overwhelming challenge in data-based measurement and evaluation across an organization.

Most systems supporting library services focus on the operational needs of the library. Once a transaction is complete, many systems either delete or hide the resulting data, as it's seldom needed for future operations. Personal information is a prime example. Once a patron completes a transaction, the system doesn't have an operational need to retain the data, so deleting it for reasons of patron privacy may seem the obvious thing to do.

But librarians are discovering uses for data archived from operational systems beyond the initial transaction. Unfortunately, library systems can make it difficult to extract, match, and clean the data needed for a specific project. When a librarian must make a decision quickly, such hurdles often mean falling back on McClure's "adhocracy" method.

Data Warehousing and Patron Privacy

How did the corporate sector address stovepiping? They created the data warehouse, a place to collect and normalize data from different systems. Extracted data is brought into the data warehouse and matched using common variables. The unified structure of the data warehouse makes future analysis easier. The key concept in the data warehouse is that it stores data useful in

supporting decision-making, which is different from the data required for the daily operations of the library.

For example, a digital reference service in an academic library has one data source about the users and another for the questions and answers. The university maintains the student data with current status information (major, grade level, etc.). The library wants to understand who's asking what questions and, more importantly, what resources are useful in providing answers. If the library only matches the university's current status information to past questions and answers, they won't know what the patron was like while using the service.

A data warehouse allows the library to capture the transaction along with data about the patron's characteristics *at the time*. People change addresses, grades, majors, and positions within an organization. The data warehousing process takes a snapshot of the patron status, attaches it to information about resources used, and brings both into a separate data source. The data warehouse is always ready for analysis and provides a more accurate picture of library use.

What about patron privacy? Current legislation makes libraries increasingly concerned about the privacy of individuals using their services, and some libraries delete large amounts of data about library use (Murphy, 2003). At the same time, libraries are called upon to justify themselves through data and outcome-based evaluation. Libraries must balance the desire to protect patron information with the need to be accountable to stakeholders.

It's possible to accomplish both goals in a data warehouse through the use of *demographic surrogates*. Just as bibliographic surrogates represent materials within library systems, demographic surrogates represent users in a library data warehouse. The personal information about a user is scrubbed and replaced with demographic data that is attached to information about the item used. In building surrogates, library administrators can select demographic fields that best serve management needs. For more information about creating demographic surrogates, see the special issue on Bibliomining of *Information Technology and Libraries* (Nicholson, 2003).

The data warehouse has other advantages. It can integrate tools and common reports used for analysis so more staff in the library can explore the data. Since the data warehouse is separate from the operational systems, running routines that heavily tap into it will not interfere with library operations.

Data Warehousing Projects in Libraries

A leading data warehousing application is the University of Pennsylvania's Data Farm. Spearheaded by librarian Joseph Zucca, the Data Farm has grown over the past six years. Zucca started by bringing data feeds from different systems, both human and automated, into the Data Farm. He matched these feeds and cleaned the data, removing personal identifiers. The Data Farm captures these data in non-aggregated form to allow flexible exploration.

Zucca describes what's in the Data Farm by focusing on the concept of a service event, which he defines as "some user interaction with the library." Service events include these data points:

- Resource (book, database, Web site, librarian),
- Some activity (usually a server activity, like a checkout, but it could be a term paper clinic or another human activity),

- Attributes of the resource (call number, language code, pub. date, or the reference librarian's specialty or liaison role),
- Attributes of the resource user (school, status, year, GPA, gender),
- Environment (location, date/time, in/out of the library, in-person or by chat), and
- Event outcomes.

In populating the Data Farm, Zucca matches each of these data points to others as appropriate and strips personal identifiers to produce a record of the service event, which goes into the Data Farm.

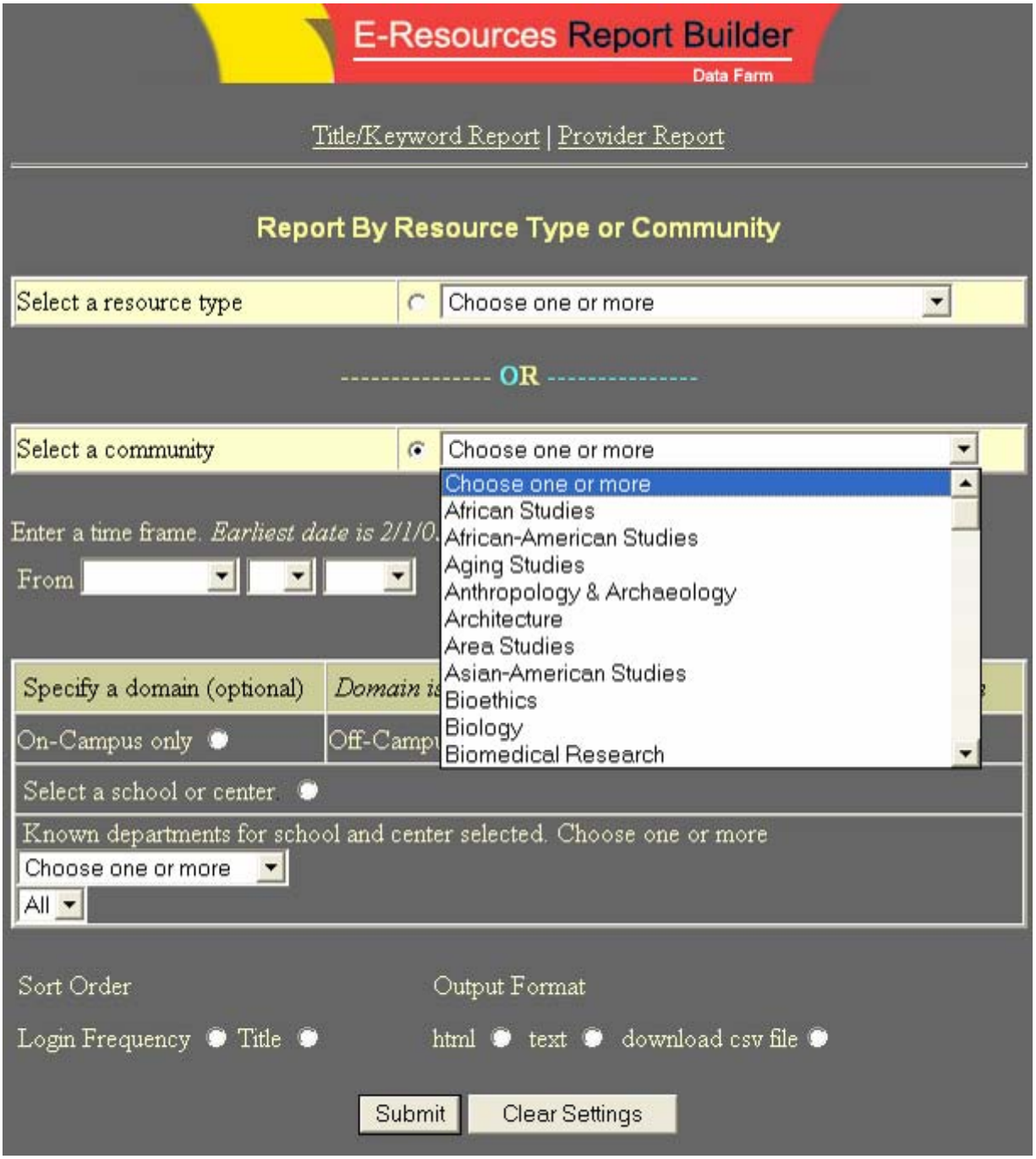

Figure 1: Choices from the Penn Library Data Farm

Zucca's goal is to integrate assessment into daily library operations. He finds that, as the Data Farm grows and more library staff use it, the library employs more data-driven assessments in making decisions. Zucca believes that "assessment is not a series of one-off projects or the preserve of some special office in the library. It's a priority of the staff, so the staff needs to be equipped."

For example, the University of Pennsylvania Library used the Data Farm to make decisions about moving works to off-site storage. After beginning the project, the librarians realized they needed to do mindful weeding, instead of just transferring books from one place to another. Without a data warehouse, the librarians can either wait for systems staff to identify low-circulating works or just shrug their shoulders, take a guess, and pile items onto the cart. The Data Farm changed that. Zucca explains, "The religious studies selector can monitor the rate of intake and use within her stack ranges over different periods of times and devise a profile to help inform her transfer decisions. All the data is up to the minute, quick to generate, structured in NATC schemes to make reports usable, and within easy reach of the managers."

The greater challenge faced by Zucca—and every library—is determining the impact of library services. We may know someone used a database or Web site, but what difference did it make? The rapidly growing field of outcome-based evaluation is focused on answering this question. (See related article on outcomes in this issue, page OUTCOMES.) There are few system-based measures that reveal outcomes, so these must come from users through other types of studies. Zucca states he's working to "create tools that help staff collaborate with faculty on outcomes assessment. We have projects like the latter underway. In the end, library outcomes raise policy issues for the institution as a whole, which issues are themselves independent of management tools like the Data Farm."

Most libraries don't have the resources to create a Data Farm. This opens the door for vendors of integrated library systems to get involved, and some are embracing the idea (See article about ILS data mining in this issue, page ILS). Like the Data Farm, SirsiDynix's Normative Data Project (NDP) gathers and normalizes data from different sources. NDP originated with data from libraries running the Unicorn ILS. The project matched Unicorn data with Census data, demographics from GIS data, and data about U.S. libraries from the National Center for Education Statistics (NCES). Participants pay a nominal fee to use NDP, which allows SirsiDynix to develop new tools for examining the data.

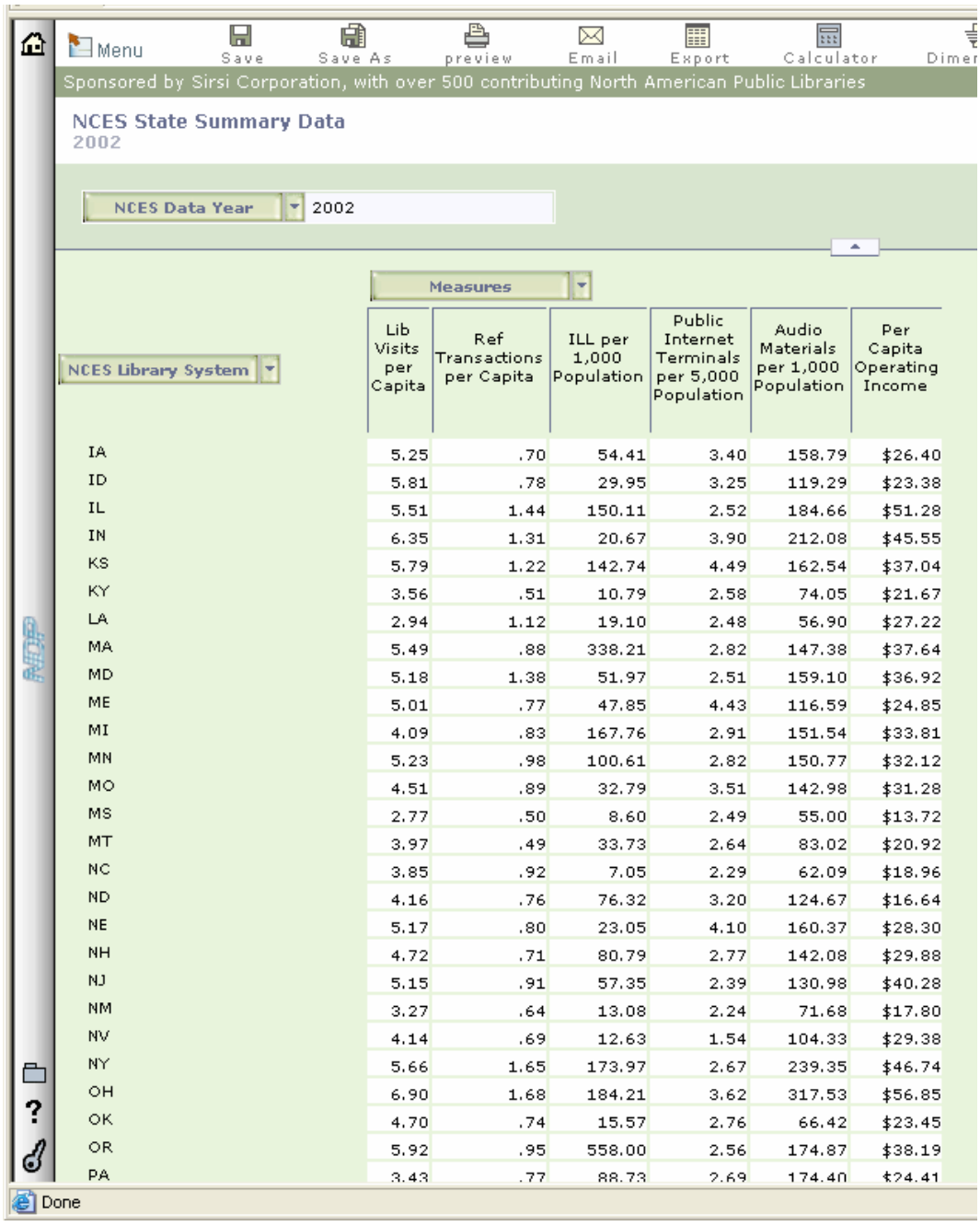

Figure 2: Data from the Normative Data Project for Libraries

Bob Molyneux, Chief Statistician for SirsiDynix, gets giddy when talking about the NDP. Molyneux exclaims, “The first time I saw it work I said to myself ‘that's impossible’ because I had spent 20 some years analyzing data, and I had never seen anything like it. Clearly, it is possible and I was looking at a revolution.”

NDP’s success brings up another important issue – the need for standards. Without standards regarding how data are collected, matched, cleaned, and kept, it’s much harder to combine

various data into a large database. (For more on standards initiatives, see the article on ERM in this issue, page "ERM.") In fact, the NDP is working to bridge different automation systems by developing a schema to bring data from a Horizon system into the NDP. The problem of matching data from different systems is complex, not only in use and patron data, but also in merging locally customized cataloging data. NDP can serve as a model for library consortia and library networks wishing to share data across different automation platforms.

As projects like the Data Farm and NDP demonstrate, surfacing patterns of use means librarians need data at the individual use level. General aggregations hide underlying patterns that are available to resource vendors, who use them in making business decisions. Libraries may not be a money-making proposition, but they need the same item-level data and reporting standards to make solid management decisions and respond to demands for accountability in using public funds.

Exploring the Data

The (sometimes unexpected) challenge after gathering data is finding interesting patterns and stories within it. Creating a data warehouse doesn't alleviate this. While it puts everything in one place, the amount of data is overwhelming. Corporate America discovered this, too. After building data warehouses, managers and other stakeholders did not see any benefit to the piles of data collected in the data warehouse. They lacked data mining techniques to find patterns in the data.

Data mining is the exploration of a large dataset for non-trivial, novel, and useful patterns, using different statistical, analytical, and visualization tools. The process starts by collecting relevant data on a particular topic. The collected data is matched into a single, large database and cleaned. Cleaning takes most of the time in a data mining process and often requires repeated attempts, as data mining tools highlight flaws in the data. Data mining programs, such as Clementine, SAS Enterprise Miner, or the open-source WEKA, offer many options that can be executed on the same data set.

Data mining options can be either descriptive or predictive. *Descriptive tools* help the librarian describe and compare the past and current patterns in the data. These include:

- Traditional aggregates and averages,
- Clustering and market-basket analysis to identify groups of items or users that belong together because of other aspects in the data,
- Online analytical processing (OLAP) to explore tables of data, clicking on headings and rows to "drill down" into the data or "rolling up" data into higher level categories to better understand groupings, and
- Visualization, or graphical representations of data to highlight patterns.

Predictive tools determine the unknown from what is known. For example, past patterns of use may predict future resource needs, which is valuable in purchasing decisions. Some predictive tools include:

- Correlation and regression to find variables that go up or down together or variables that go up as others go down,
- Rule generation, or creating a series of "if..then" rules describing patterns in the data, and
- Neural networks, which take in a large number of variables and predict a single result from past performance.

Some descriptive tools can be used in a predictive manner and vice versa, and there are many other data mining tools besides these. For an accessible introduction to data mining, see Berry and Linoff's (2000) book on this topic.

Advanced Data Analysis Applications in Libraries

Some of the projects described earlier have dedicated data mining tools. The Normative Data Project, for example, includes OLAP and visualization options to allow exploration that goes beyond basic reporting.

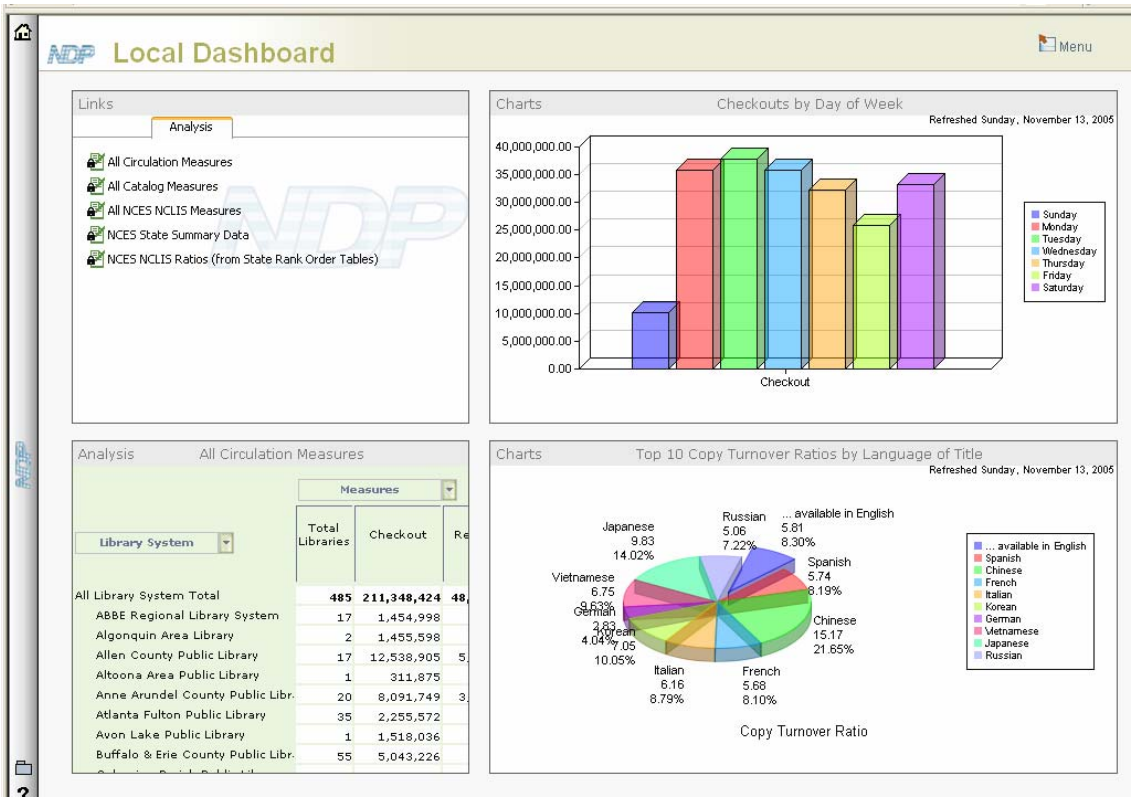

Figure 3: Multiple views of data through the Normative Data Project

Bob Molyneux reports the NDP toolkit is "liberating because it makes it possible for sophisticated analysis to be done without having to know how to program or about analytical techniques, as common sense will suffice." He adds, "What we have here, then, is a database of information working librarians can use to make better decisions. It is a decision support system. Given that public libraries in the U.S. had an income of almost $9 billion and spent $8.3 billion on operations, according to the latest estimates, better decision information will result in better decisions and better use of that money, particularly in budget-constrained times."

Like Molyneux, OCLC Research seeks to leverage library data in new ways. Current data mining projects use WorldCat and library holdings, circulation and interlibrary loan (ILL) statistics, and system transaction logs to identify collection trends, themes, and search behaviors. (See www.oclc.org/research/projects/mining for more.)

One such project is OCLC's WorldMap™, which presents WorldCat titles and holdings geographically, by state, province, and country of publication. WorldMap™ also displays other statistics, i.e., number of libraries in a country, collection, and expenditure data. Users first see a world map, select a desired variable, and WorldMap™ shades countries with colors to reflect values for that variable (Figure 4).

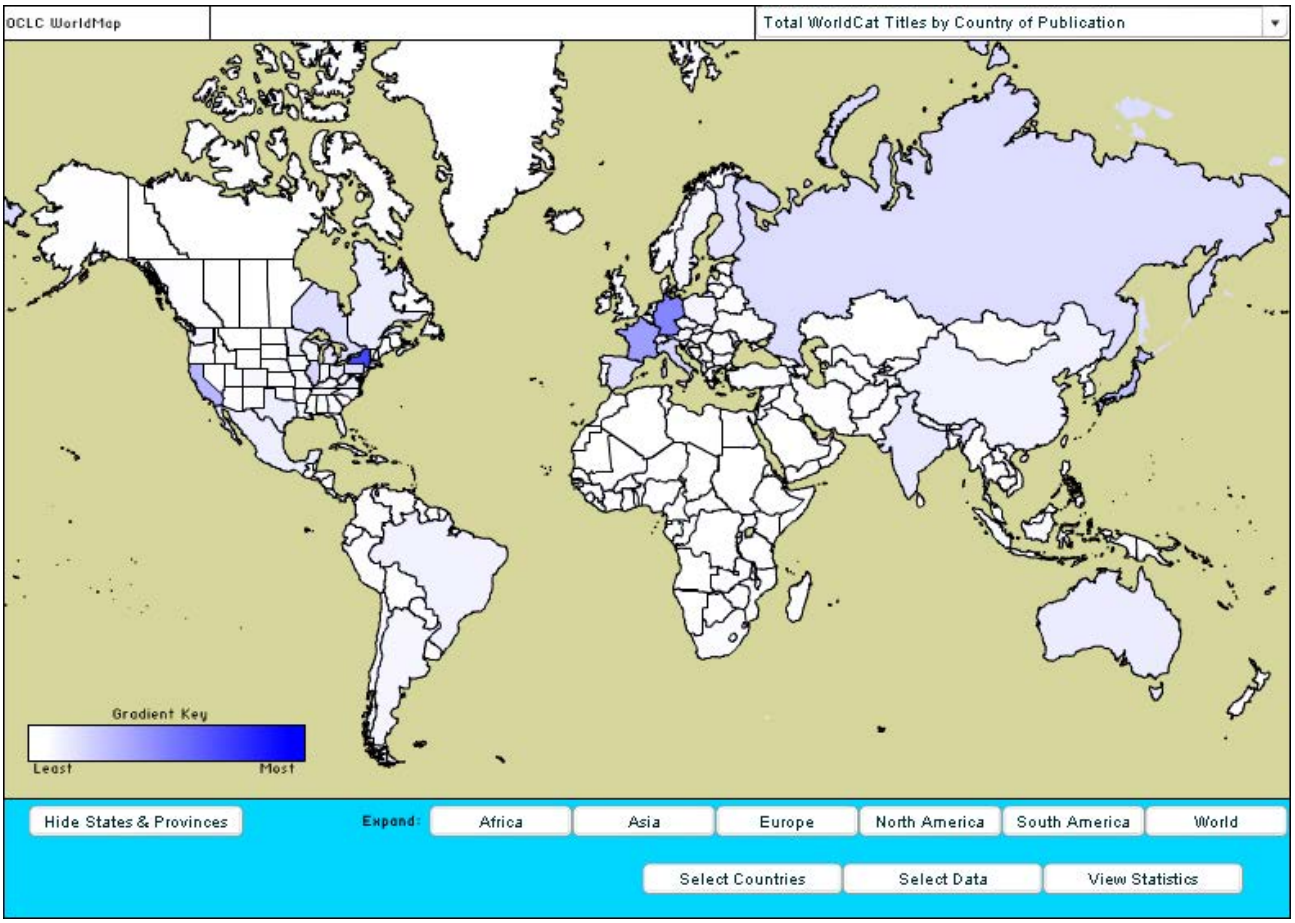

Figure 4: OCLC's WorldMap Project

The user can then select a country to see the its underlying data, either displayed on the map or in a table. Rather than starting with lists and tables, OCLC wanted to provide a graphical entry point to the data. WorldMap™ is a combination of OLAP and visualization that results in a more natural way of exploring information.

Another data visualization project is the Public Library Geographic Database (PLGDB), part of the GeoLib Program, directed by Dr. Christie Koontz, College of Information at Florida State University. PLGDB is free and allows researchers and libraries to explore map-based library data from different sources. As an example, the figure below shows a map of library locations and the percentage of the local population under age 5.

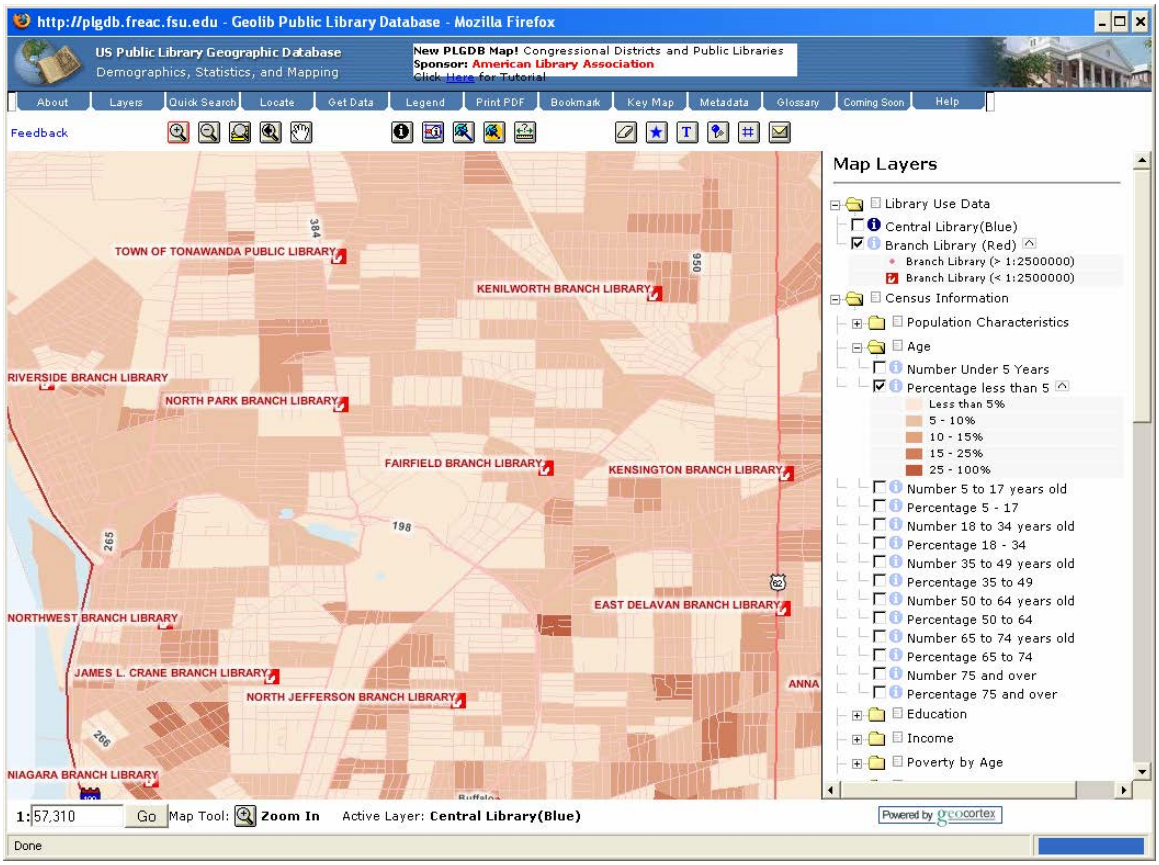

Figure 5: Map from the Public Library Geographic Database

Nancy Smith, coordinator for the Prairie Area Library System in Rockford, Illinois, uses the PLGDB for marketing and research about patron needs. “In Illinois, we have varying service area lines, which are ascribed by funders, and using the PLGDB allows librarians to estimate customer market areas or the geography in which our users really live,” Smith explains, “but the real power is in identifying potential library customers.”

Pursuing Patterns through Bibliomining

Using visualization or OLAP tools to manually explore data is akin to driving around a neighborhood looking for a dream house. A data mining tool, on the other hand, is like a real estate agent that creates a list of potential exciting choices. These data mining tools can help direct a libarian to patterns of interest in the data.

Another discipline based on patterns from library and information science is bibliometrics. Bibliometric concepts focus on patterns in the creation of works using data such as citations and authorship. The concept of bibliomining combines data mining and bibliometrics to improve library decision-making (Nicholson and Stanton, 2003). Data warehouses are at the core of bibliomining, allowing data needed for both data mining and bibliometrics to live within the same space so librarians can discover patterns that span both creation and use (Nicholson, 2006).

The library's story can be told in many ways. In the past, we have relied upon anecdotes and broad aggregations, yet stakeholders wish more evidence-based justification. Drawing together selected artifacts of our daily efforts with data-based patterns that support our anecotes and provide evidence for our decisions will allow us a more powerful way to convince funders of our value.

Scott Nicholson is an Assistant Professor at the Syracuse University School of Information Studies in Syracuse, New York.

Web Extras

University of Pennsylvania Library's Data Farm
http://metrics.library.upenn.edu/prototype/datafarm/

The Normative Data Project
http://www.libraryndp.info/

OCLC WorldMap™
http://www.oclc.org/research/researchworks/worldmap/prototype.htm

Public Library Geographic Database Mapping
http://www.geolib.org/PLGDB.cfm

The Bibliomining Information Center
http://bibliomining.com/

References

American Library Association. (2005). *The Campaign to Save America's Libraries*. Retrieved November 21, 2005, from http://www.ala.org/ala/issues/campaignsal.htm

Berry, M. J. A., & Linoff, G. (2000). *Mastering data mining : the art and science of customer relationship management*. New York: Wiley Computer Pub.

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

Murphy, D. (2003, April 7, 2003). Some Librarians Use Shredder to Show Opposition to New F.B.I. Powers. *New York Times*, p. 12.

Nicholson, S. (2006). The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services . *Information Processing and Management 42*(3), 785-804.

Nicholson, S. (2003). The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making. *Information Technology and Libraries*, *22*(4), 146-151.

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