

Making data work harder

BY LYNN SILIPIGNI CONNAWAY, Consulting Research Scientist, OCLC Research

LIBRARIES' ELECTRONIC SYSTEMS COLLECT AND store large amounts of bibliographic, statistical and other types of data. The extraction and analysis of these data can provide librarians with useful information for making informed decisions for collections, services and systems. In general, this type of activity has come to be known as "data mining."

While much of these data are collected automatically by system transaction logs, libraries have invested enormous amounts of time, effort and resources in creating rich, structured descriptions of the materials in their collections. There is a wealth of untapped value in this information, in the form of intelligence that can be used to support library decision-making. This intelligence needs to be mined and applied to create value for libraries and users... in short, libraries need to make their data work harder.

OCLC Research conducts several data mining projects as part of its Content Management and Collection and

used to calculate the intellectual level of each title by weighting the holdings counts by type of library. One collection study compared an ARL library's circulation statistics and interlibrary loan data to its WorldCat holdings, in order to provide information for collection development and remote storage decisions. WorldCat holdings also were used to identify last copies for preservation and digitization, including an analysis of the Google Print Library Project.³

Another activity, the publisher name-authority project, is underway to identify the variant forms of publisher names in WorldCat. Researchers will attempt to link the variant forms of the publisher name to the most frequently used form of the publisher name in WorldCat. The successful completion of this project will allow more in-depth data mining of resources in WorldCat by linking publishers to specific subject areas and intellectual levels.

A third project examined WorldCat records represent-

There is a wealth of untapped value in library data, in the form of intelligence that can be used to support library decision-making. This intelligence needs to be mined and applied to create value for libraries.

User Analysis research agendas. Data mining utilizes WorldCat resources and library holdings, circulation statistics, interlibrary loan (ILL) data and system transaction logs, as well as other sources, to identify collection trends and themes and user search behaviors.

In the collection studies, researchers identify and characterize book-collection holdings by library type, i.e., ARL, academic non-ARL, public, special or school. They then analyze and compare aggregate book-holdings data for each type of library by subject areas, publication dates and publishers. The OCLC Conspectus¹ and the North American Title Count² were used to determine the subject areas covered by each collection. Holdings data from WorldCat provide information that can be

ing digital resources. Once records were extracted, they were categorized by type of resource represented. This extraction was not straightforward, since digital resource cataloging practices are varied. As of July 2004, WorldCat contained at least 751,837 records for digital resources. Books, computer files and government documents represent the greatest number of digital records in WorldCat. The digital record with the lowest OCLC number in WorldCat was created on September 11, 1975 by the American Antiquarian Society, and describes a data file recorded on a single tape reel, containing 1860 and 1880 U.S. census data on residents of Worcester, Massachusetts. The digital resource represented by the highest OCLC number (at the time this work



was conducted) was created on July 1, 2004 by Mississippi State University, and describes a master's thesis, published as a PDF file. This type of analysis can help us better understand how libraries have evolved within a historical, sociological and cultural context, and warrants future research.

The OCLC WorldMap™ presents a geographical representation of WorldCat title holdings by state, province and country of publication. Possible extensions include displaying other library statistics (such as the number of libraries in a country or ARL collection and expenditure data) and/or representing data in tabular, as well as graphic, format. We hope to make this prototype system publicly available in the near future.

A fifth project analyzed NetLibrary transaction logs to identify the date and time that electronic book (eBook) titles were accessed, as well as the type of library from which they were accessed. Search words and terms, and the type of searches used, e.g., keyword, subject, author,

title, etc., were also identified. The information provided by these data resulted in design changes to the search interface and contributed to revised eBook acquisition strategies and pricing models.

Data mining provides information that can be used by librarians to make intelligent, data-driven decisions for developing and managing collections; comparing library book-collection holdings to specific types of aggregate libraries; identifying last copies for preservation, remote storage and digitization; and developing user-centered services. OCLC Research is in a unique position to help libraries make their data work harder in support of stronger libraries and better services. ■

1. The OCLC Conspetus is a subject hierarchy consisting of divisions, categories and subject descriptors.
2. North American Title Count is a database of classified titles held by participating U.S. and Canadian libraries in selected subject areas. We used the 2001 edition, produced by the Library Research Center of the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign.
3. Google Print Library Project. <http://print.google.com/googleprint/library.html>. (Accessed 20 May 2005.)