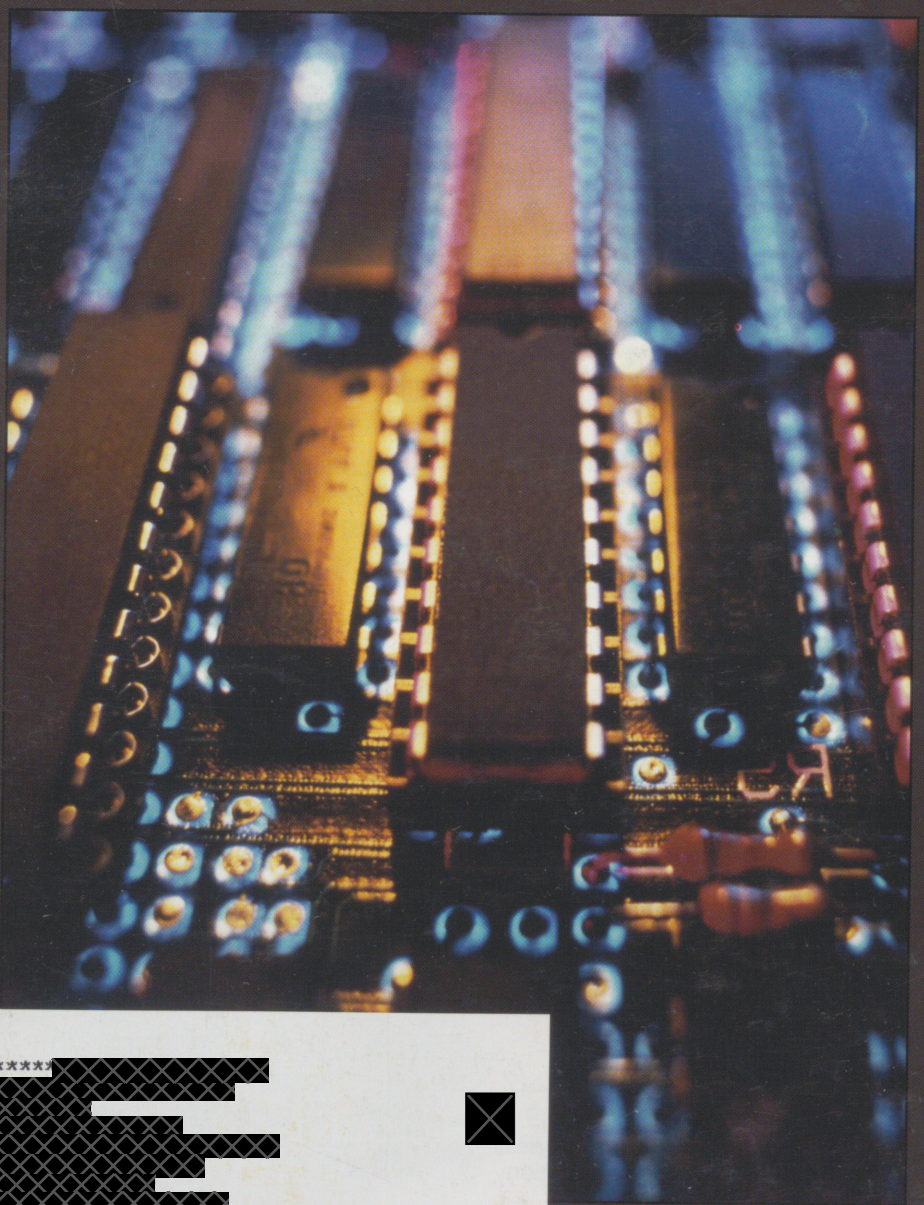
Information Technology AND Libraries

ISSN 0730-9295

Volume 22 Number 4

DECEMBER 2003



IN THIS ISSUE

- Data Warehousing and Data Mining for Library Decision Making
- Matching Subject Portals with the Research Environment
- Behavior-Based Library Recommender Systems

Most Trusted Databases
for every state and region

Online databases and print reference directories from Harris give your patrons top-quality business information in an easy-to-use, powerful interface.

Give your patrons the best resource for information on U.S. companies with a specialty database from Harris InfoSource, a D&B company. A trusted source for over 2,000 libraries, Harris has the information on U.S. businesses your patrons need accessible through intuitive web-based software.

We've been compiling business information for over 30 years, and our commitment to quality is shown in our relationships with chambers of commerce and development corporations nationwide. With a Harris database, your patrons will be able to quickly find addresses, phone numbers, contact names and more on U.S. companies.

Try our database for yourself and discover why:

- Harris electronic databases are the easiest to use on the market.
- These exclusive specialty databases provide the right option for your library.
- Your patrons' searches will be made simple with less hassle for you.

With databases available in a variety of geographic areas and specialty markets, you will find the product that fits your patrons' needs—and your budget.

MANUFACTURING DATABASES

The premier source of information on America's industrial sector

Published in cooperation with the National Association of Manufacturers

MAJOR COMPANIES *Market Driver*™ DATABASE

The top U.S. firms that drive business spending

ALL-BUSINESSES DATABASES

Manufacturing and service firms combined in a single database

REFERENCE DIRECTORIES

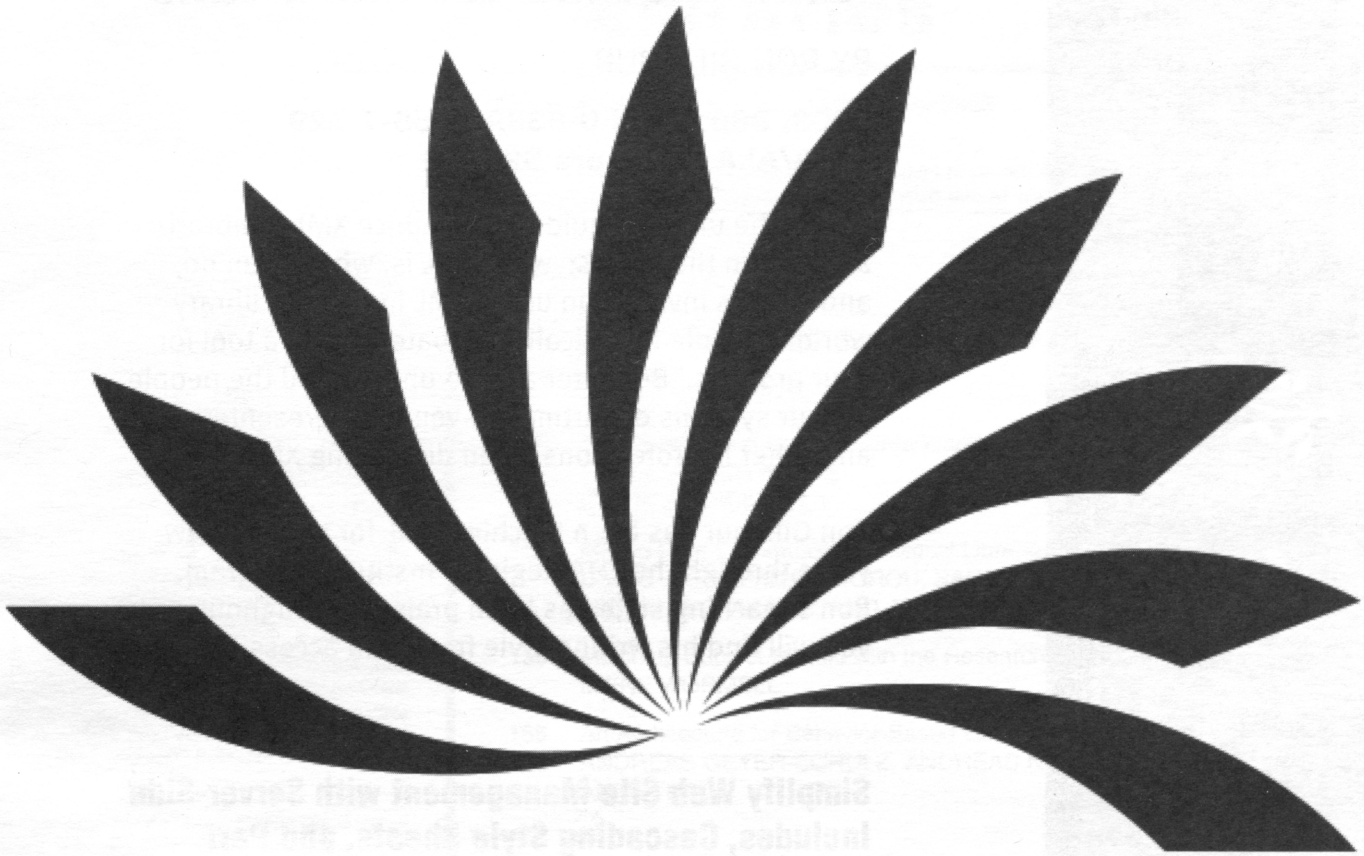
Classic print reference directories cover the nation



Call **800.888.5900** to set up a complimentary trial

Go to www.harrisinfo.com/libraries for more information

HARRIS INFOSOURCE™
a D&B company



Content Transformation Solutions

- **SGML, HTML, XML**
- **DTD Development**
- **Imaging**
- **Web-site Development**
- **eBook & PDF**
- ***and so much more...***



Information Publishing Group, 1930 SouthWest 48th Lane, Cape Coral, FL 33914
239.549.2384. • iris.hanney@techbooks.com • www.techbooks.com

LITA Guides

Reaches Out Serves Educates

All LITA publications and a printable order form can be found at www.lita.org/litapubs
Fax orders to (770) 442-9742 or call 1-866-746-7252.

XML: A Guide for Librarians

BY RON GILMOUR

2003. 88p. ISBN 0-8389-8255-7 $29 (LITA/ALA Members $26.10)

This is the ultimate guide to introduce XML to librarians. Learn the basics: what XML is, what it can do, and what is involved in using XML in the real library world. Be able to critically evaluate XML as a tool for your projects. Be better able to understand the people in your systems departments, vendor representatives, and other IT professions when discussing XML.

Ron Gilmour has been teaching XML for the past few years through the LITA Regional Institutes program. Ron's teaching style has been praised throughout and you will find his writing style fresh and accessible.

Simplify Web Site Management with Server-Side Includes, Cascading Style Sheets, and Perl

BY ANDREA PETERSON

2002. 76p. ISBN 0-8389-8199-2 $29 (LITA Members $26.10)

Open Source Software for Libraries

BY CINDY MURDOCK, ERIC SCHNELL, KEVIN S. CLARKE, GORDON PAYNTER, ERIC LEASE MORGAN

2002. 72p. ISBN 0-8389-8203-4 $29 (LITA Members $26.10)

Usability Assessment of Library-Related Web Sites: Methods and Case Studies

EDITED BY NICOLE CAMPBELL

2001. 124p. ISBN 0-8389-8157-7 $25 (LITA Members $22.50)

Getting the Most Out of Web-Based Surveys

BY DAVID WARD

2000. 60p. ISBN 0-8389-8108-9 $20 (LITA Members $18.00)

LITA Publications

Library & Information Technology Association
a division of the American Library Association

Dan Marmion, Editor
University of Notre Dame, Notre Dame, Indiana (dmarmion@nd.edu)
Judith Carter, Managing Editor
OCLC, Inc., Dublin, Ohio (carterj@oclc.org)
Tom Zillner, Book Review Editor
Wisconsin Library Services, Madison, Wisconsin (tzillner@wils.wisc.edu)
Andy Boze, Software Review Editor
University of Notre Dame, Notre Dame, Indiana (Boze.1@nd.edu)

EDITORIAL BOARD: **Judith Gelernter,** The Union Club, New York; **Eddy Hogan,** California State University, Monterey Bay; **Alan Manifold,** Purdue University; **Scott Nicholson,** Syracuse University; **Paul Soderdahl,** University of Iowa; **Marc Truitt,** University of Houston; **Joan Frye Williams,** Information Technology Planning, Management; **Melinda Stowe,** University of Rochester, ex officio.

Information Technology and Libraries (ISSN 0730-9295) is published quarterly in March, June, September, and December by the American Library Association, 50 E. Huron St., Chicago, IL 60611. It is the official publication of the Library and Information Technology Association, a division of the American Library Association. SUBSCRIPTIONS: Subscription price, $30, is included in membership dues. Nonmembers may subscribe for $55 per year in the U.S.; $60 in Canada, Mexico, Spain, and other PUAS countries; $65 in other foreign countries. Send subscription requests to *Information Technology and Libraries*, Subscription Department, American Library Association, 50 E. Huron St., Chicago, IL 60611; 800-545-2433; fax: (312) 944-2641; subscription@ala.org. Single copies, $20. Periodical-class postage paid at Chicago, Illinois, and at additional mailing offices. POSTMASTER: Send address changes to *Information Technology and Libraries*, 50 E. Huron St., Chicago, IL 60611.

Information Technology and Libraries publishes material related to all aspects of libraries and information technology, including digital libraries, metadata, authorization and authentication, electronic journals and electronic publishing, telecommunications, distributed systems and networks, computer security and intellectual property rights, technical standards, geographic information systems, desktop applications, online catalogs and bibliographic systems, optical information systems, software engineering, universal access to technology, futuristic forecasting, library consortia, vendor relations, and technology and the arts. *ITAL* welcomes unsolicited manuscripts. Submissions should follow the guidelines stated under "Instructions to Authors" on page 95 of the June issue or on the LITA Web site at www.lita.org/ital. Manuscripts of feature articles, communications, and tutorials should be addressed to Dan Marmion, 221 Hesburgh Library, The University of Notre Dame, Notre Dame, IN 46556-5629. Books for review should be addressed to Tom Zillner, WILS, 464 Memorial Library, 728 State St., Madison, WI 53706. Software for review should be addressed to Andy Boze, Hesburgh Library, The University of Notre Dame, Notre Dame, IN 46556-5629.

ADVERTISING: The Goldman Group, 14497 N. Dale Mabry Hwy., Ste. 205N, Tampa, FL 33618; telephone (813) 264-2772; fax (813) 264-2343.

As a matter of policy, *ITAL*, as the scholarly organ of LITA, does not review LITA publications.

PRODUCTION: ALA Production Services (Troy D. Linker, Kevin Heubusch; Ellie Barta-Moran, Angela Hanshaw, Kristen McKulski, and Karen Sheets), American Library Association, 50 E. Huron St., Chicago, IL 60611.

Publication of material in *Information Technology and Libraries* does not constitute official endorsement by LITA or the ALA.

Abstracted in *Computer & Information Systems, Computing Reviews, Information Science Abstracts, Library & Information Science Abstracts, Referativnyi Zhurnal, Nauchnaya i Tekhnicheskaya Informatsiya, Otdyelnyi Vypusk,* and *Science Abstracts Publications.* Indexed in *CompuMath Citation Index, Computer Contents, Computer Literature Index, Current Contents/Health Services Administration, Current Contents/Social Behavioral Sciences, Current Index to Journals in Education, Education, Library Literature, Magazine Index, NewSearch,* and *Social Sciences Citation Index.* Microfilm copies available to subscribers from Proquest, Ann Arbor, Michigan.

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1992.×

Information Technology AND Libraries

Volume 22, Number 4 December 2003 ISSN 0730-9295

Abstracts and contents pages of recent issues of *ITAL* can be found on the LITA Web site at www.lita.org/ital.

144 Guest Editorial: Introduction to This Special Issue on the Bibliomining Process
SCOTT NICHOLSON

146 The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making
SCOTT NICHOLSON

152 Mining User Communities in Digital Libraries
CHRISTOS PAPATHEODOROU, SARANTOS KAPIDAKIS, MICHALIS SFAKAKIS, AND ALEXANDRA VASSILIOU

158 Matching Subject Portals with the Research Environment
IRENE WORMELL

165 An Architecture for Behavior-Based Library Recommender Systems
ANDREAS GEYER-SCHULZ, ANDREAS NEUMANN, AND ANKE THEDE

COMMUNICATIONS

175 Traces in the Clickstream: Early Work on a Management Information Repository at the University of Pennsylvania
JOE ZUCCA

179 A Study of the Use of the Carlos III University of Madrid Library's Online Database Service in Scientific Endeavor
CARLOS A. SUÁREZ-BALSEIRO, ISABEL IRIBARREN-MAESTRO, AND ELÍAS SANZ CASADO

183 Mapping the Output of Topical Searches in the Web of Knowledge and the Case of Watson-Crick
EUGENE GARFIELD, A. I. PUDOVKIN, AND V. S. ISTOMIN

189 Index to Volume 22

192 Index to Advertisers

cover design Kevin Heubusch ● **interior design** Dianne M. Rooney

Copyright ©2003 American Library Association.

All material in this journal subject to copyright by ALA may be photocopied for the noncommercial purpose of scientific or educational advancement granted by Sections 107 and 108 of the Copyright Revision Act of 1976. For other reprinting, photocopying, or translating, address requests to the ALA Office of Rights and Permissions.

Guest Editorial: Introduction to This Special Issue on the Bibliomining Process

Scott Nicholson

Libraries are under siege. Material costs continue to rise, forcing more selective collection development; funding agencies demand data-based justification for services; and the Internet is replacing the library as the primary source for information. In addition, legal threats to the privacy of patron records are pushing librarians to take extremely destructive action to their institutional records.

One solution to many of the problems is the bibliomining process. Libraries can use a data warehouse to capture information about their materials without associating personally identifiable information with their use. Statistical and data-mining techniques can then be used with the data warehouse to understand patterns of use. Understanding these patterns allows for:

- better decisions about collection development;
- thorough data-based justification of library services;
- customized library services to compete with the Internet; and
- a more complete understanding of how the library is used.

These methods and tools are heavily used in the corporate sector to systematically capture, clean, and customize services for users. The significant difference between corporate use and the use of these tools in libraries is that of user privacy; libraries are more interested in patterns of use exhibited by groups than in behaviors exhibited by individuals. By definition, a pattern is something done by a group of people; therefore, the goal is only to discover patterns of behavior. During the data-cleansing process, the connections are broken so that these patterns cannot be traced back to individual users.

This is not the first special issue on this topic; in 1996, *Library Administration and Management* had a themed issue on the mining of library data. However, most of the data came from automation systems. While those systems are still a valuable resource for data, the introduction of digital library services through the Internet has greatly increased the amount of data available. As the percentage of library budgets spent on electronic materials grows, so does the need for updated management tools to understand the use of those materials. The goal of this special issue is to raise awareness of some current and forthcoming methods and tools useful in understanding library use.

Scott Nicholson (srnichol@syr.edu) is an Assistant Professor at the Syracuse (N.Y.) University School of Information Studies.

Two articles in this issue deal primarily with data warehousing. The first, "The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making," presents an overview of the entire process with some discussion of methods that can be employed to create the data warehouse. The other article is "Traces in the Clickstream: Early Work on a Management Information Repository at the University of Pennsylvania," written by Joe Zucca, who has been in charge of the construction of a data warehouse at the University of Pennsylvania Library.

Also included in this special issue are two brief communications about some applications of the bibliomining process. In "A Study of the Use of the Carlos III University of Madrid Library's Online Database Service in Scientific Endeavor," Carlos A. Suárez-Balseiro and others from Madrid explore patterns of e-journal use by different departments in an academic setting. Eugene Garfield, founder of the Institute for Scientific Information and the Web of Knowledge, and his colleagues present a new way of using citation data with visual data mining through HistCite to gain a clearer understanding of essential literature on a topic.

This special issue includes three full articles exploring different bibliomining methods for gaining insight from library data. Christos Papatheodorou and his coauthors from Greece describe two methods of extracting profiles of user groups in "Mining User Communities in Digital Libraries." Irene Wormell from Sweden presents a technique for developing library portals that are representative of user needs in "Matching Subject Portals with the Research Environment." Finally, in "An Architecture for Behavior-Based Library Recommender Systems," Geyer-Schultz and his colleagues from Germany present a system that will recommend works for users based upon patterns of past use.

This international collection of articles all use library data to help in library decision-making processes. Some of these research articles are focused on service-building while others are designed to assist library personnel in meeting the needs of users. The one common factor in all of these articles is the need to maintain library data. Library administrators contemplating the wholesale deletion of library records are advised to consider the long-term impact of their decision; the interim step of a data warehouse can maintain some data important for decision-making and the building of services while still protecting user confidentiality.

To learn more about this topic, visit the Bibliomining Information Center at http://bibliomining.org.



If you have been searching for an easy way to authority control your library's current cataloging, try LTI's ***Authority Express*** service.

With ***Authority Express,*** a library uses the Internet to transmit a file of newly cataloged bibliographic records to LTI (via ftp). LTI immediately processes the records through its state-of-the-art authority control system. Then, at the library's convenience, it logs into LTI's ftp server to retrieve fully authorized catalog records, along with linked LC name and subject authority records.

Authority Express

- Keeps authority control current at an affordable price
- Integrates easily into existing workflows
- Lowers cost by reducing staff time spent on catalog maintenance
- Provides 1 hour turn around for up to 1,000 catalog records

"Authority Control for the 21st Century"



LIBRARY TECHNOLOGIES, INC.

2300 Computer Avenue, Suite D-19 Willow Grove, PA 19090
(215) 830-9320 Fax: (215) 830-9422
(800) 795-9504 email: LTI@LibraryTech.Com

Visit our website at:
www.LibraryTech.Com

The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making

Scott Nicholson

Bibliomining, or data mining for libraries, is the application of data mining and bibliometric tools to data produced from library services. This article outlines the bibliomining process with emphasis on data warehousing issues. Methods for cleaning and anonymizing library data are presented with examples.

One of the challenges put forth to the library profession by Michael Buckland is to gain a better understanding of library user communities.[1] Most current library evaluation techniques focus on frequencies and aggregate measures; these statistics hide underlying patterns. Discovering these patterns is key in understanding the communities that use library services. In order to tailor services to meet the needs of different user groups, library decision-makers can use the bibliomining process to uncover patterns of data-based artifacts of use. The bibliomining process, which consists of data warehousing and data mining, will be explored in this brief article.

The term bibliomining was first used by Nicholson and Stanton in discussing data mining for libraries.[2] In the research literature, most works that contain the terms library and data mining are not talking about traditional library data, but rather using library in the context of software libraries, as data mining is the application of techniques from a large library of tools. In order to make it more conducive for those concerned with data mining in a library setting to locate other works and other researchers, the term bibliomining was created. The term pays homage to bibliometrics, which is the science of pattern discovery in scientific communication.

Bibliomining is the application of statistical and pattern-recognition tools to large amounts of data associated with library systems in order to aid decision-making or justify services. The bibliomining process consists of:

- determining areas of focus;
- identifying internal and external data sources;
- collecting, cleaning, and anonymizing the data into a data warehouse;
- selecting appropriate analysis tools;
- discovery of patterns through data mining and creation of reports with traditional analytical tools; and
- analyzing and implementing the results.

The process is cyclical in nature: as patterns are discovered, more questions will be raised that will start the process again. As additional areas of the library are explored, the data warehouse will become more complete, which will make the exploration of other issues much easier.

Scott Nicholson (srnichol@syr.edu) is an Assistant Professor at the Syracuse (N.Y.) University School of Information Studies.

Determining Areas of Focus

The first step in the bibliomining process is to determine the area of focus. This area might come from a specific problem in the library or may be a general area requiring exploration and decision-making. The first decision is whether to conduct directed or undirected data mining.[3] Directed data mining is problem-focused: there is a specific problem that drives the exploration—for example, "Budget cuts have reduced the staff time for contacting patrons about delinquent materials. Is there a way to predict the chance patrons will return material once it is one week late in order to prioritize our calling lists?" Undirected data mining is used when a library manager wants to get a better idea of a general topical area; one possible area of focus would be: "How are different departments and types of patrons using the electronic journals?"

There are several dangers with undirected data mining, as this type of exploration involves the use of many tools repeatedly to locate any available patterns. Many data-mining techniques are based on probability, and therefore there is a small chance that each pattern found is incorrect: the more times this type of tool is used, the greater the chance of finding an invalid pattern. Another problem arises with data-mining tools that locate the most frequently occurring patterns: in a complex data set, the most frequent pattern may only occur a few times. Therefore, all patterns that are found will need to be explored to make sure they are valid, make sense to the librarians involved, and are not trivial. Undirected data mining may produce an overwhelming number of patterns to explore. Given the time it takes to collect and clean the data, undirected data mining should be considered only when a strong data warehouse is in place.

Identifying Data Sources

After determining the areas of focus, the next step is to identify appropriate data sources. The bibliomining process requires transactional, nonaggregated, low-level data. This can be made much more difficult, if not impossible, if librarians delete the operational data in an attempt to protect user privacy. For example, the *New York Times* recently reported the daily destruction of library records by public library staff members.[4] However, the data warehouse can be used to maintain a data-based his-

tory of the library while at the same time protecting the privacy of patrons; discarding everything makes it impossible to provide the data-based justification of the use of library services. As funding for public services continues to be cut, maintaining the ability to justify and defend the existence of library services is essential.

There are two types of data sources that should be considered. Internal data sources are those already within the library system. As most library systems have their data stored in different data silos around the organization (patron database, transactional data, Web server logs), discovering and extracting the internal data might be challenging.

External data sources are those that are not located within the library system. In an academic setting, these might include demographic information related to a specific ID number that is located in the computer center or personnel management system. For a public library setting, this might include demographic information for zip codes from census data.

Creating the Data Warehouse

Assuming that the data do exist, the challenge remains of combining different data sources into a single data warehouse that does not contain publicly identifiable information. A data warehouse is a database that is separate from the operational systems and contains a cleaned and anonymized version of the operational data reformatted for analysis. To create the warehouse, the librarian writes queries to extract the data from the identified sources, combines those data using common fields, cleans the data, and writes the resulting records into either a flat file or a relational database designed specifically for analysis.[5] Once this procedure has been tested, it can be automated to pull data from the operational systems into the data warehouse on a regular basis.

Protecting Patron Privacy

On the surface, it may seem that the data warehouse will destroy patron privacy by combining different library systems into one source. The *New York Times* reported a recommendation that libraries delete all transactional data; this also makes it very difficult to evaluate and justify library services.[6] Other libraries take no action, in which case the data lingers in operational systems and on backup tapes. Going through the data warehousing process requires the library to examine their data sources; by explicitly determining what to keep and what to destroy, libraries can save the demographic information needed to evaluate communities of users without keeping records of the individuals in those communities.

This extraction and cleaning process is the key to protecting patron privacy during data warehousing. As the records are drawn from internal and external systems, matches are made to connect data, and then the personally identifiable information is discarded. This personal information should never be put into the data warehouse, so it will not be backed up, saved, or otherwise archived. After the data warehouse is created, the original data can then be deleted, in accordance with current advice to protect the privacy of patrons. The goal of data warehousing is to create a data source that contains decision-making information that cannot be used to recreate the original transactional records.

Here are two examples of the cleaning process that maintain important information for decision-making without keeping personally identifiable information:

- When an item is returned, many libraries delete all information about that transaction. However, there is valuable decision-making information that is lost. While the operational system is a user-focused data source, the data warehouse is an item-focused data source. Therefore, before deleting the transactional information, a record should be created in the data warehouse that combines information about the item with demographic information about the patron. This will capture the important information about the transaction without identifying the patron involved, and will still allow the discovery of patterns about patrons without compromising personally identifiable information (see figure 1—follow the shaded cells to see how data are replaced).
- The server transaction log contains invaluable information about how the Web-based services of the library are used. These logs contain IP addresses that could be used to track the individual computers used to access these services. The useful information that does not track personally identifiable information consists of which pages were looked at in the same session. Instead of deleting the IP address field, some type of code should be used to replace each unique IP address. One way to protect privacy is to replace identical IP addresses within a certain time window with a code generated by the first time/date stamp associated with the IP. For example, an IP in the log registered at 10:32/10-29-02 could be coded as 102902-1032-A. Activity coming from that IP without a fifteen-minute break would then get the same code.

  In addition, some transaction logs will keep login information; just as with circulation records, these logins should be replaced with the matching demographic information about the user in the data warehouse (see figure 2).

Whenever capturing demographic information, librarians must ensure that no combination of the demographic

variables will lead to the identification of an individual. This should be explored in the development of these capturing tools; demographic groups should be combined or dropped until all combinations of the saved demographics lead to groups of patrons (figure 3). These demographic categories should be audited on a regular basis and adjusted as needed to represent the current user population.

Librarians are advised to consult with legal counsel before starting a data-warehousing project. Privacy of

Original Circulation Records		
Book ID	Subject	Patron
QA76.9	Computer Science	392-33
PS159.G8	American Literature	575-49
HF5415.125	Marketing	392-33

Original Patron Database			
Patron	Name	Class	Dept.
373-34	Abby Lavender	Grad	Psych
392-33	Kenneth Moore	Ugrad	Math
575-49	Sophie Richards	Faculty	English

Data Warehouse-Combined Cleaned Circulation Records			
Book ID	Subject	Patron Class	Patron Dept.
QA76.9	Computer Science	Ugrad	Math
PS159.G8	American Literature	Faculty	English
HF5415.125	Marketing	Ugrad	Math

Figure 1. Cleaning Transactional Records

Original Web Server Transaction Log

IP Address	Time/Date	Referring Page	Page Retrieved
12.90.201.23	10:32/10-29-02	Google.com	Index.html
98.28.189.49	10:33/10-29-02	Index.html	Resources/oclc.asp
12.90.201.23	10:35/10-29-02	Index.html	Reference.html
12.90.201.23	10:36/10-29-02	Reference.html	Databases.html
98.28.189.49	10:37/10-29-02	Firstsearch.html	Resources/oclc.asp

Data Warehouse—Cleaned Web Transaction Records

IP Identifier	Time/Date	Referring Page	Page Retrieved
102902-1032-A	10:32/10-29-02	Google.com	Index.html
102902-1033-A	10:33/10-29-02	Index.html	Resources/oclc.asp
102902-1032-A	10:35/10-29-02	Index.html	Reference.html
102902-1032-A	10:36/10-29-02	Reference.html	Databases.html
102902-1033-A	10:37/10-29-02	Firstsearch.html	Resources/oclc.asp

Figure 2. Cleaning Web Server Transactional Records

patron records is currently legislated on a state-by-state basis in the United States; therefore, librarians may have to adjust these techniques to meet the laws of their state. Libraries attached to a research institution may need to get clearance from their institutional research board before engaging in this type of examination.

Building the Data Warehouse

Building the data warehouse takes much more time than mining the data.[7] This can make the process seem never-ending. It is essential that all involved realize that this portion of the process is an investment; once the programs for collection and cleaning are written, they will require only minor maintenance in the future as long as the library uses the same system. Because of this initial time investment, it is suggested to start with a narrowly defined bibliomining topic and work through the entire process. This iterative process also has the advantage of allowing those developing the data warehouse to improve their collection and cleaning algorithms early in the life of the bibliomining project.

Selecting Appropriate Analysis Tools

Once the data warehouse has been developed, the stage is set for analysis. As mentioned earlier, traditional aggregations and ratios can be easily calculated to create traditional reports. However, interesting and useful patterns may be hiding behind these aggregate measures; these patterns allow library managers to better understand their individual user groups. For example, use of interlibrary loan (ILL) data for collection development can be problematic due to the short-term needs of some ILL users. Bibliomining could be used to look for patterns in ILL use in order to remove the short-term ILL needs from consistent collection deficiencies.

Management Information Systems (MIS)

These systems provide a manager with the ability to ask basic questions of the data. Many integrated library system (ILS) packages have some type of basic MIS built in. However, the data powering these systems comes from the operational systems and have not been cleaned or matched to any external data. In addition, most databases attached to libraries are treated as independent, unconnected data silos. Therefore, an MIS built on top of a data warehouse made for the library will be much more powerful and provide information that the library needs to see (instead of what the ILS vendor wants them to see).

Another addition to MIS is a critical factor alert system. As managers and administrators work with the reports produced from the data, key variables that provide the pulse of the library can be identified. Automatic notification programs can be attached to these variables, so that if they stray outside of a specified zone, the managers can be automatically notified of a potential issue. For example, if hourly circulation is below or above a certain level, a manager could be immediately notified so staffing changes could be made.

Online Analytical Processing (OLAP)

OLAP puts an interactive view of the data on the decision-makers desktop. Under the surface, the OLAP tool has run thousands of database queries to combine all of the selected variables along with all of the selected measures. These reports are presented in an easy-to-understand menu-driven system. The user will pick one of many variables from a list to examine, such as use of e-journals, and can then select basic dimensions, such as time and subject. OLAP will then present a high-level view of this data in a tabular report, perhaps by year and general classification. The user can then click one of the dimensions to expand the report—in the example above, if the user clicked on a year, the tool would expand the year into quarters, leaving the subject headings the same and recalculating the data. The user can then click on another field to drill down into the data.

OLAP tools are powered by a data warehouse. All of the fields are defined ahead of time, and the system runs many queries before anyone uses it. Therefore, response to the manager using the OLAP front-end for reports is instant, which encourages exploration. During exploration, the manger can capture any view of the data and turn it into a regular report.

Data Mining

Another way to analyze the data, and part of the inspiration of the term bibliomining, is data mining. The goal of data mining is the discovery of valid, novel, and actionable patterns in large amounts of data using statistical and artificial intelligence tools.[8] Data mining came out of corporate America in the early 1990s after data warehousing became popular; companies had stored large amounts of data and looked for ways to take advantage of these warehouses.

There are two main categories of data-mining tasks: description and prediction. In description, the goal is to understand the data from the past and the present. The patterns discovered are used to seek out affinity groups of variables common to different patrons or clusters of

Original Demographics

	Ugrad	Grad	Faculty
English	27	5	8
C. Sci.	14	3	7
Math	33	1	6
Psych.	24	6	7
Bus.	24	14	5

Cleaned Demographics

	Ugrad	Grad	Faculty
English	27	5	8
C. Sci./Math	47	4	13
Psych.	24	6	7
Bus.	24	14	5

Figure 3. Collapsing Demographic Variables

demographic groups that exhibit certain characteristics. Prediction is used to make a statement about the unknown based upon what is known. It can be used to predict the future or to make statements about the present. The two types of tasks for prediction are classification, where the goal is to place an item into a category; and estimation, where the goal is to produce a numeric value for an unknown variable. The task selected will dictate the tools chosen.

Data Mining Software Packages

There are a number of software packages that provide access to different data-mining tools. For most of these packages, one imports the data in a flat file or other common data format. Once the data have been imported, the analyst can implement different tools and explore results. The statistical packages SAS and SPSS both have optional data-mining programs, although these can be quite expensive. An open source data-mining tool is Weka, which contains many of the same options at a much more reasonable cost; however, it is not as user-friendly.[9]

The simplicity of running model after model holds a danger. Many of the individual tools are based on probabilistic statistical methods. The methods require certain types of data, such as data from a normal distribution, in order to provide accurate results. The danger is that the tools do not have any type of check for these requirements built into the system. Therefore, an unsuspecting user can run the tools on the data and the tools will provide an answer; however, that answer may not be meaningful if the data do not meet certain assumptions. To learn more about the individual algorithms in a data-mining package, one can refer to the texts by Barry and Linoff.[10] Information about the free Weka data-mining suite can be found in Witten and Frank.[11]

Analysis and Implementation

Once the results, either static reports or decision-making models and aids, have been developed, they must be validated. The first step is to test and tweak the model with data that were not used during the development process; this technique allows for the creation of a more robust model. The most important validation, however, is to have a librarian who is familiar with that particular library context examine the models. The librarian should concur with the results found; the patterns can be the data-based validation of the tacit knowledge gained from working in a library situation. If the librarian disagrees with a pattern, it is essential to explore the data streams carefully; this type of disagreement usually stems from a flaw in the data or a misapplication of a tool.

The final step is to implement the report/model. It is essential to monitor the variables that power the models over time; if the mean of a variable strays too far because of changes in the library, the model may have to be reevaluated.

Conclusion

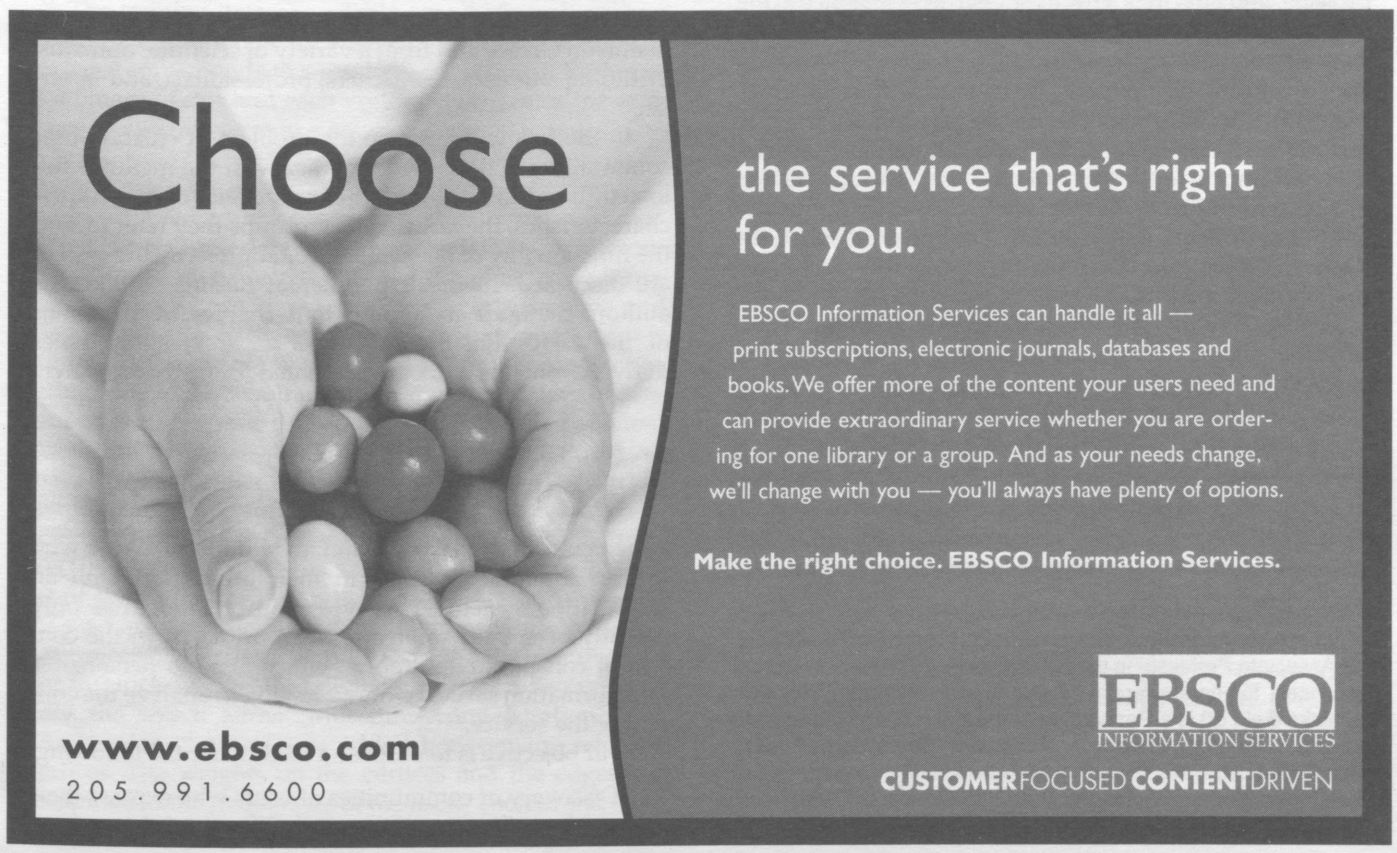
The goal of this article was to explain the bibliomining process. Emphasis was placed on data warehousing and patron privacy issues because they are required before anything else can begin. It is essential to capture our data-based institutional records while at the same time protecting the privacy of users. By using a data warehouse, both goals can be met. Once the data warehouse is in place, the library can use a plethora of reporting and exploration tools to gain a more thorough knowledge of their user communities and resource utilization.

References

1. Michael Buckland, "Five Grand Challenges for Library Research," *Library Trends* 51, no. 4 (spring 2003): 675–86.

2. Scott Nicholson and J. Stanton, "Gaining Strategic Advantage through Bibliomining: Data Mining for Management Decisions in Corporate, Special, Digital, and Traditional Libraries," in *Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance*, H. Nemati and C. Barko, eds. (Hershey, Penn.: Idea Group Pub., 2003).

3. M. Berry and G. Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support* (New York: Wiley, 1997).
4. "Librarians Use Shredder to Show Opposition to New FBI Powers," *New York Times*, Apr. 4, 2003.
5. W. Inmon, *Building the Data Warehouse*, 3d ed. (New York: Wiley, 2002).
6. "Librarians Receive Advice on Law and Reader Privacy," *New York Times*, Dec. 12, 2002.
7. M. Berry and G. Linoff, *Mastering Data Mining* (New York: Wiley, 2000).
8. Berry and Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support*.
9. I. Witten and E. Frank, *Practical Machine Learning Tools and Techniques with Java Implementations* (San Francisco: Morgan Kaufmann, 1999).
10. Berry and Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support*; Berry and Linoff, *Mastering Data Mining*.
11. Witten and Frank, *Practical Machine Learning Tools and Techniques*.



Mining User Communities in Digital Libraries

Christos Papatheodorou, Sarantos Kapidakis, Michalis Sfakakis, and Alexandra Vassiliou

The interest in the analysis of library user behavior has been increasing rapidly since the advent of digital libraries and the Internet. In this context, the authors analyze the queries posed to a digital library and recorded into the Z39.50 session log files, and construct communities of users with common interests, using data-mining techniques. One of the main concerns of this study is the construction of meaningful communities that can be used for improving information access. Analysis of the results brings to the surface some of the important properties of the task, suggesting the feasibility of a common methodology.

Information services aim to satisfy the needs of their users in a way that ensures precision and effectiveness. Many of these services use intelligent information retrieval and filtering techniques to personalize and customize their content to the users' interests and preferences.[1] Several information providers exploit user modeling techniques to understand and evaluate the usage of their services. The study of user behavior has become a crucial point in a number of digital library projects, and specifies a number of critical factors during the design and development process of a digital library.[2] Data mining offers powerful techniques for discovering nontrivial and useful patterns in voluminous datasets. Many such techniques have been applied to information services and especially to the Web, offering personalization and improving information access.[3] The application of such techniques in library systems comprises the bibliomining domain, which aims to upgrade almost all the decision-making processes concerning library and information management.

The goal of this paper is to show how the administrator of a digital library can analyze user behavior and extract the data necessary for improving information access. In particular, the authors are interested in formulating community models, which represent patterns of usage of digital libraries and can be associated with different types of users. The authors' main concern is the association of the digital library queries. Similar queries, recorded into Z39.50 session log files, are grouped into clusters. The clusters map user community models that represent the users' demands and querying habits. Such models could be used in a query-expansion process contributing to efficient retrieval. Generally, the query analysis can be beneficial to both the digital library and its users in many ways:

- *Service optimization*—helping the administrators reorganize the digital library content, authorities, and user interfaces, making them more suitable for different user groups
- *Decision support*—helping the administrator form an effective query expansion strategy for the digital library
- *Personalization*—helping the users identify information of interest to them by recommending similar subjects

This project is motivated by the need to improve the querying mechanisms provided by the digital library of the Hellenic National Documentation Center (NDC) (http://theses.ndc.gr), which is one of the most significant in Greece, consisting of many collections that are unique world wide and of international interest. The digital library of NDC is targeted to a number of user groups, mainly in Greece and from a variety of scientific domains, including students, researchers, professionals, and librarians.

In the following section the problem of creating user communities is described together with the methods followed. In addition, the digital collections of NDC, their characteristics, the targeted user groups they refer to, and the functionality of the available operations by the system are discussed. In the Experimental Results section, the authors' methods are applied to two different collections of the NDC. Finally, a number of interesting issues derived from this work are presented for further research.

Problem and Method

User communities correspond to groups of users who exhibit common behavior in their interaction with an information system.[4] This type of user model has only recently been paid some attention, primarily in the context of collaborative filtering, which aims to personalize an information service without having to analyze the content of the service.[5]

Our objective is to propose a method for the following:

- Discovery of communities of users who express their needs using common vocabulary (mainly in con-

Christos Papatheodorou (papatheodor@ionio.gr) is an Assistant Professor and **Sarantos Kapidakis** (sarantos@ionio.gr) is an Associate Professor in the Department of Archive and Library Sciences, Ionian University, Corfu, Greece. **Michalis Sfakakis** (msfaka@ekt.gr) is a Researcher at the National Documentation Center, Athens, Greece. **Alexandra Vassiliou** (alex@lib.demokritos.gr) is Head of the Documentation Department at the Library, National Center for Scientific Research, Demokritos, Athens, Greece.

trolled terms, such as subject headings or authors, where terms are selected from a big global authority file), by clustering the logged queries. Each community is described by a subset of the digital library authority file.
- Recommendation of particular terms for query expansion. The suggested terms correspond to relevant search terms used by the community users.

Consider the community discovery problem as a typical data-mining task. Hence, the stages for creating user communities from Z39.50 server logs are the same as those of any other data-mining task: data collection and preprocessing, pattern discovery, and knowledge postprocessing.

Data Collection and Preprocessing

Preprocessing aims to enable data to be used as input to the next stage of pattern discovery. The Z39.50 protocol organizes the user queries into sessions containing one or more queries. For each log-file record that refers to searches, the following data is kept: session code, date and time of the query, origin IP address, code of the collection accessed, and the search terms of the access under consideration.

The training dataset—a table for which each column (feature) corresponds to a search term and each row (objects) corresponds to a session—is prepared. Each row is a binary vector and each vector cell indicates the existence of a search term in a session.

Pattern Discovery

Given the training data in the appropriate form, interesting patterns are extracted with the use of machine learning techniques, such as clustering, classification, and association rule discovery. The technique chosen for this study was the use of unsupervised learning through clustering, due to the requirement for the formulation of user groups with no prior decisions. Two simple and efficient graph theory-based clustering algorithms are proposed that search for either all the cliques or all the connected components of a graph that represents the sessions' characteristic features—in other words, search terms.

The process starts by constructing a graph $G(A,E)$. The set of vertices A corresponds to the search terms used in the user queries. The set of edges E corresponds to the co-occurrence of search terms as observed in the user queries. For instance, in the NDC digital library, if a session contains the search terms "Museums-Administration" and "Archaeology," an edge is added between the relevant vertices. The weights on the vertices and the edges are computed as the frequencies of the users' choices and their combinations respectively. Edge frequencies are normalized by dividing them with the maximum of the frequencies of the two vertices that they connect. The effect of the normalization is to remove the bias for characteristics that appear very often in all users.

The connectivity of the resulting graph G is usually high. For this reason we make use of a *connectivity threshold*, attempting to reduce the number of edges in the graph. The connectivity threshold represents the minimum weight allowed for the edge existence. When this threshold is high, the graph will be sparse and the derived clusters will have small size and high quality. When the threshold is lower the derived clusters will be larger.

After the edge reduction, the clique-mining method accepts as clusters, or user communities, all the existing cliques: all the subgraphs of the graph G in which every pair of nodes has an edge between them. Despite the large complexity of the clique-finding problem, the implemented algorithm is very fast.[6]

Alternatively, the connected components mining method could be used. This parses the graph G and searches for all the subgraphs in which every pair of nodes is connected by a path of edges between them. The derived components represent disjoint groups of search terms and each of them corresponds to a user community.

The main advantages for using the connected components mining method are mainly its efficiency (the algorithm complexity is almost linear on the number of sessions) as well as its ability to discover patterns in very sparse data sets.[7] As will be shown in the Experimental Results section, a sparse graph that contains a number of connected components may be too sparse to contain any sizeable cliques. However, the clique-mining method provides more cohesive and coherent clusters. One more difference between the two clustering methods is that clique mining produces overlapping clusters, while the connected components are, by default, disjoint clusters.

Pattern Postprocessing and Evaluation

There are several issues concerning the expressiveness of the clusters produced by the proposed algorithms, as well as the recommendation of the appropriate subject headings to the digital library users. The main questions that have to be answered are:

- Which clusters actually constitute user communities?
- Which subject terms provide meaningful descriptions of the user communities (that is, community models) and could be recommended in a query expansion process?

In the case of clique mining, the generated clusters are evaluated by varying the connectivity threshold, and measuring the following two properties:

- *Coverage*—the proportion of search terms participating in the clustering, since due to the connectivity threshold not all the search terms are members of the generated clusters.
- *Overlap*—the extent of overlap between the clusters. This is the average value of the size of the intersection of any pair of clusters divided by the size of their union (Jaccard co-efficient).

In the case of discovering connected components, there is no overlap between them, so the criterion for selecting clusters as meaningful user communities is the coverage of the search terms. Moreover, the components that are exactly the same with user sessions are dropped, as they do not form user communities but simply user sessions.

The terms describing each user community come from frequently used search terms that exist in the digital library authority files. These terms characterize each community; that is to say, they constitute the *community model*. The terms that make up a community model could be recommended to the users as alternative or additional search terms in a query expansion mechanism.

Description of the Usage Data

The NDC digital library hosts many collections covering various scientific domains. All the collections are structured using the UNIMARC format with almost the same level of quality attributes: specificity, completeness of fields, syntactic correctness, but a different level of the quality consistency attribute as described in every category. From their 300,000 metadata records, there are links giving online access to fourteen thousand digitized documents composed of two million scanned pages and few other object formats.

The Web-based retrieval system that was monitored implements a Z39.50 client and connects to a Z39.50 server. The users start their sessions by selecting and connecting to a collection. After connecting to a collection, a user may express a search request or browse specific access points and then retrieve the documents.

For this experiment, the following two categories of digital library collections were chosen, including two collections each, based on the completeness of their metadata descriptions as well as the different ways that the users access them (table 1 presents the access point usage in these NDC collections):

- Hellenic Archaeological Records—grARGOS (C1) and International Archaeological Records—intARGOS (C2). They consist of archaeological records, including library material with diverse types of data and a good consistency level, targeted to a specific scientific user group (such as researchers on archaeology) and have the largest authority files among all collections. However only 5 percent of the users pose queries through the subject heading access point, while more than 30 percent use the author access point. This is due to the nature of the collection and the high degree of specialization of the users, and because the authors of the works included in these collections are known by their direct relation with their subjects, places, and findings—often coming from excavations. For example, Evans is directly related with the palace of King Minos at Knossos, Crete. This explains how the author access point has in fact a thematic character in ARGOS user sessions.
- Hellenic School Libraries (C3) and Hellenic Public Libraries Union Catalog (C4). They are union catalogs for library material from many domains, consisting of general-purpose records, a very good consistency level, are targeted to librarians, and have the smallest authority files. The users access them using all the main access points uniformly. The high percentage of the usage of subject heading access point is explained because the target group of these collections are librarians—a homogeneous user group experienced in searching and retrieving information.

The operations performed by the users on the content of the two selected collection categories were logged for a period of thirty-two months (September 2000 to April 2003), keeping the sessions that used the authors and subject heading access points for the two collection categories. Table 2 shows the numbers of the logged sessions, queries, and search terms per collection, during the mentioned period.

Experimental Results

In the following paragraph, the results from the application of the clique-mining method on the datasets concerning the author access point are presented, together with the results from the application of the connected

Table 1. Percentage of Access Point Usage per Collection

Access Point	Total	C1	C2	C3	C4
Any	50.5	39.0	30.0	30.0	35.0
Author	19.3	28.0	36.0	23.0	26.0
Title	16.0	26.0	26.0	21.0	18.0
Subject Heading	9.5	4.0	5.0	22.0	15.0

components mining method on the datasets concerning the subject heading access point.

Author Access Point Analysis—Cliques

The processing of the data sets from both collection categories concerning the author access point resulted in cliques for various values of the connectivity threshold. Depending on the value of the connectivity threshold, the coverage of the clusters and the overlap varied. Table 3 presents the results along those two dimensions. The column named Number of Cliques indicates the amount of cliques that correspond to a particular connectivity threshold. As expected, the coverage for small threshold values is large due to the large number of cliques. A similar behavior follows the overlap. For both collections around the threshold value 0.5, about 10 percent of the search terms appear in the cliques (coverage equals to 0.12 for ARGOS and 0.14 for public school libraries), while there is little overlap between the cliques (for ARGOS equals to 0.0009 and for public school libraries equals to 0.0007). This observation suggests the selection of this threshold value for the formation of the desired representative vocabulary. At this connectivity threshold value, sixty-nine clusters are generated for the ARGOS collection category and seventy-four for public school libraries.

Table 4 presents an indicative example of two user community models per each collection category.

Subject Heading Access Point Analysis—Connected Components

The clique-mining algorithm has been applied on the datasets produced by the logs for the two access points and for each collection category. However, for both collection categories, the graph generated from the subject heading access point dataset was too sparse. Even with no connectivity threshold, the algorithm could not produce any sizeable cliques. Therefore, a different method was used for the subject heading access point dataset, the connected-components-mining-method.

The component-mining algorithm has been applied on the two datasets that refer to the subject heading access point, producing connected components of variant sizes as shown in table 5. The column named Component Size indicates the number of the included search terms in each component, while the column named Component Number corresponds to the amount of the generated connected components of a particular size.

Table 2. Logged Sessions, Queries, and Search Terms per Collection and Access Point

	ARGOS		Public School Libraries	
	Authors	Subject Heading	Authors	Subject Heading
Sessions	8,172	1,141	4,965	2,148
Queries	15,069	1,742	10,566	5,622
Search Terms	10,020	1,434	6,114	2,336

Table 3. The Results of the Clique-Mining Algorithm on Author Access Point

Connectivity Threshold	ARGOS			Public School Libraries		
	No. of Cliques	Coverage	Overlap	No. of Cliques	Coverage	Overlap
0.3	407	0.64	0.0010	454	0.69	0.00100
0.4	266	0.44	0.0010	289	0.46	0.00100
0.5	69	0.12	0.0009	74	0.14	0.0007
0.7	41	0.08	0.0004	30	0.06	0.0000

Table 4. A Sample of User Community Models on Author Access Point

Community	ARGOS	Public School Libraries
Connectivity threshold	0.5	0.5
1	Kunz, Hagu, Orszag, Kubcza	Tomboulides, Reeve, Slott, Pinar
2	Brancacci, Dubella, Cinuzz, Marciol, Adrian, Valde	Harrison, Hutchinson, Morgan, Aitke, Littlewod, Rinvolucre, Bygat

A significant observation on the datasets is that most sessions include just one query in both collections. Therefore, most derived components are of small size. In particular, the components having more than eight search terms are unique in both collections. Furthermore, an interesting phenomenon in both case studies is the existence of one giant component. For ARGOS, the giant component consisted of ninety-three search terms, while in the public school libraries the giant component consisted of 967 search terms. These components do not seem to be homogeneous and meaningful. They are the union of sessions consisting of search terms with high appearance frequencies, but they also include almost arbitrary terms with low occurrence frequencies, present in the same sessions. However, it is important to identify these special clusters and not treat them as normal cohesive communities.

The search terms coverage decreases as the size of the components increases. This behavior is to be expected because there are very few large components that include only a small part of the search terms. Figure 1 shows the coverage decrease with respect to the component size. The public school libraries curve is less steep than the ARGOS curve due to the usage of the subject heading access point. As mentioned above, public school library users are librarians and search the collections using the subject heading access point more than average users do. So, the derived components complement each other and include more than 40 percent of the search terms. On the other side, ARGOS users are archaeologists and prefer the Author access point. Therefore, ARGOS-user sessions consist of few queries with very different search terms.

According to figure 1, around the component size with value 3 (coverage value 28.5 percent) the ARGOS coverage curve becomes less steep, while the public school libraries coverage curve becomes less steep around the component size of value 3 (coverage value 45.9 percent). This observation suggests the selection of this size value for the formation of the desired user-community models. Hence, the analysis leads to the selection of components with size greater than or equal to three search terms for the ARGOS collection, and more than four search terms for the public school libraries collection. Table 6 presents an indicative example of two user-community models per each collection. The derived community models are disjoint sets of subject headings and constitute specific vocabularies that describe the communities. These subject headings could be recommended and used by the digital library in a query expansion process.

Conclusions

Efficient and effective access to online information becomes increasingly critical as the amount of services for

Table 5. Components Size and Number per Collection on Subject Heading Access Point

Component Size	Component Number ARGOS	Public School Libraries
1	50	56
2	21	31
3	11	9
4	13	5
5	6	4
6	3	2
7	2	2
8	1	
9	1	
14	1	
16	1	
22	1	
93	1	
960		1

Table 6. A Sample of User Community Models on Subject Heading Access Point

Community	ARGOS	Public School Libraries
Component Size	3	4
Number of Communities	41	14
1	Peloponnese, Medieval Peloponnese, Medieval Greece, Paleologos	Special education, Disabilities, Handicap, Children Disabilities, Higher Education, Education history
2	Asia Minor, Smyrna, Balkan Peninsula, Asia Minor -- Economic conditions, Asia Minor -- history, Smyrna disaster 1922, Smyrna -- Economic conditions, Smyrna Fire 1922	Literature, Greek literature, Classical literature, Classical literature -- critics, Classical philology European literature

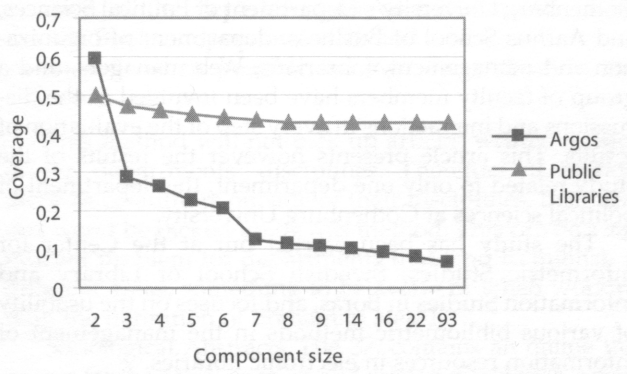


Figure 1. Coverage Curves on Subject Heading Access Point

delivering information over the Internet is continuously increasing. The objective of this study was to analyze the behavior of the users of a digital library and to provide useful information to its administrator, in order to improve the provided services. The approach that was adopted was to construct user communities corresponding to groups of users with similar behavior. The end result was a behavioral pattern for each community, which provides much richer information to the administrator than the commonly used statistical figures about the usage of a service.

One of the important issues to arise in this work was the necessary engineering of the data collected. Query-based information-retrieval services need various language-engineering tools and domain-specific classification hierarchies in order to reduce the dimensionality of the problem and generate the final data set.

Another important issue is the choice of the clustering method that will construct the communities. Many graph theoretic methods have been used for the identification of user communities.[8] The advantages of the approach described in this paper are its simplicity and its computational efficiency. Component-mining method generates disjoint communities that map each user uniquely to a community. This might be desirable in services where each user community needs to be treated separately and data are sparse. However, it will not be desirable in many services where users naturally belong to more than one community. In such cases the clique-mining algorithm is more suitable. Further research to evaluate other clustering algorithms is also worthwhile.

In conclusion, it can be said that the use of clustering methods in the discovery of behavioral patterns for user communities is feasible and can provide very valuable information to the administrator of a digital library.

References

1. C. Basu, H. Hirsh, and W. Cohen, "Recommendation As Classification: Using Social and Content-Based Information in Recommendation," in *Proceedings of the 15th National Conference in Artificial Intelligence* (Menlo Park, Calif.: American Assoc. for Artificial Intelligence, 1998), 714–20; D. Billsus and M. Pazzani, "User Modeling for Adaptive News Access," *User Modeling and User-Adapted Interaction* 10, no. 2–3 (2000): 147–80; C. R. Anderson and E. Horvitz, "Web Montage: A Dynamic Personalized Start Page," in *Proceedings of the 11th Intl. World Wide Web Conference* (New York: ACM Pr., 2002).

2. A. Peterson Bishop, "Working toward an Understanding of Digital Library Use: A Report on the User Research Efforts of the NSF/ARPA/NASA DLI Projects," *D-Lib Magazine* 4, no. 1 (Oct. 1995); S. D. Payette and O. Y. Rieger, "Z39.50 The User's Perspective," *D-Lib Magazine* (Apr. 1997); F. Esposito et al., "Adding Machine Learning and Knowledge Intensive Techniques to a Digital Library Service," *International Journal on Digital Libraries* 2, no. 1 (1998): 3–19; N. A. Van House et al., "User Centered Iterative Design for Digital Libraries: The Cypress Experience," *D-Lib Magazine* (Feb. 1996); M. Sfakakis and S. Kapidakis, "User Behavior Tendencies on Data Collections in a Digital Library," in *Research and Advanced Technology for Digital Libraries: 6th European Conference on Digital Libraries* (ECDL 2002), Rome, Italy, Sept. 2002: Proceedings, Lecture Notes on Computer Science, vol. 2458 (Heidelberg, Germany, Springer-Verlag, 2002), 550–59.

3. S. Chakrabarti, M. van der Berg, and B. Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," in *Proceedings of the 8th Intl. World Wide Web Conference*, (Toronto: Elsevier, 1999), 545–62.

4. J. Orwant, "Heterogeneous Learning in the Doppelgänger User Modeling System," *User Modeling and User-Adapted Interaction* 4 (1995): 107–30.

5. J. S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence* (San Francisco: Morgan Kaufman, 1998), 43–52; D. Pennock et al., "Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach," in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence* (San Francisco: Morgan Kaufman, 2000), 474–80.

6. G. Bron and J. Kerbosch, "Finding All Cliques of an Undirected Graph," *Communications of the ACM* 16, no. 9 (1973): 575–77.

7. M. Perkowitz and O. Etzioni, "Adaptive Web Sites: Automatically Synthesizing Web Pages," in *Proceedings of the 15th National Conference in Artificial Intelligence* (Menlo Park, Calif.: American Assoc. for Artificial Intelligence, 1998), 727–32.

8. N. Imafuji and M. Kitsuregawa, "Effects of Maximum Flow Algorithm on Identifying Web Community," in *4th ACM CIKM International Workshop on Web Information and Data Management* (New York: ACM Pr., 2002): 43–48; G. Paliouras et al., "Discovering User Communities on the Internet Using Unsupervised Machine Learning Techniques," *Interacting with Computers Journal* 14, no. 6 (2002): 761–91.

Matching Subject Portals with the Research Environment

Irene Wormell

This article presents methods for testing the usefulness of bibliometric methods for the evaluation of information resources located at subject portals. Two subject portals for social sciences have been selected as objects for the study: SamWebb at Gothenburg University Library in Sweden and Bisigate at the Aarhus Business School Library, Denmark. To show how to capture the local users´ views and requirements in the development of portals, this article explores the results of the analyses targeting one of the selected institutions, Gothenburg University's Department of Political Sciences. The study produced various types of lists as well as maps for monitoring the research and publication pattern of the department. These reports allow exploration and visualization of the research results of the institution in a form that is easy to read and understand for portal users. The content of the lists and maps was designed to provide information about which journals are relevant for the ongoing research activities in the department, and to identify useful links to professional institutions, organizations, persons, most cited publications, and authors. The study gathered quantitative data to measure how well the information resources of the portals match the research profile of the institutions.

Within the frame of the current digital libraries program in the Nordic countries, several subject portals have been set up for the purpose of making the collective information resources of digital materials available to users on a nationwide basis. Besides providing access, these portals present new services in libraries and the reallocation of human resources.

This study was designed to test the usefulness of bibliometric methods for the evaluation of information resources located at the subject portals, and for raising interest in portals in the local research community. SamWebb at Gothenburg University Library in Sweden and Bisigate at the Aarhus Business School Library, Denmark, both subject portals for social sciences, were selected for the study.

To capture the local users´ views and requirements of the portals, the analyses targeted two selected institutions: Gothenburg University's Department of Political Sciences, and Aarhus School of Business, department of organization and management. Librarians, Web managers, and a group of faculty members have been involved in the discussions and interactions in every step of the evaluation of results. This article presents however the results of the study related to only one department, the department of political sciences at Gothenburg University.

The study has been carried out at the Center for Informetric Studies, Swedish School of Library and Information Studies in Borås, and focuses on the usability of various bibliometric methods in the management of information resources in electronic libraries.[1]

Method

A bibliometric approach was chosen to collect evidence leading to journals, publications, researchers, and institutions of potential interest for those using portals.

The bibliometric analyses were based mainly on the internal information received from the selected departments in electronic and printed form: these consisted of lists of publications and annual reports with various categories of publications listed for 1999 to 2000. Starting with these lists, relevant authors and publications have been identified using international bibliographic databases. The citation database, Social Science Citation Index (SSCI), File 7 in Dialog, provided the source articles and the citation analyses. To trace the publications of the departments—also outside citation databases—the analyses were extended to other core databases within the field: Sociological Abstracts, File 37, and Gale Group Business ARTS (Applied Research, Theory, and Scholarship), File 88, a scholarly business database designed to help identify current research, analysis, trends, and expert opinions in such disciplines as management, science, humanities, and social sciences. With complementary searches in these two different types of databases, it is assumed that the analysis covers a significant portion of all of the international publications of the departments, including monographs, case studies, conference papers, and dissertations.

In the citation databases, the authors´ affiliation is identified by the field code CS (corporate source). In other bibliographic databases, however, this data is not accessible. Therefore, in these databases the department publications had to be identified by other means. Documents for the analysis have been selected in the following way:

- step 1: use subject headings, descriptors, and identifiers to build a set of relevant documents within the whole domain;
- step 2: this large set of documents, in combination with the authors´ name, will comprise the final selec-

Irene Wormell (Irene.wormell@hb.se) is a Professor at the Swedish School of Library and Information Science, Borås, Sweden.

tion of the department-relevant documents (using the Boolean operator AND); and
- step 3: when possible, arrange the authors in thematic groups.

This method will not pick up articles written outside of the departments' disciplines. Therefore, if authors serve as co-authors outside of their disciplines, their articles won't be picked up by this method. This can be a significant problem for departments that tend to publish in different areas, something to be aware of when working with the definition of domain in step 1.

Sociological Abstracts, File 37, contains an online thesaurus for the social sciences and related disciplines, a standard search and indexing tool in the field. To identify the relevant international key words and index terms for the research area of each department, this thesaurus was used to see which related (R) and broader (B) terms are given in the database for the subject term (S) Political Science.

The study produced various types of lists and maps for monitoring the research and publication pattern of the departments. The contents of these were designed to provide the following information:

- which journals are (the most) relevant for the ongoing research activities of the department; and
- useful links to professional institutions, organizations, persons, and most-cited publications and authors.

Comparative Citation Analyses

To see how the research and publication pattern at the department has developed during the last twenty-seven years, a series of parallel analyses of two periods, 1974 to 1989 and 1990 to 2001 was undertaken. This made it possible to compare how the publication pattern in the department has changed between the two periods.

Changes in the rank order between the two time periods indicate changes not only in the research direction or research staff, but also in the publication pattern. Ranked lists by subject were developed for the two time periods, involving authors (AU), journals (JN), subject codes indicating the topical area of the journal (SC), cited authors (CA), cited works (CW), and cited references (CR). These lists can illustrate trends and developments in the research history of the department (see figure 1 for an example).

The rank order of the journals in which the researchers are publishing and citing is probably the most useful information to be gained for the evaluation of subject portals, because it indicates the topicality of the current research in the departments in the time period 1999 to 2001. The analysis includes both citation databases and other main international databases. Merging the tops of the ranked lists of publishing and citing journals, it is possible to compile a list of candidate journals matching the (current) research of the departments (see figure 2).

In the bibliometric literature there are several studies and methods reported for the identification of the most

Top Journal Titles for 1974–1992

Rank	No. Items	Title
1	17	Electoral Studies
2	3	Acta Sociologica
3	2	Quality and Quantity
4	2	Sociologisk Forskning
5	1	American Econ. Review
6	1	British J. of Political Sc.
7	1	Europ J. of Polit. Research
8	1	Human Relations
9	1	Internasjonal Politikk
10	1	J. of Palestine Studies
11	1	Polity
12	1	Public Opinion Quarterly
13	1	Sage Professional Paper in Contemp. Politics
14	1	Social Psychol. Quarterly
15	1	Urban Affairs Quarterly

Top Journal Titles for 1993–2001

Rank	No. Items	Title
1	7	Scandinavian Polit. Studies
2	5	Europ J. of Polit. Research
3	3	Electoral Studies
4	2	Europ J. of Communication
5	2	Governance—An Internat. J. of Policy
6	2	Internat. J. of Public Opinion Resear
7	2	J. of Theoretical Politics
8	1	Annals of yhe American Academy of Political An.
9	1	British J. of Polit. Science
10	1	Gender anf Society
11	1	Geoforum
12	1	Housing Studies
13	1	Internat. Polit. Sc. Review
14	1	Javnost—The Public
15	1	J. of Peace Research
16	1	Party Politics
17	1	Policy and Politics
18	1	Politics and the Life Sciences
19	1	Rationality and Society
20	1	Scandinavian Housing and Planning Research

Figure 1. Ranked lists presenting journal names in rank order where researchers published articles between 1974–1992 and 1993–2001

relevant set of journals for a library, special collection, research area, or a business unit. They use combinations of advanced citation and co-citation analyses with qualitative analyses and local strategic discussions to capture both the local and global aspects of a serial collection.[2] In this study, the publication patterns of the departments were used as bases for the identification of the most relevant journals for their research activities. Because of the small data set (number of international publications in Gothenburg: seventy-one articles) a more extensive analysis of the citations was not feasible in this project. However, the final report presents in detail citation and co-citation analyses where relevant data are captured for the local and global aspects of a serial collection.[3]

Electoral Studies
- Scandinavian Political Studies
- European J. of Political Research
- Acta Sociologica
- Housing Studies
- International J. of Public Opinion Research
- The Public Opinion Quarterly
- Tidsskrift For Samfunnsforskning
- American J. of Sociology
- J. of Politics

Politics And Society
- Urban Studies
- British J. of Political Science
- European J. of Communication
- Governance—An International J. of Policy

J. of Theoretical Politics

Quality and Quantity
- Sociologisk Forskning
- Annals of the American Academy of Political An.

Gender and Society

Geoforum

Human Relations

Internasjonal Politikk
- International Political Science Review
- Javnost—The Public
- J. of Palestine Studies
- J. of Peace Research
- Party Politics
- Policy and Politics
- Politics and the Life Sciences
- Polity
- Public Opinion Quarterly
- Rationality and Society

Sage Professional Paper in Contemporary Politics

Scandinavian Housing and Planning Research
- Social Psychology Quarterly
- Urban Affairs Quarterly

Figure 2. Top of the list of candidate journals for the Gothenburg University Department of Political Sciences based on the publication pattern of the department from 1999 to 2001

Other Results

As a spin-off effect of the bibliometric analyses, the research and publication pattern of each department was mapped. Because the analyses were carried out in two time periods, 1974 to 1992 and 1993 to 2001, they facilitated the study of trends and developments in the research history of the departments.

Figure 3 shows the publication analyses of the Gothenburg University Department of Political Sciences.

As a means of verifying the methods, the names were checked with the department's list of publications to see how this list matched what they reported.

Figure 4 cites the names of those journals where the articles (seventy-one) were published between 1974–2001. The thirty-two listed journals in rank order indicate the relevance of each journal for the department: their research, reported in the citation databases, appearing mainly in the form of articles published in the international core journals of the field.

All source articles related to the names of the researchers listed in the Publication List 1999–2000 have been searched and analyzed in the cited authors (CA) index (see figure 3). The following lists in figure 5 show the top of the ranked lists for the two time periods: cited authors with the number of documents where they occur in the list of references. This is a representation of the researchers and their peers within the research area of the department, based on the citation frequency analysis.

Only articles in which at least two of the department's researchers were referred to (co-cited) have been included in the co-citation analysis. A total of one hundred articles fulfilled this requirement in the citation database SSCI.

In figure 6, the co-citation map shows the clusters of researchers with citation links between them. (The names of Gothenburg´s researchers are all in capitals.) The map is based on a co-citation frequency analysis measuring the relationship between documents (published by Gothenburg's researchers) and calculated on the numbers of co-citations between pairs of authors. The size of the dots (author names) indicate citation frequency, and the lines point to author pairs. For the production of graphs, the co-citation frequency analysis has been processed in Bibexcel, a special software for bibliometric analysis, developed by Professor Olle Persson at Umeå University, INFORSK, Sweden. It can be downloaded for free for academic purposes from www.umu.se/inforsk.

Assisted by the department´s list of publications and assignments, the research at the department was grouped into three thematic areas. In order to allow visualization of the relationships between groups, colors have been used to indicate group membership.

- The main research area (group 1): electoral studies, political behavior (gray dots)
- Historically core research area (group 2): public policy, public administration (black dots)
- New research area (group 3): comparative politics, international relations (white dots)

This grouping of the research highlights the analysis of the new research directions of the department while at the same time focuses on emerging new disciplines. In group 2, for example, it illustrates how new disciplinary areas such

Rank	No. Items	Term
1	16	Sarlvik, B.
2	9	Granberg, D.
3	7	*Holmberg, S.
4	7	*Pierre, J.
5	6	Widfeldt, A.
6	4	*Gilljam, M.
7	3	*Esaiasson, P.
8	2	*Bjereld, U.
9	2	*Demker, M.
10	2	*Hinnfors, J.
11	2	*Johansson, F.
12	2	*Lundqvist, L.
13	2	*Rothstein, B.
14	2	Szucs, S.
15	1	Anderson, E.
16	1	*Bjereld, U.
17	1	Brantgarde, L.
18	1	Dahlkvist, M.
19	1	Hadenius, S.
20	1	Hakansson, N.
21	1	Hibbs, Da.
22	1	Hoyer, S.
23	1	Jahn, D.
24	1	Janson, S.
25	1	Jonasdottir, A.
26	1	Jonasson, B.
27	1	Jonson, L.
28	1	Kumlin, S.
29	1	*Lundqvist, L.
30	1	Magnusson, A.
31	1	Moring, T.
32	1	*Oscarsson, H.
33	1	Park ,S.
34	1	*Persson, S.
35	1	*Peters, Bg.
36	1	Petersson, O.
37	1	*Rothstein, N.
38	1	Vegelius, J.
39	1	*Wangnerud, L.
40	1	Weibull, L.
41	1	Westerstahl, J.
42	1	Wise, L.

* Names are from the list of publications 1999–2000.

Figure 3. Ranked list of author names published in the international core journals in affiliation with the Gothenburg University Department of Political Sciences from 1974 to 2001

Rank	No. Items	Term
1	20	Electoral Studies
2	7	Scandinavian Political Studies
3	6	European J. of Political Research
4	3	Acta Sociologica
5	2	British J. of Political Science
6	2	European J. of Communication
7	2	Governance—An Int'l J. of Policy
8	2	International J. of Public Opinion Research
9	2	J. of Theoretical Politics
10	2	Quality and Quantity
11	2	Sociologisk Forskning
12	1	American Economic Review
13	1	Annals of the American Academy of Political An.
14	1	Gender and Society
15	1	Geoforum
16	1	Housing Studies
17	1	Human Relations
18	1	Internasjonal Politikk
19	1	International Political Science Review
20	1	Javnost—The Public
21	1	J. of Palestine Studies
22	1	J. of Peace Research
23	1	Party Politics
24	1	Policy and Politics
25	1	Politics and the Life Sciences
26	1	Polity
27	1	Public Opinion Quarterly
28	1	Rationality and Society
29	1	Sage Professional Paper in Contemporary Politics
30	1	Scandinavian Housing and Planning Research
31	1	Social Psychology Quarterly
32	1	Urban Affairs Quarterly

Figure 4. Ranked list of journals publishing articles by the researchers of the Gothenburg University Department of Political Sciences

as politics in Sweden, welfare state, and corporatism (light gray dots) are linked to the historically core research areas of public policy and public administration (black dots).

The co-citation map in figure 7 illustrates internal relations within the department: citation links between the researchers of the department. The map is based on a co-citation frequency list (with a range of 33–1). The thickness of lines indicates strong professional relations: researchers are active within the same topical area and they are co-cited in publications several times. The size of dots has no significance here. The gray dots represent an emerging new research area at the department, gender studies, which is related to the activities of the new generation of researcher. It is still a new co-citation cluster, but clearly indicated on the map.

Please note that the two unconnected names in figure 7 have no co-citations pairs, but they are included in the map for the purpose of showing the names of all researchers active within the field of gender studies at the department.

The connection between the source journals could be traced from the authors' co-citation pattern. It resulted in a map (figure 8) that shows the clusters of journals cited by the researchers of the department, a means of identifying the central journals used in the research activities of the department. The size of the dots signifies citation frequency of articles published by the department. Distance between the dots indicates the degree of co-citation between the journals.

1974–1992			1993–2001		
No. Rank	**Items**	**Term**	**No. Rank**	**Items**	**Title**
1	6	Granberg, D.	1	13	Gilljam, M.
2	5	Holmberg, S.	2	12	Holmberg, S.
3	5	Lazarsfeld, P.	3	9	Downs, A.
4	4	Campbell, A.	4	9	Esaiasson, P.
5	3	Berelson, B.	5	6	Asp, K.
6	3	Gilljam, M.	6	6	Bennulf, M.
7	3	Kelley, S.	7	6	Converse, P.
8	3	Wattenberg, M.	8	6	Granberg, D.
9	2	Asp, K.	9	6	Petersson, O.
10	2	Berelson, B.	10	6	Sarlvik, B.
11	2	Budge, I.	11	5	Rothstein, B.
12	2	Butler, D.	12	4	Lewin, L.
13	2	Clausen, A.	13	4	Listhaug, O.
14	2	Converse, P.	14	4	Oskarson, M.
15	2	Crewe, I.	15	4	Putnam, R.
16	2	Esaiasson, P.	16	4	Sjoblom, G.
17	2	Key, V.	17	4	Westerstahl, J.
18	2	Sarlvik, B.	18	3	Aberbach, J.
19	2	Vandereijk ,C.	19	3	Bjereld, U.
20	2	Wattenberg, M.	20	3	Conover, P.

Figure 5. Ranked list of the top twenty cited authors in the department's publications

Findings and Application

These bibliometric methods, based on a publication pattern analysis of the local user environment, will improve the match between the resources of the subject portals of university libraries and the current research activities of the departments. The goal is to help information resource managers map the emerging new subject developments and new research areas implied in each department´s dynamic scientific and social life. Portal managers and librarians, in spite of their ambition, may not have easy access to this kind of information through their traditional

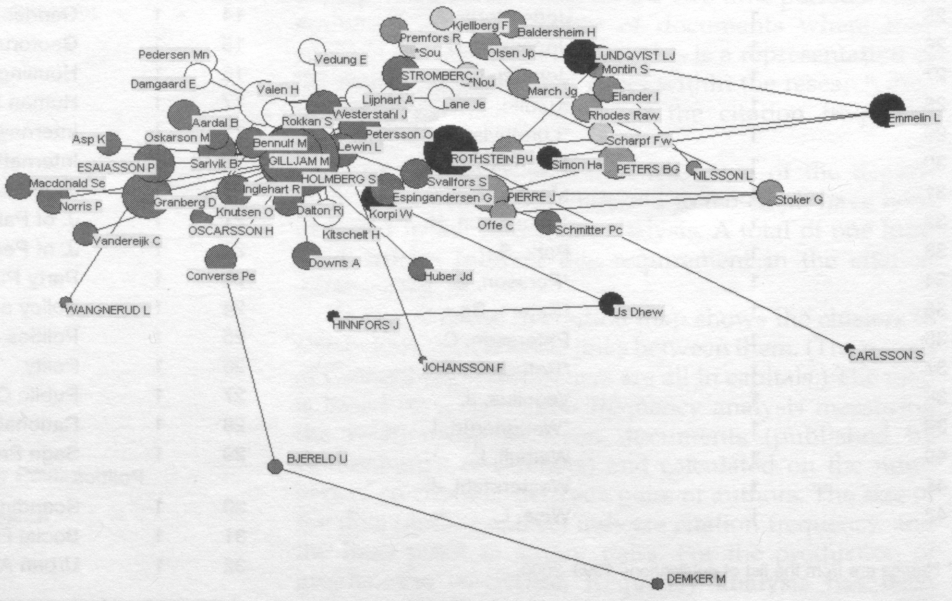


Figure 6. Co-citation map with names of researchers from Gothenburg (in capital letters) and researchers at other institutions with whom they are co-cited. Colors refer to different thematic clusters to which the authors belong (source: Social SciSearch, 2001)

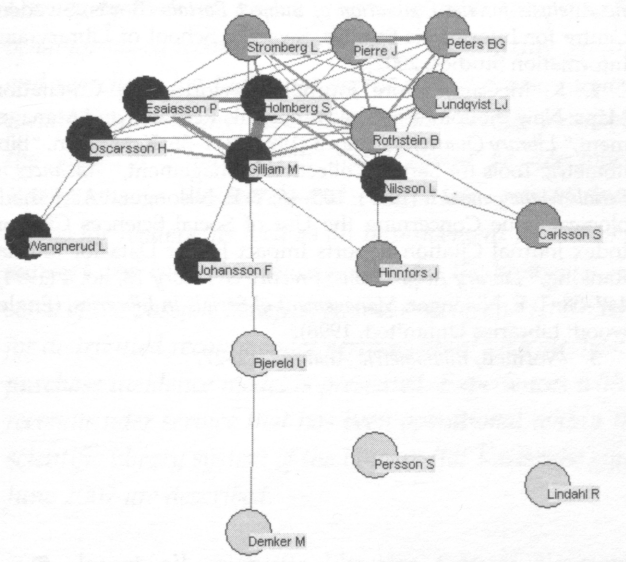


Figure 7. Internal relationships among researchers at the Gothenburg University Department of Political Sciences based on co-citation analysis (source: Social SciSearch, 2001)

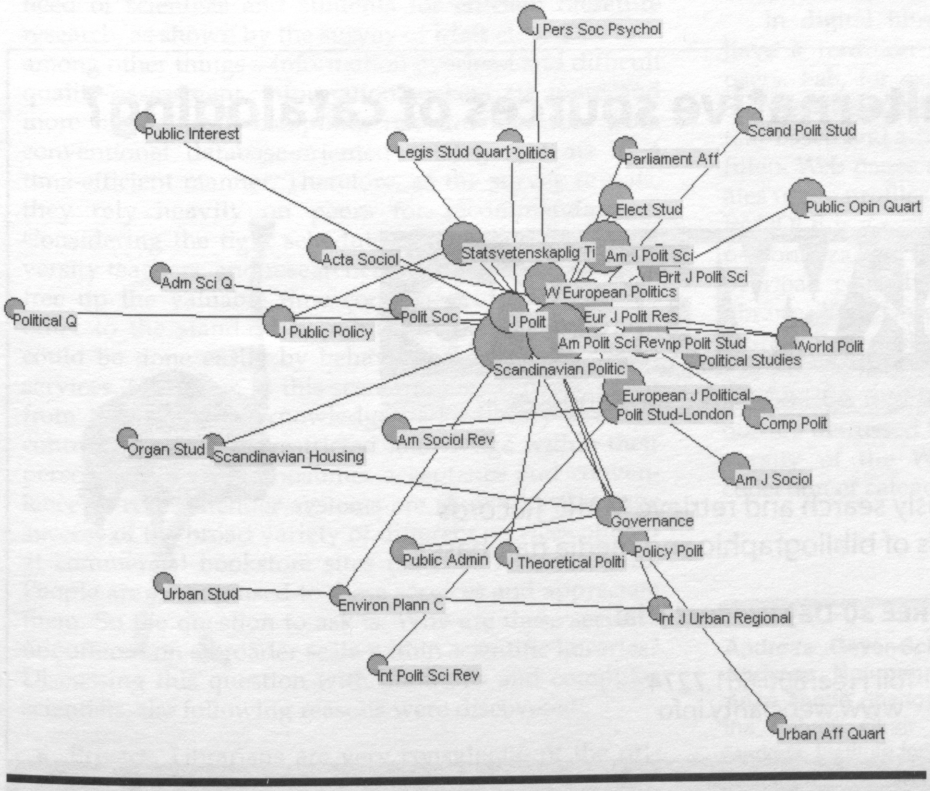


Figure 8. The cluster of journals cited in source documents: publications of the Gothenburg University Department of Political Sciences (source: Social SciSearch, 2001)

communication channels. Knowledge about new projects, and the publication pattern for dissemination of their results to the international scientific community, is essential for the utilization and the effective management of subject portals. Therefore, the purpose of this pilot project was to generate ideas and extend the vision for subject portals.

Gothenburg University (like other universities) regards the portal as a resource platform in the new learning environment, where librarians create tools to assist students in the identification and use of relevant research resources. The methods presented here for knowledge production in one selected department can function as guidance and support for the researchers of that department in discovering their own subject area, and in getting acquainted with the information resources relevant to the local and global perspectives in their study or research.

The presented analyses are based on small data sets, related to one department at the university. It would also be relevant to gain similar bibliometric information about larger units of the academic world, such as all departments (within the proposed field of study) in the country or region. In this case, the bibliometric analyses would be based on much larger data sets, and would also allow exploration of the national and regional perspectives in publication patterns and networking in the given subject area.

One should consider the use of this type of information, and the necessity of creating navigation instruments for the users of portals at various levels of academic activities.

The next step in the development of portals should be to provide education and training for the library staff in bibliometric analysis, and in the use of modern presentation and visualization techniques. Having these skills, they could collaborate with teachers and researchers in the development of useful facilities for the portals, thus promoting the integrated use of bibliographic information in the learning processes.

The project served to generate interest for the portals and to explore its potential for use among both faculty members and students. The study and discussions were useful also for enlarging the horizons of the

librarians and their services: traditional custodian roles have to be transferred to more analytical work, combined with skills in presentation and visualization of information.

References

1. Irene Wormell, "Informetrics—A New Area of Quantitative Studies," *Education for Information* 18 (2000): 131–38; Irene Wormell, "Databases As Analytical Tools," *Encyclopedia of Library and Information Science*, suppl. 33, s.v. databases, (New York: Marcel Dekker Inc., 2001), 77–92; Irene Wormell, *Bibliometric Analysis for the Evaluation of Subject Portals* (Borås, Sweden: Centre for Informetric Studies, Swedish School of Library and Information Studies, 2002).

2. K. McCain, "Core Journal Networks and Co-citation Maps: New Bibliometric Tools for Serials Research and Management," *Library Quarterly* 61 no. 3 (1991): 311–36; K. McCain, "Bibliometric Tools for Serials Collection Management," *Advances in Serials Management* 6 (1997): 105–46; T. E. Nisonger, "A Methodological Issue Concerning the Use of Social Sciences Citation Index Journal Citation Reports Impact Factor Data for Journal Ranking," *Library Acquisitions: Practice & Theory* 18, no. 4 (1994): 447–58; T. E. Nisonger, *Management of Serials in Libraries*, (Englewood: Libraries Unlimited, 1998).

3. Wormell, *Bibliometric Analysis*, 12–27.

Looking for alternative sources of cataloging?

BookWhere™

The Definitive Z39.50 Client for Windows



Simultaneously search and retrieve MARC Records
from thousands of bibliographic and media databases.

FREE 30-Day Product Trial

Toll Free: 800.661.7274
www.webclarity.info

WebClarity
SOFTWARE INC.

Looking Beyond . . .

An Architecture for Behavior-Based Library Recommender Systems

Andreas Geyer-Schulz, Andreas Neumann, and Anke Thede

Library systems are a very promising application area for behavior-based recommender services. By utilizing lending and searching log files from online public access catalogs through data mining, customer-oriented service portals in the style of Amazon.com could easily be developed. Reductions in the search and evaluation costs of documents for readers, as well as an improvement in customer support and collection management for the librarians, are some of the possible benefits. In this article, an architecture for distributed recommender services based on a stochastic purchase incidence model is presented. Experiences with a recommender service that has been operational within the scientific library system of the Universität Karlsruhe since June 2002 are described.

Almost all scientific libraries feature electronic library management systems. With their online public access catalogs (OPACs), they possess all the requirements in almost the same manner as digital libraries for electronic value-added services. A very promising add-on for traditional libraries are recommender systems, the necessity for which arises from the need of scientists and students for efficient literature research, as shown by the survey of Klatt et al.[1] Due to—among other things—information overload and difficult quality assessment, information seekers are more and more incapable of compiling relevant literature from conventional database-oriented catalog systems in a time-efficient manner. Therefore, as the survey reveals, they rely heavily on peers for recommendations. Considering the tight schedule of many students, university teachers, and researchers, it is worth the effort to free up the valuable time consumed in steering each other to the standard literature of their fields, which could be done easily by behavior-based expert advice services. Moreover, in this scenario, they can also profit from the combined knowledge of all library users in contrast to the more restricted knowledge within their personal networks. Consumer acceptance and convenience of recommender systems are shown by the huge success of the broad variety of different services offered at commercial bookstore sites (such as Amazon.com). People are getting used to these services and appreciate them. So the question to ask is: Why are these services not offered on a broader scale within scientific libraries? Discussing this question with librarians and computer scientists, the following reasons were discovered:

- Privacy. Librarians are very considerate of the privacy of their patrons. Transaction-level data as well as reading histories must be protected.
- Budget restrictions. Public libraries in general run under tight budget restrictions. New electronic services for millions of users might require prohibitively high additional information technology (IT)-investments.
- Data size. The number of documents contained in many public or academic library systems is at least one order of magnitude higher than in most commercial organizations. This implies that transaction-level data is scattered on more documents.

While one would expect that more data implies a better chance for finding meaningful patterns, it becomes increasingly difficult to detect these patterns due to their sparsity, and because the computational complexity of counting such association rules is exponential in the number of objects. Standard association-rule algorithms reduce the complexity by deleting all objects that do not receive sufficient support. In a library context, the sparsity of the data, unfortunately, makes this approach not feasible. Increasing the support threshold to reduce the computational complexity will lead to pruning all meaningful but weak association patterns that may be below the support threshold, but that are still statistically significant. This article presents a strategy to overcome these obstacles with behavior-based recommendations that can be efficiently generated from anonymous session data on off-the-shelf PC systems.

In digital libraries, recommender systems already have a tradition of supporting the search process of users. Fab, for example, was developed as part of the Stanford Digital Library Project.[2] Fab combines a content-based and a collaborative recommender system that filters Web pages according to content analysis and creates usage profiles for user groups with similar interests. PADDLE is a system that introduces customization and personalization features to deal with the information overload caused by the mass of documents in digital libraries.[3] The University of California at Berkeley Digital Library Project allows users to build personalized collections of their documents of interest.[4] Recommendation services for digital libraries and their evaluation methods are discussed by Bollen and Rocha.[5] The virtual university of the Wirtschaftsuniversität Wien offers a collection of categorized Web links and is equipped with

Andreas Geyer-Schulz (geyer-schulz@em.uni-karlsruhe.de), **Andreas Neumann** (neumann@em.uni-karlsruhe.de), and **Anke Thede** (thede@em.uni-karlsruhe.de) are Researchers at the Schroff Chair of Information Services and Electronic Markets, Institute for Information Engineering and Management, Department of Economics and Business Engineering, Universität Karlsruhe (TH), Germany.

many different recommendation and personalization services.[6] Other projects deal with the problem of information overload and intelligent systems for information retrieval in digital libraries, such as query-relevance feedback and information alert systems.[7]

Today, many commercial sites compete by offering a variety of different value-added services that can successfully be added to digital library applications. In fact, online bookstores not offering at least some of these services are not supposed to survive. Amazon.com as market leader offers many different types of recommendation services. But others, such as bol.com, jumped on the bandwagon of recommendations as well, thus achieving a broader range of expertise about documents than a bookstore clerk can possibly provide.

In contrast to digital libraries, no related work can be found in the field of traditional scientific libraries, although these are up-to-date and still the ones offering the broadest and largest variety of literature. Current digital library projects, compared to traditional libraries, are often specialized, comprise less documents, and focused on different types of documents like Web pages not present in traditional libraries. Although this paper is focused on the context of a scientific library, the system it discusses and the underlying behavior patterns are by no means restricted to scientific environments only. It should be stressed that this system could, in principle, augment any existing library system. Moreover, since the university library of the Universität Karlsruhe (TH) acts as information provider for the South-West German library network with twenty-five libraries, it provides recommendations for a variety of libraries including community libraries like the City Library of Karlsruhe and the Central Library of Baden (Badische Landesbibliothek).

This article is structured as follows. The Distributed Architecture for Library Recommender Systems section shows the scalable architecture of a recommender system for legacy libraries. This distributed-agent-based service is operational at the university library of the Universität Karlsruhe (TH). The Generation of Recommendations section presents the mathematical model adapted from Ehrenberg's repeat-buying theory used to generate the recommendations and address the questions of privacy and data size involved, especially compared to other currently available techniques. An evaluation of the system is given in the Evaluation section.

Distributed Architecture for Library Recommender Systems

The recommender services implemented at the library of the Universität Karlsruhe (TH) are based on a generic, three-level architecture whose main idea is described by the pattern of a library with agents.[8] An agent, as described in Russell and Norvig, perceives its environment through sensors and acts upon it using its effectors.[9] Humans as well as software systems can be agents. An example of an agent is a part-picking robot with a conveyer belt and parts as its environment. The robot perceives pixels of varying color and intensity on the belt and acts by picking up the parts and placing them into bins, with the goal of correctly sorting the parts. Figure 1 shows the architecture of the system. The first, lower level consists of the data sources that reside at the legacy library system as well as at the recommendation system's site. Additionally, external library sources are also integrated into the legacy library application (twelve library catalogs hosted at the university library, thirteen external catalogs). The dashed lines indicate the system boundaries. The second level, separated by the dotted line from the first level, contains the agents for the library system as well as for the recommendation service. The third level represents the user-interface system.

Different active agents collaborate in order to make up the library and the recommendation service. The agents' environment consists of the lower-level data sources as well as the upper-level user-interface agent. The agents detect changes in the environment through their sensors and take appropriate actions to such changes. The arrows in figure 1 indicate which environment resources these agents collect information from and which resources are, in turn, altered by them. Bold arrows represent information that is transported over system boundaries. The data exchanged between agents from the recommendation system and the legacy library system are transported using the hypertext transfer protocol (HTTP). The OPAC agent represents the online catalog functionality of the legacy library application. The agent waits for user-agent requests for document information, extracts that information from the corresponding library sources, and returns the data to the user agent, where it is presented to the user. At the same time, the request is inserted into the transaction logs for later analysis.

A second agent is involved in implementing the display of recommendations to the user: the visualization agent of the recommendation system. By incorporating corresponding links in the reply of OPAC, the visualization agent is also invoked by the user-interface agent. The visualization agent checks in the repository of pregenerated recommendation Web pages to ascertain whether recommendations are available for the currently requested document. Then, depending on the request, it returns a link pointing to the recommendation Web page or directly to the list of recommendations to be displayed to the user. The transaction agent and the aggregation agent cooperate to generate the recommendations. The transaction agent regularly sends the latest transaction logs to the aggregation agent. The aggregation agent ana-

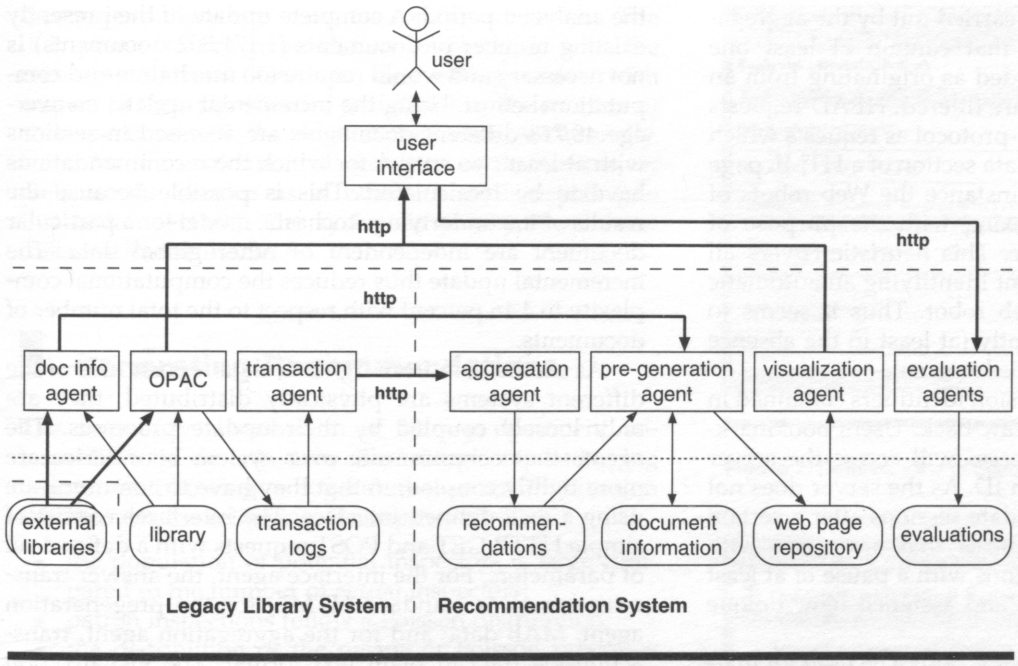


Figure 1. Architecture for a library with recommender services

lyzes the information contained in the logs and calculates recommendations for each possible information object. The recommendations are saved into a repository in the recommendation system.

The information objects of the university library system are described using the German standard representation for libraries' metadata, the Maschinelles Austauschformat für Bibliotheken (MAB) format for books and journals.[10] MAB is a textual format that can easily be transmitted using HTTP. Each single information object can be uniquely identified by the combination of its document number and the name of the library or catalog to which it belongs. Recommendations are generated based on the unique identifier without any additional knowledge about the document description. The task of the pregeneration agent is to examine newly generated recommendations, provide the MAB description for each of the recommended items, and generate Web pages ready to be presented to the user for each recommendation list. The agent manages a local repository of the document-description data containing the relevant subset of the information defined by the MAB format for the creation of recommendation lists. Whenever it encounters an information object for which no local description data is available, the pregeneration agent communicates with the doc-info agent residing at the library system. The doc-info agent extracts the MAB description from the corresponding library and returns it to the pregeneration agent, which in turn saves a copy of this information in its local repository to save bandwidth and time for future requests of the same document.

The evaluation agent works independently of the other agents. Its environment comprises the user interface from which it collects opinions of general and expert users and a separate repository for saving the evaluations for later analysis.

The library is regarded as a market for information where each detailed-document information constitutes a product. The inspection of an information object's details is regarded as a purchase at the cost defined by the amount of time the user (consumer) spends on viewing the information. Consecutive inspections of document details by the same user can be identified through the session identifier that is created upon the first search for documents and carried along in all subsequent page requests from the same user interface (that is, the Internet browser). The session identifier is coded in the uniform resource identifiers (URI) of the requests and is thus contained in the Web server's transaction logs. The aggregation agent analyzes the transaction logs and identifies the market baskets for each user. These baskets contain all detailed-document inspection calls having the same session identifier. The recommendations are computed using the contents of the market baskets as described in more detail in the Generation of Recommendations section.

Due to the nature of the Internet, extraction of market baskets from the transaction log has to be preceded by several preprocessing steps that deal with the following problems:

- Filtering of Web robot and other automated requests. The detection of robot requests is not a straightforward task and has been the topic of several recent research papers.[11] Not all robots identify themselves through an appropriate user-agent specification. Other patterns such as regular or subsequent requests with only a few seconds between them, as well as analysis of the remote machine's name or IP address, can be used to filter robots. In

the preprocessing phase carried out by the aggregation agent all sessions that contain at least one HEAD-request are regarded as originating from an automated process and are filtered. HEAD-requests are defined by the HTTP-protocol as requests which transmit only the meta-data section of a HTML page in order to support for instance the Web robots of search engines for indexing with the purpose of reducing network traffic. This heuristic covers all requests with a user agent identifying an automatic process as a known Web robot. Thus it seems to catch Web robots efficiently (at least in the absence of robots pretending to be humans).

- Bookmarked session. Session identifiers contained in URIs have one major draw-back: Users bookmarking this URI for later access will access the server with an outdated session ID. As the server does not have the ability to invalidate sessions after a certain timeout and detect requests that are made with invalid session IDs, sessions with a pause of at least fifteen minutes are split and assigned new, unique session IDs.
- Requests without session IDs cannot be used for market-basket analysis and are deleted.
- Public-terminal access. Public terminals that can be used without personal login may generate requests with the same session identifier but with different users if the explicit closure of the session is omitted. If users change quickly and the network identifications of the public-access terminals are unknown, the appropriate end of a user session is difficult to identify. Still, combinations of specific documents inspected by different users into a larger market basket will occur randomly. Random co-inspections of the same documents are filtered by the stochastic model used for calculating recommendations; thus no false recommendations are produced from the data contained in the multi-user session.

Preprocessing reduces the number of lines in the log file by approximately three quarters. All log entries not containing an HTTP-GET request to the library's search interface are filtered out. This includes, for example, image files or in-site navigation requests. No relevant behavioral data is lost. The aggregation agent still has to calculate recommendations from, on average, about 142,000 detailed inspections per week. The transaction logs are transmitted on a daily basis at 1 A.M. when user network traffic is low. Computation of the new recommendations is performed regularly four times a month. This update rate is a compromise between the timeliness of the recommendations and the computational expense of each update. The update is done during off-peak hours at night. An incremental update was implemented that accesses only those documents inspected by a user during the analyzed period. A complete update of the presently existing number of documents (1,171,502 documents) is not necessary and would require too much time and computational effort. Using the incremental update, on average 48,718 different documents are accessed in sessions with at least two entries for which the recommendations have to be recalculated. This is possible because the results of the underlying stochastic model for a particular document are independent of other global data. The incremental update thus reduces the computational complexity to 4.16 percent with respect to the total number of documents.

As can be seen from figure 1, data repositories of the different systems are physically distributed; they are only loosely coupled by their update processes. The agents that communicate over system boundaries are more tightly coupled, in that they have to communicate using a well-defined interface. The interfaces consist of simple HTTP-GET and POST requests with a defined set of parameters. For the interface agent, the answer transmitted back is standard HTML; for the pregeneration agent, MAB data; and for the aggregation agent, transaction-log data in plain text format. The visualization agent uses a special technique for displaying a link to the recommendation page. This link is incorporated in the document-detail page delivered by OPAC that is only visible whenever recommendations are actually available for the current document, and can be activated by clicking on an image. The document-detail page always contains a link to the corresponding recommendation page that is activated by clicking on an image. The image is a GIF requested from the visualization agent, and returns a transparent, 1-pixel GIF if there are no recommendations, or an image with "Empfehlungen (Others also use. . .)" if recommendations are present (see figure 2). The advantage of this method is that failure of the recommendation server does not affect the usability of the library OPAC. In the worst case, a broken image icon will be displayed, but all remaining functionalities, as well as the transmission speed of the page, remain unaffected.

Currently, the recommendation server resides on a standard PC with a 1.2 GHz AMD Athlon processor and 1.5 GB main memory, running a customized Linux (Kernel Version 2.4.20) based on a Mandrake 8.0 distribution. The software agents are implemented in Perl (5.6) using MySQL as the underlying database system. The system hardware costs approximately $2,500; no license cost for software applies due to the consequent usage of open-source software. Even a small community library should be able to afford these expenses.

The user interface is shown in figures 2 and 3. Figure 2 presents the detailed-document view with all the important information. On the right hand side, a link to the recommendation list is shown indicating that

recommendations are available for this document. The link points to the page shown in figure 3, with the recommendations listed in descending order with respect to their significance as defined by the number of co-inspections with the original document. Additionally, icons for the evaluation of each pair of documents, as well as the service as a whole, are displayed (see Evaluation section). The system is accessible at www.ubka.uni-karlsruhe.de.

Generating Recommendations

The frequency distribution of library document co-inspections is described by Ehrenberg's repeat-buying theory as following a logarithmic series distribution (LSD) under the following assumptions:

- the population of potential inspectors is large compared to the number of actual inspectors;
- patron inspections follow a Poisson distribution;
- the distribution of the means of Poisson processes follows a truncated gamma-distribution; and
- the market is stationary.[12]

Ehrenberg's classic book on repeat-buying theory remains a readable standard reference and a suitable introduction to consumer panel analysis for the practitioner without a mathematical background.[13] However, in order to be applicable in a library context, two additional insights that are not at the center of Ehrenberg's theory are necessary.

Ehrenberg's theory faithfully models the noise part of buying processes. That is, repeat-buying theory is capable of predicting random co-purchases of consumer goods. Intentionally bought combinations of consumer goods—a six-pack of beer, spareribs, potatoes, and barbecue sauce for dinner, for example—are outliers. In this sense, Ehrenberg's theory acts as a filter to suppress noise (stochastic regularity) in buying behavior.[14]

The second key insight that allows an adaptation of repeat-buying theory for anonymous user-groups is the economic theory of self-selection, which today is routinely used in the design of incentive contracts. A readable example of how this theory is used in hiring and human-resource management is found in Spence.[15] Geyer-Schulz et al. argue that this theory allows observation of aggregate buying processes for self-selected groups of consumers, and thus detects outliers in aggregate processes.[16] This is essential for a library context because it preserves the privacy of the individual and thus addresses the problem of privacy raised in the introduction. A detailed account on this work is in progress.

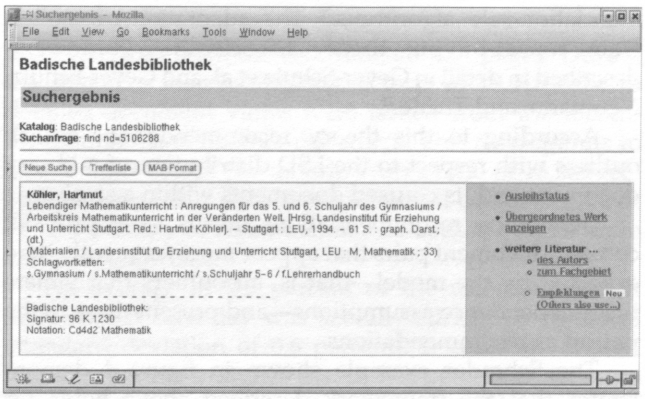


Figure 2. Detailed view of documents

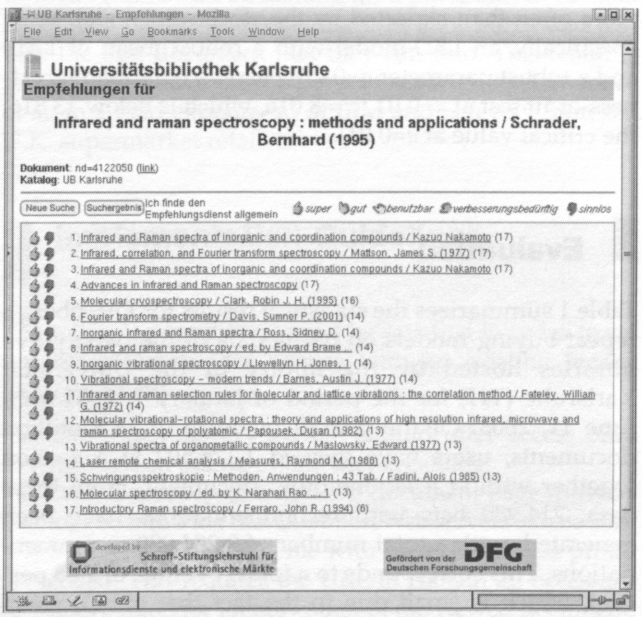


Figure 3. Recommendation list

Ehrenberg's repeat-buying theory was successfully empirically checked for many classes of nondurable consumer products (soap, coffee, toothpaste) and is used today for consumer-panel analysis in marketing.[17] An overview of stochastic-purchase incidence models can be found in Wagner and Taudes.[18] This theory with its strong independence assumptions was utilized by Böhm et al. to describe the regularities of anonymous users in the context of a virtual university.[19] Further, Ehrenberg's theory was tested successfully for the automatic generation of product recommendations at a business-to-business computer accessories dealer and within this context compared to

association-rule algorithms.[20] The adaptation of models from repeat-buying theory to scientific libraries is described in detail in Geyer-Schulz et al. and Geyer-Schulz, Neumann, and Thede.[21]

According to this theory, recommendations are all outliers with respect to the LSD distribution of a library document and its co-used documents within a session.

Generating recommendations means to identify all co-used document pairs that appear more frequently than expected by the model—that is, all outliers that violate the independence assumptions—and present these to the patron as recommendations.

The Schrader example shown in figure 4 demonstrates that the framework described above holds for libraries. The observed frequency distribution f (x obs) corresponds to the ranking by decreasing number of repeated co-purchases in figure 3. Recommendations are the outliers, or products, that have been bought together more often than expected by the stochastic model. More specifically, an LSD-model with a robust mean of 1.410 and a robust parameter q=0 479 passes a chi square-goodness-of-fit test at a=0 01 (c^2= 8 016, which is below 13 816, the critical value at a=0 01).

Evaluation

Table 1 summarizes the statistical results for Ehrenberg's repeat-buying models on usage data of the twenty-five libraries hosted by the library of the Universität Karlsruhe (TH) for the period of January 1, 2001 until June 11, 2003. Of the overall quantity of 14.5 million documents, users have viewed 1,171,502 in a session together with at least one other document. From these data, 214,980 lists with recommendations have been generated, with a total number of 2,204,980 recommendations. This corresponds to a total coverage of 1.48 percent, which is small due to the fact that only a small percentage of these documents were actually inspected during the observation period. The coverage with respect to those documents actually inspected in two or more sessions is 18.4 percent. Since only a small number of the documents within the catalogs are inspected by the majority of the patrons, the percentage of recommendations in respect to detailed inspections shows an even higher number. A second, separate statistic (this time computed for the time period from January 1, 2001 to June 18, 2003) shows that for 46.21 percent of the 3,083,840 visited detailed inspection pages recommendations can be provided.

For evaluation of user acceptance of the recommendation service as a whole, a five-item Likert scale was inserted into the recommendation list (see figure 3). The user could choose between "dispensable" (1), "needs improvement" (2), "usable" (3), "good" (4), and "super" (5). During the period from February 7, 2003, to June 23, 2003, a total of 723 votes were collected. The distribution of votes is shown in figure 5, with a mean of 4.26, and a variance of 0.97. The development of the mean can be seen in figure 6. The results show that a very high percentage of users of the recommendation service are very satisfied or satisfied with the availability and quality of the service. The strong stability of the mean over time shows that no major changes in user acceptance can be expected during future observation periods.

Although user acceptance seems to be high, such Internet-based opinion surveys raise the problem of positive bias by self-selected users and of what constitutes a representative sample for such surveys. One way of controlling for such bias ex post facto is in testing the distribution of additional known attributes of the sample in order to detect significant deviations in the distribution with respect to the population. Due to the anonymity of the survey, standard demographic attributes cannot be checked. However, the ratio of delivered votes to recommender usage was checked to see whether it remained constant. Additional candidates to be checked would be the distribution of votes and recommendation list accesses with regard to different topics or library catalogs. With each additional attribute tested, the repre-

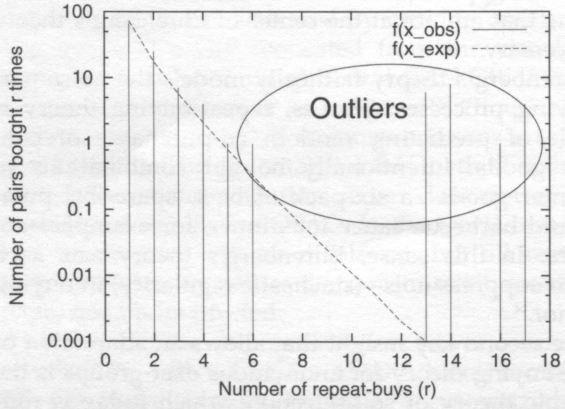


Figure 4. Logarithmic plot of the frequency distribution of infrared and raman spectroscopy by Schrader

Table 1. Detailed results for observation period, January 1, 2001 to June 11, 2003

	q undef.	no x^2 (< 3 classes)	Sign. % = 0.05	Sign. % = 0.01	Not sign.	3
3	756,852 (0)	336,070 (142,424)	21,159 (21,019)	24,445 (21,832)	32,976 (29,705)	1,171,502 (214,980)

sentatively of the sample becomes more credible. However, this remains to be checked.

Figure 7 shows the distribution of the number of daily votes in the upper graph and the distribution of the daily recommendation list calls for the same period in the lower graph. First, the curves in figure 7 were smoothed using a seven-day moving average to filter the week-specific patterns. From the smoothed values, the ratio of votes-per-recommendation-list calls was calculated for each day in percent. The mean of these ratios is 5.43 percent with a standard deviation of 1.09 percent. The box plot in figure 8 shows the distribution of these ratios. The extreme ratios are 3.19 percent and 7.87 percent. The median ratio is 5.45 percent and the two quartiles around the median lie between 4.53 percent and 6.22 percent. The plot shows that the ratios are nearly normally distributed with the mean near the median. Actually, 65.77 percent of all ratios are within a distance of one standard deviation from the mean. No significant outliers were observed. This shows the acceptability of the hypothesis that there is no significant change in the number of votes collected relative to the number of recommendation lists called during a day.

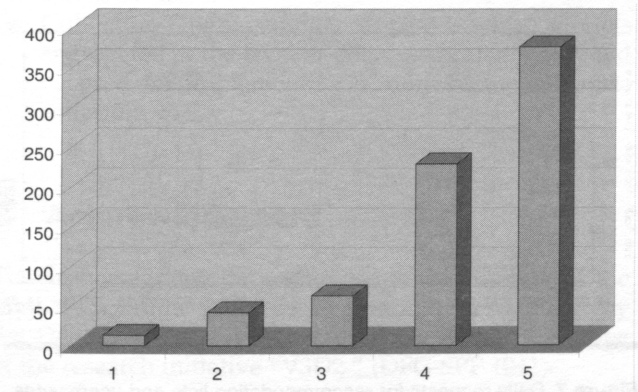


Figure 5. Opinions of users about the recommendation service (dispensable [1] to super [5])

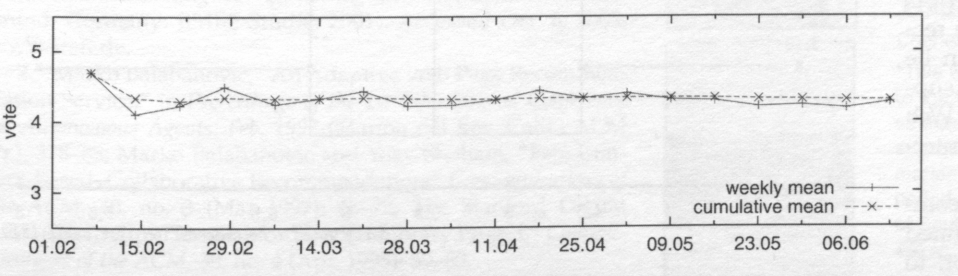


Figure 6. Mean value of users' opinions about the recommendation service (dispensable [1] to super [5])

Implicit preferences can be drawn from the number of links followed from the recommendation pages. During the first five months in 2003, a total of 15,115 calls to detailed-document views were issued from the recommendation list of another document. This corresponds to an average of 3,023 per month. With a total of 16,269 recommendation lists shown per month, it can be seen that in 18.6 percent of recommendation lists, the links were actually followed.

An average of 2.6 percent of the recommendation links shown each day was actually clicked on by users (standard deviation of 0.6 percent). It could be argued that this number indicates that in 97.4 percent of the cases the recommendation service was dispensable. But there are many diverse reasons for users not to be interested in recommendations. Recommendations are useful when browsing through a topic for interesting literature. But users might instead be looking for a particular document (such as a course book) or might be in a hurry and just take the first document that they run into. This result is still slightly above the 1.8 percent reported by Lawrence et al. in a field experiment of IBM and Safeway Stores, a U.K. supermarket retailer.[22]

Implementation Guidelines

The following conditions must be satisfied in order to set up this recommender service with an existing legacy library system:

- The legacy library system should offer access only through an online public access catalog.
- Each inspection of a detailed document page must be logged into a log file. Common Web servers can be configured to write a log file with each URL access protocolled. This requires the URL to contain all necessary information for identification of the viewed document.
- Subsequent inspections from the same user client must be able to be identified. This can be realized using either cookie technology or by personalizing the users' URLs by inserting unique session identifiers. As the session identifiers appear in the URL, the information can be extracted from the Web-server log file. The cookie mechanism requires the script that performs the document inspection to identify known cookies and save the information for later analysis.

- The point in time of each inspection must be known. The exact time of the Web-page request is contained in the Web-server log.
- The detailed-document page must be modified to include the extra image link that checks whether recommendations are available for this particular document. The link contains the name of the recommendation-list script as well as all information necessary for identification of the viewed document and all remaining parameters necessary for calling the detailed document page. This permits the recommendation page to present a link back to the detailed document page, which is important for reasons of usability.
- The recommendation server must be accessible from the Internet as it delivers the link image as well as the recommendation list directly to the user client.
- The legacy library Web server and the recommendation server must be connected via HTTP or FTP for transferring the above one or more textual log files. Regular execution of programs can be configured in most operating systems; a simple program implementing an HTTP file upload or an FTP file transfer can be used. The recommendation server must accordingly provide a simple Web script that handles HTTP file uploads respectively an FTP server.
- The recommendation server must be able to translate the document identifier to information about the document (title, author). There are several possibilities for bringing this about. The optimal choice depends on how and where this information is stored (for example, direct database connection, HTTP transmission with data coded in plain text or XML).

On the recommendation server, a job must be started on a regular basis, such as once a week, to examine the newly transferred log files, do the preprocessing, and update the recommendations from the new information product baskets. The necessary data can be stored into simple text files so no database applications are needed. Two different scripts implement the link image and the recommendation list. The scripts can be implemented in any language that supports communication with the Web server via the common gateway interface (CGI).

The Web pages with the recommendation lists can either be generated on the fly at each user access or, to increase performance, be pregenerated after calculation of the new recommendations. The choice depends on the frequency of recommendation updates, on the performance of the Web-side generation, and on the availability of the necessary hard-disk space.

Future Research

The transfer of repeat-buying theory suggests that today's libraries have much in common with retail chain management, at least as far as the underlying stochastic

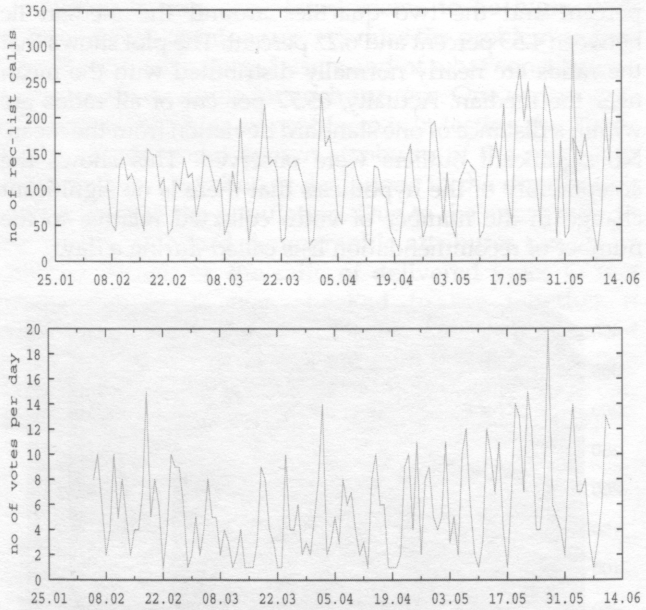


Figure 7. Daily requests for recommendation lists and users votes

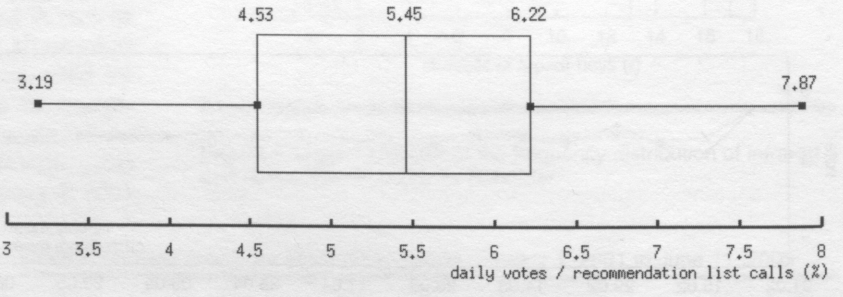


Figure 8. Box plot of the ratios of smoothed daily user votes per smoothed recommendation list calls (in percent)

purchase models are concerned. For further research, the following topics will be addressed:

- Higher-order associations. Ehrenberg's model generalizes to higher-order associations that can be used to generate path-specific recommendations for individual users.
- Fashions, trends, dynamics. By using sliding time windows over the data set nonstationarities can be identified by simple statistical tests. This could be used to summarize purchase histories for individual documents of similar document categories in order to visualize reading trends and fashions.
- Personalization. Combining self-selection according to experience in a scientific field allows recommendations to be generated that are tailored to the individual interest and ability profile of a patron. For example, in an academic environment such recommendations support individual tutoring of students with heterogeneous educational backgrounds.
- Content analysis and clustering for catalog management. Significant outliers provide a new source for document clustering which may lead to an improved access by users.
- Collection management. Repeat-buying models aggregated at the level of document categories could be used for implementing a more customer-centric collection policy.

Acknowledgment

The authors gratefully acknowledge the funding of the project "Scientific Libraries in Information Markets" by the Deutsche Forschungsgemeinschaft within the scope of the research initiative "V3D2 " (DFG-SPP 1041).

References

1. Rüdiger Klatt et al., *Nutzung und Potenziale der innovativen Mediennutzung im Lernalltag der Hochschulen* (Dortmund, Germany: BMBF-Studie, 2001). Accessed Oct. 6, 2003, www.stefi.de.

2. Marko Balabanovic, "An Adaptive Web Page Recommendation Service," in *Proceedings of the 1st International Conference on Autonomous Agents, Feb. 1997* (Marina del Rey, Calif.: ACM Pr.), 378–85; Marko Balabanovic and Yoav Shoham, "Fab: Content-Based, Collaborative Recommendation," *Communications of the ACM*, 40, no. 3 (Mar 1997): 66–72; The Stanford Digital Libraries Group, "The Stanford Digital Library Project," *Communications of the ACM*, 38, no. 4 (Apr. 1995): 59–60.

3. David Hicks, Klaus Tochtermann, and Andreas Kussmaul, "Augmenting Digital Catalog Functionality with Support for Customization," in *Proceedings of 3d International Conference on Asian Digital Libraries* (Berlin: Springer, 2000), 155–61.

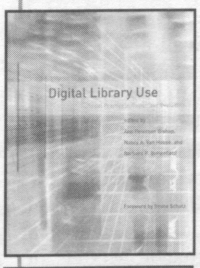
4. Robert Wilensky et al., "Reinventing Scholarly Information Dissemination and Use," Technical Report, University of California, Berkeley, 1999. Accessed Oct 6, 2003, http://elib.cs.berkeley.edu/pl/about. html.

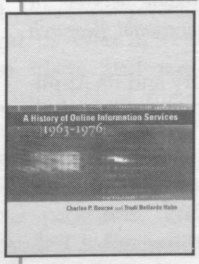
5. Johan Bollen and Luis M. Rocha, "An Adaptive Systems Approach to the Implementation and Evaluation of Digital Library Recommendation Systems," in J. Borbinha and T. Baker, eds., *Proceedings of the 4th European Conference on Digital Libraries*, vol. 1923 of *LNCS* (Heidelbert, Germany: Springer, 2000), 356–59.

6. Andreas Geyer-Schulz, Michael Hahsler, and Maximilian Jahn, "A Customer Purchase Incidence Model Applied to Recommender Services," in R. Kohavi et al., eds., *Proceedings of the WebKDD–Mining Log Data across All Customer Touchpoints*, vol. 2356 of *Lecture Notes in Artificial Intelligence LNAI* (Berlin: ACM, Springer-Verlag, 2002), 25–47.

7. G. Salton and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science* 41, no. 4 (1990): 288–97; Ioannis Papadakis, Ioannis Andreou, and Vassileios Chrissikopoulos, "Interactive Search Results," in M. Agosti and C. Thanos, eds., *Research and Advanced Technology for Digital Libraries*, vol. 2458 of *LNCS* (Berlin: Springer, 2002), 448–62; Ann Apps and Ross MacIntyre, "Prototyping Digital Library Technologies in zetoc," in M. Agosti and C. Thanos, eds., *Research and Advanced Technology for Digital Libraries*, vol. 2458 of *LNCS* (Berlin: Springer, 2002), 309–23; Norbert Fuhr et al., "Daffodil: An Integrated Desktop for

New from The MIT Press





Digital Library Use

Social Practice in Design and Evaluation

edited by Ann Peterson Bishop, Nancy A. Van House, and Barbara P. Buttenfield
foreword by Bruce Schatz

Examines digital libraries as sociotechnical systems, networks of people and technology interacting with society.

Digital Libraries and Electronic Publishing series
328 pp., 15 illus. $40

A History of Online Information Services, 1963-1976

Charles P. Bourne and Trudi Bellardo Hahn

"This work will be a significant contribution to the field! Because it is carefully researched and documented, it will be the standard source for all historians of information science and technology of the period." — Robert V. Williams, University of South Carolina

496 pp., 11 illus. $45

To order call **800-405-1619**.

http://mitpress.mit.edu

Supporting High-Level Search Activities in Federated Digital Libraries," in M. Agosti and C. Thanos, eds., *Research and Advanced Technology for Digital Libraries,* vol. 2458 of *LNCS* (Berlin: Springer, 2002), 597–612.

8. Andreas Geyer-Schulz and Michael Hahsler, "Pinboards and Virtual Libraries—Analysis Patterns for Collaboration," *Technical Report 1, Institut für Informationsverarbeitung und -wirtschaft, WirtschaftsUniversität Wien, A-1090* (Wien: Institut für Informationsverarbeitung und -wirtschaft, 2001). Accessed Oct. 6, 2003, wwwai.wu-wien.ac.at/~hahsler/research/virlib_working2001/virlib.pdf.

9. Stuart Russell and Peter Norvig, "Introduction and Survey of AI," in *Artificial Intelligence: A Modern Approach—The Intelligent Agent Book* (Upper Saddle River, N.J.: Prentice-Hall, 1995), 31.

10. Die Deutsche Bibliothek, *MAB2: Maschinelles Austauschformat für Bibliotheken,* 2d ed. (Leipzig/Frankfurt am Main: Die Deutsche Bibliothek, 1999).

11. Robert Walker Cooley, "Web Usage Mining: Discovery and Application of Interesting Patterns and Web Data" (Ph.D. diss., University of Minnesota, 2000); Pang-Nin Tan and Vipin Kumar, "Discovery of Web Robot Sessions Based on Their Navigational Patterns," *Data Mining and Knowledge Discovery* 6, (2002): 9–35.

12. A. S. C. Ehrenberg, *Repeat-Buying: Facts, Theory, and Applications,* 2d ed. (London: Charles Griffin, 1988).

13. Ibid.

14. Andreas Geyer-Schulz et al., "Behavior-Based Recommender Systems As Value-Added Services for Scientific Libraries," in Hamparsum Bozdogan, ed., *Statistical Data Mining and Knowledge Discovery* (Boca Raton, Fla.: Chapman and Hall/CRC, 2003), 433–54.

15. Michael A. Spence, *Market Signaling: Information Transfer in Hiring and Related Screening Processes* (Cambridge, Mass.: Harvard Univ. Pr., 1974).

16. Geyer-Schulz et al., "Behavior-Based Recommender Systems."

17. P. Charlton and A. S. C. Ehrenberg, "Customers of the LEP," *Applied Statistics* 25 (1976): 26–30; C. Chatfield, A. S. C. Ehrenberg, and G. J. Goodhardt, "Progress on a Simplified Model of Stationary Purchasing Behavior," *Journal of the Royal Statistical Society* A 129 (1966): 317–67; N. Powell and J. Westwood, "Buyer-Behaviour in Management Education," *Applied Statistics* 27 (1978): 69–72; H. S. Sichel, "Repeat-Buying and the Poisson-Generalised Inverse Gaussian Distributions," *Applied Statistics* 31 (1982): 193–204; SRS, "The SRS Motorists Panel," Technical Report (London: Sales Research Service, 1965); Aske Research, "The Structure of the Toothpaste Market," Technical Report (London: Aske Research Ltd., 1975).

18. Udo Wagner and Alfred Taudes, "Stochastic Models of Consumer Behaviour," *European Journal of Operational Research* 29, no. 1 (1987): 1–23.

19. Walter Böhm et al., "Repeat-Buying Theory and Its Application for Recommender Services," in Otto Opitz and Manfred Schwaiger, eds, *Exploratory Data Analysis in Empirical Research,* vol. 22 of *Studies in Classification, Data Analysis, and Knowledge Organization,* Gesellschaft für Klassifikation e.V. [German Classification Society], (Heidelberg: Springer-Verlag 2002), 229–39.

20. Andreas Geyer-Schulz, Michael Hahsler, and Anke Thede, "Comparing Association-Rules and Repeat-Buying-Based Recommender Systems in a B2B Environment," in *Between Data Science And Everyday Web Practice, Studies in Classification, Data Analysis, and Knowledge Organization* (Heidelberg: Springer-Verlag, 2003).

21. Geyer-Schulz et al., "Behavior-Based Recommender Systems," 433–54; Andreas Geyer-Schulz, Andreas Neumann, and Anke Thede, "Others Also Use: A Robust Recommender System for Scientific Libraries," in T. Koch and I. T. Solvberg, eds., *Research and Advanced Technology for Digital Libraries, 7th European Conference, ECDL 2003, Trondheim, Norway, 2003: Proceedings,* Lecture Notes in Computer Science #2769 (Berlin: Springer Verlag, 2003), 113–25.

22. R. D. Lawrence et al., "Personalization of Supermarket Product Recommendations," *Data Mining and Knowledge Discovery* 5 (2001): 11–32.

SIMA Camelot

Integrated Library System

- Highly affordable
- Powerful, full-featured
- Easy to use, easily customized
- Install on your network (LAN/WAN) or ours (SIMAnet for Internet access)
- Catalog, Utilities, Reports plus optional Web-access catalog
- Serials/Acquisitions/Accounting
- Circulation with Bar Coding

SIMA, Inc.

An application service provider

www.simainc.com 703-569-0993

simainc@simainc.com

Traces in the Clickstream: Early Work on a Management Information Repository at the University of Pennsylvania

Joe Zucca

For the past three years, the University of Pennsylvania (Penn) library has been building a data repository and developing computer functionality to support management information needs. This article traces the origin and evolution of Penn's evolving management information system (MIS) program, known as the data farm. It addresses problems pertaining to the collection, storage, anonymization, and normalization of data, and looks at current work on a database-driven model for future MIS functions.

In recent years, the range of academic library services powered by relational databases and servers has increased substantially. Our customers access electronic journals and search article indexes on the Web. They wand barcodes to charge-out books, use debit cards to copy or print documents, swipe ID badges to enter library buildings . . . and the list goes on. These interactions with the library give rise to a flood of transaction data recorded by the databases and computers that drive service. The machines register when and where services are used and contain information about individuals and the resources people access. Occasionally they even capture clues about an individual's work environment; the traces of a cookie or a referring URL can provide such indicators; the location of a photocopier may offer others. The powerful tools of library service can support equally powerful tools of library assessment. In the right organizational culture, those assessment tools can inform and thus improve planning and decision-making. They can serve less benign purposes as well absent policies and procedures designed to safeguard identities

Three years ago, motivated by the need to improve the measurement of electronic resource use, the Penn library began working with various kinds of transaction data, and in particular Web server logs, as sources of management information. Though narrowly defined in its early stage, this effort at log analysis is beginning to resemble the form and function of an MIS. The following is a brief survey of the embryonic MIS at the Penn library, an aggregation of data files and databases, form interfaces, and Web pages called the library data farm.

Initial Challenges

By 1998, the Penn library offered a large collection of online article indexes, full-text databases, and electronic journals through its Web site. As this quickly evolving, costly set of services expanded, frustrations mounted over the difficulty of measuring its use and impact. The lack of good management information, if only in the form of frequencies or other descriptive statistics, was viewed as a serious shortcoming, one that could hamper the library's accountability to the university's schools and impede planning and budgeting.

For several years, the library staff has worked diligently to compile the few statistics that vendors provide and trace the spiral of Web activity based on Web-site visits and page counts. But these attempts at measurement have had obvious drawbacks. Vendor statistics have been meager, erratic, poorly defined, and incompatible across products. Web measures have provided information about machine load, but have revealed nothing about resource use. Together, the external and internal sources have contributed little clarity to the picture of digital information use. In addition, the compilation of these crudest of measures has proved too time consuming and labor intensive to carry out with any regularity.

Given the absence of third-party solutions or working models, any approach to the management information problem would require a lot of experimentation with local data sources and systems. The approach would have to be independent of information providers and sufficiently robust to generate useful statistics with a modicum of labor. In short, it would have to provide a means of:

1. increasing the resolution of library statistics, especially with regard to demographics and cost analysis;
2. counting with reasonable accuracy and sustainable methods;
3. reducing the level of effort involved in harvesting and converting data into information; and
4. improving the consistency and reliability of data collection and administration.

The key to meeting these objectives exists in the self-monitoring capabilities of the library management system (the Penn LMS is the Voyager product from Endeavor Information Systems) and other components of the information technology (IT) infrastructure, specifically standardized and configurable logging processes.

Prior to the data farm, the library viewed logs as a byproduct rather than an end of service, and thus afforded logging low operational priority. At Penn—and surely elsewhere—logs were often discarded or lost whenever servers ran short on storage, or machines were rebuilt or

Joe Zucca (zucca@pobox.upenn.edu) is the Assessment, Planning, and Publications Librarian, University of Pennsylvania Library, Philadelphia.

redeployed. Even seemingly trivial attempts to count the number of HTML pages served often fell short for want of data. If logs were to provide the basis for measurement, then logging processes would have to be routinized to sustain and ensure the availability of log inputs. The data farm traces its origin to these narrow objectives.

Initial Steps

In winter 2000, work on log analysis began in earnest, focusing first on logs generated by the servers supporting open and proxied Web access. Server logs are enormous. As a consequence, they have high administrative and storage demands. In their raw form, they contain only a small percentage of useful information, and require substantial refinement before analysis can take place. To address the handling and storage issues, staff devised Perl scripts that roll log data nightly from production servers into a two-hundred-gigabyte networked storage system established for assessment purposes. Other Perl modules parse the huge, nightly dumps, then extract and save the outputs into smaller, content-rich files. The networked storage array has three partitions: one with folders accepting the massive raw logs, another with folders for processed log extracts, and a third with folders for reports.

Improving Data Resolution

Resource Identifiers

The pieces of extracted content stored in processed folders include the resolved host names of incoming IP addresses, date and time stamps, and referring URLs. Each element provides a different key to sharpening the resolution of the usage data. The most important key is a certain referring URL designed to log each time a user requests a licensed resource, such as ABI/Inform. This counting referrer captures resource names before routing visitors to a vendor or another remote site considered statistically important. In this way, a channel was opened in the log stream for marking and later counting events that would otherwise evade our scrutiny.[1]

Demographic Information

The second key to higher resolution data involves identifying the whereabouts of visitors to the library Web site and harvesting demographic information from these location sources. Network IP addresses make it possible to trace all on-campus Web activity to offices, labs, and residential facilities associated with Penn's twelve schools and other budget authorities. To demographically analyze the on-campus segment of traffic, library staff constructed a detailed map of host names derived from numerical IP addresses and the campus assignments database. All off-campus requests for licensed services require proxy authentication with Penn credentials, and library staff has recourse to good demographic information for this portion of Web activity by linking the Penn credential to school and status descriptors contained in the library's patron database. As nightly log dumps from the main Web server and proxy server are processed, scripts match IP information to the campus IP map, and credentials to library patron records. The scripts then execute processes that replace IP data and Penn IDs with demographic markers for schools or campus centers. With this step, staff is able to attribute nearly all traffic involving licensed resources to a demographic source and anonymize library data files.

Normalized Inputs

The process of cleansing and normalizing anomalies in the raw log files presented significant problems. The names of resources parsed out of the referring URL often appear in variant forms (four or five different instances of ABI/Inform, for example) or with control characters introduced by browsers. The variability of title entries caused sorting errors and garbled the first log outputs. To deal with this, resource names taken from the raw log are rendered into continuous strings of alphabetic characters and matched against lists of similarly rendered title variations and normalized equivalents. A script substitutes normalized versions for aberrant entries, with about 98 percent accuracy; in the end only a handful of titles appear in data farm reports in more than one form. The normalization list also provides publisher name and information about formats (for example, that a resource contained abstracts, full text, PDFs, or other features).

The final file extracts used to build statistical reports comprise fields with demographic variables derived from the IP map and patron database, date and time stamps, normalized resource names, publishers, and format codes.

Reports

Within a few weeks, these processes generated a large, growing data repository, ready for input into SAS software for detailed analysis. This enabled library staff to count the number of times individuals logged on to licensed resources, with breakdowns by day, time, and location. It was possible to cross tabulate the use of specific resources by school, and to distinguish for the first time on-campus from off-campus access. Using fund information extracted from the library's LMS, staff could identify per-login costs and begin gauging the cost efficiency of licensed products.

The repository was partitioned into monthly reports. Each appears in HTML format and in a delimited text form for downloading. The

entire data set was then mounted on a library server for easy access by staff. By the end of this phase, the time frame required to extract, process, and distribute data had been reduced from weeks to hours, and the resolution of library statistics improved dramatically.

Expanding Functions

The data farm operation quickly began developing Perl modules to extract and parse data from a variety of transaction files and to spawn statistics for mounting on the Web. These additional transaction data include digital and physical resources. Since 2000, the data farm has been producing anonymized monthly reports following the procedures or refined versions of the procedures previously described. Demographic information (school, in some cases department, or administrative center) and, wherever possible, the service location, are standard features of each report. A list of reports appears in figure 1.

From Static to Dynamic/Integrative Measurement

Until recently, data-farm reports appeared solely as tables in static HTML format and as delimited text files which are downloadable from the Web. Beginning in January 2003, however, library staff began logging transactions asynchronously and also in real time into Oracle tables, thus constructing a set of query-able databases of usage information. The databases continue to produce summary statistics in static monthly snapshots, but now dynamic report-building capabilities supplement the automated summaries. The report builders offer staff the option of generating more granular and focused outputs than the summaries provide. Report building functions are based on a set of templated SQL (sequential query language) queries that are made against the underlying Oracle structures with the use of Web forms. Thus, staff can build reports in the data farm without having to know SQL programming. So far, the data farm uses Oracle databases to generate measures for the library's five image collections and to track reference/instructional activity. A third Oracle-based application, presently in development, will enable staff to generate reports detailing the use of all electronic resources originating from the library Web, whether licensed or freely available.

Report	Metrics
Database & e-journal use (non-proxy)	Logins sorted by title and frequency (in html and text format)
Database & e-journal use (proxy)	Logins sorted by title and frequency (in html and text format)
OPAC activity	No. of known item searches, browses, redirects, help use, sessions, pages hits, frequency of search parameters, searches by IP domain
Circulation (book checkouts)	Checkouts by status, by subject class and status, by subject class and IP domain
Photocopier/printer use	Pages copied and printed by library location (down to the floor) and IP domain
E-book use (Oxford & Cambridge/ Mellon-funded Project)	Titles viewed, titles downloaded by IP domain
Gate counts (based on ID card swipes)	Count of visitors cross tabulated by visitor status and school. Cross tabulations of visits by day and hour
Image collection use	Frequency of image views, search types, views by IP domain. Summary data plus report builder
Browser analysis	Frequency of unique IP addresses hitting the main Web site and proxy server by browser type and version
Consultative & instructional session	Frequency of consultative & instructional sessions systemwide, with demographic breakdowns, mode of session (eg in-person, chat) measures of session length and length or prep time, locations. In summary and session by session form.
Annual data collection	Metrics used by ARL and IPEDS statistical reports
Visits to www.library.upenn.edu*	Frequency of unique hosts hitting the library Web
Online books page use*	Pages served
Blackboard courseware use*	Pages served

*These reports are generated by WebTrends software and the only uses of a third-party application in the data farm.

Figure 1. Examples of Data Farm Reports

In summer 2003, Penn will migrate the data farm to a Sun-Solaris server and a dedicated Oracle installation. The goal will be to develop table space integrating all the transaction data available to the data farm, enabling staff to move from simple, descriptive statistics to more interesting forms of data mining. In addition to integration, the Oracle migration will provide a number of operational benefits, including:

- greatly reduced data storage demands;
- faster development and processing of descriptive statistics than log-mining supports;
- the enabling of links to external data sources, such as LMS, for fund information and various kinds of bibliographic variables;
- streamlined processes to anonymize log data; and
- a more robust IP and ID mapping system for demographic resolution.[2]

Moving beyond Transaction Data

The data farm is gradually expanding to include quantitative information generated by staff rather than by machines. An example is the Library Reference/Instructional database previously mentioned. Here a set of interactive forms provides a mechanism for librarians to record information about reference consultations and various kinds of instructional contacts. Input variables describe the type of consultation or classroom contact, the length of preparation and session time, demographic parameters, location information, and outcome indicators. The forms also contain free-text sections that librarians can use to store information about sources and search strategies they may use in future reference contacts. The database provides a downloadable text file containing staff inputs that can be opened and massaged in Excel for individual staff reporting needs. The database also spawns summary statistics, which are updated daily, for managers to review system-wide. A similar form-based system has been constructed for collecting annual operational data, including those used in reports to the Association of Research Libraries and Integrated Postsecondary Education Data System (IPEDS).

In the coming year, the data farm will attempt to mount a library staff census: a human resources database describing library positions and employees. The census will provide a means of tracking changes in staff deployment and library organization over time, and it may feed content to other databases for creating directories or Web pages.

Finally, the data farm acts as a repository of knowledge derived from survey research, ad hoc statistical projects—such as the North American Title Count—and various types of program assessments conducted by staff. A longer-term goal for the data farm is to index these knowledge sources and link them through a basic search structure.

Staffing and Resources

Graduate interns from the department of computer and information science of Penn's School of Engineering and Applied Science perform most of the Perl and SQL programming used by the data farm. In a typical year, the library employs approximately 0.5 full-time equivalent of intern assistance. The assessment, planning, and publications librarian devotes about 15 percent of his time to project planning and management. The interns work under his direction, and he reports to the director for finance and administration.

In addition to Perl and SQL, the library has made limited use of Java and XML for certain recurring functions, such as form generation, Penn ID-lookups and IP mapping. PHP is presently the application of choice for serving forms used to query the data farm's Oracle database.

Conclusion

The long term future of the data-farm project is contingent on several factors. The most important is the project's capacity to inform planning and management with useful data. Thus far, data-farm resources have figured in Web and OPAC redesign, facilities planning, and collection development considerations involving e-resource selection. Just as important, the project has provided data that have been extremely valuable in establishing the library's credibility among its funders: the deans and the finance and provostial staff of the university.

The future of the project also hinges on library staff's success at creating a robust set of MIS functions. This will require the further integration of data sources and the creation of functionality that enables a range of staff to quickly and intelligently pose questions to data. Dynamic linkages will have to be built between the data farm and operational systems that support fiscal activities, document delivery/interlibrary loan, cataloging, and circulation functions. Finally, data-farm capacities will have to be built in ways that do not negatively impact the library budget or staff workload. This constraint highlights the importance of building assessment capabilities into the engines of service. If machines take on the heavy lifting involved in data capture and processing, then staff can be free for analysis and knowledge creation. In large part the data farm project owes its existence to decisions that have enabled the library to re-purpose its service delivery technologies for measurement. In the end, Penn's MIS will mature only if the library continues to view measurement as an aspect of service delivery and leverages the architecture that supports service delivery to accommodate management information needs.

Notes

1. The technique described here to capture and count database and e-journal sessions has been modified over time. But whatever the mechanism, the strategy has specific limitations: It does not allow us to capture sessions originating from bookmarks, and it inhibits demographic analysis when sessions originate in computer labs or other locations where machines

are pooled. It has specific benefits as well. It provides uniform and known counting methods. It can be applied to all e-resources, locally created and licensed, and it is amenable to time series analysis. We have found a high degree of positive correlation between our internally-generated counts of database and e-journal use and the measures provided by vendors. Further research on this topic should be available later in 2003.

2. Oracle logging will lessen storage requirements by an order of magnitude. A 20 MB Oracle table contains the equivalent of a month of raw Apache logs that presently consume 200 MB or more.

A Study of the Use of the Carlos III University of Madrid Library's Online Database Service in Scientific Endeavor

Carlos A. Suárez-Balseiro, Isabel Iribarren-Maestro, and Elías Sanz Casado

The identification of variables affecting university research is one of the chief factors in the evaluation of these institutions. This paper explores the relationships between the use of the information resources available through the Carlos III University of Madrid Library's Online Database Service and the results of scientific endeavor within the institution. A two-dimensional analysis was performed, combining the number of database accesses as identified in the monthly activity records furnished by the IRIS CD-ROM database management module with the level of research activity represented by an aggregate index of the results of scientific endeavor, calculated by principal components methods.

University research is conditioned by a number of factors, both objective and subjective. Such factors have a direct impact on the development of the research culture underlying the intellectual dynamics needed for dialogue and collaboration among research scholars. The factors range from inherent institution and department features—number of research scientists, size of budget—to the motivations intrinsically related to the research endeavor itself, such as the pursuit of professional repute, the chance to work with other colleagues, and the opportunity to contribute to the training of new scientists.[1]

In this context, studies on the impact of the university library on the results of scientific endeavor acquire increasing importance. They serve as a measure of how useful the information resources and services provided are for the university community. In this regard and given the importance in the academic domain of the publication of scientific papers, managers responsible for university resources, especially those dealing with information resources, should focus attention on the factors that support research and affect faculty research productivity.[2] According to Kyrillidou, this is a novel approach, since, generally speaking, libraries have not undertaken the systematic evaluation of their services from the standpoint of the effect that their use has on the results of user activities.[3] Moreover, the prevalence of new technologies has led to a heightened interest in studies on the usage and quality of electronic information services and their impact on research conducted in universities.[4]

The purpose of this paper is to conduct a quantitative study of the relations between the results of scientific endeavor in Carlos III University of Madrid departments and the extent to which they use the online databases available in the university library.

Method

The data used for the study were taken from the Research Activity Information System (SINAI) developed by the Carlos III University of Madrid's Research Results Transfer Bureau (OTRI). The SINAI information is published yearly in the university's Research Memorandum and is likewise available on the Internet.[5]

The information respecting online database service was provided by the Infoware integrated electronic information resource management module incorporated in the information technology (IT) system presently in place in the university library. This module is based on Citrix Systems, Inc., Citrix Independent Computing Architecture (ICA), and Citrix MetaFrame technologies. The daily use records stored in this system can be used to quantify the total number of database accesses. In addition, they can identify the respective area or department from which the resource is accessed from the terminal IP address and the user's system administrator-assigned password. In this respect, faculty activity has been clearly differentiated from student activity. Students can only access the databases from computers located in the library, while faculty use their own computers in the departments. Both can be fully identified by filtering IP addresses. The study covered all the university's areas of knowledge that offer diplomas to date, although aggregate categories were created to adjust the departmental structure for easier comprehension of the results.

Research activity is represented by a composite index (CI) of the results of scientific endeavor found

Carlos A. Suárez-Balseiro (csbgv@bib.uc3m.es) and **Isabel Iribarren-Maestro** (iiribarr@bib.uc3m.es) are Research Fellows, and **Elías Sanz Casado** (elias@bib.uc3m.es) is Chair, Director of the Departamento de Biblioteconomía y Documentación, Facultad de Humanities, Comunicación y Documentación, Universidad Carlos III de Madrid.

from four variables related to that process. Essentially, articles in scientific journals and books were regarded to be indicators of the dissemination of research results, doctoral dissertations defended as an indicator of training of new scientists, and patents registered as an indicator of economic contributions. These variables were chosen because they are the items considered to constitute ordinary university department practice by Ministry of Education and Science provisions on evaluation of research activity.[6] The statistical technique used to maximize the representativeness of the index found was principal component analysis. The CI for each department is equal to the factorial score obtained by the department in the first component, assumed to be a summary indicator representing the largest source of variance between cases. In this respect, a department's factorial score is defined by the following expression, used for factor scores calculation in principal component analysis:[7]

$$Fij = Fi1Z1 + Fi2Z2 + ...+ FinZn = \sum_{t=1}^{n} FitZt$$

where *Fij* is the factorial score of department *j* for factor *i*. Each *Fij* is the factorial weight of variable *t* for factor *i*. The individual scores for each variable are represented as *Zt*.

Preliminary Findings

Table 1 shows the values found for the indexes of the results of scientific endeavor (CI) by department and the total number of database accesses over the period 1997 to 2000. The observed fluctuations in the number of database accesses could be explained by changes that occurred in the university library's online database service during the period, such as improvement in hardware and software, changes in databases availability, and other technical and administrative decisions. For instance, in 2000, some changes that took place in the integrated electronic information resource management module facilitated a higher number of concurrent users, and the library budget increased. It was a key factor to assign more licenses for databases. On the other hand, changes in the index of the results of scientific endeavor (CI) are a more complex matter because they are a composite index of the research process and the continuity of the present study will be necessary to explain their evolution.

These indexes (CI) take the value of the department´s factorial score for the first component. Factorial scores (i.e., CI) are normalized values (mean = 0 and variance = 1). The normality condition has been tested running the Kolmogorov-Smirnov Test. The factors scores may have less error and be more reliable measures than the original variable.

The information in the table was plotted on scatter diagrams, not to

Table 1. Index of Results of Scientific Endeavor (CI) and Database Accesses by Department

Departments	1997		1998		1999		2000	
	CI	DB Accesses	CI	DB Accesses	CI	DB Accesses	CI	DB Accesses
Law	2.16	5316	2.07	2352	2.42	1946	0.33	5661
Economics	0.23	5228	0.80	4272	0.15	3956	-0.20	5994
Statistics and Econometrics	-0.22	159	-0.51	113	-0.56	182	0.43	171
Library and Information Science and Documentation	-0.61	200	-1.14	377	-0.46	590	-0.66	861
Humanities	0.10	35	0.30	183	0.60	211	-0.50	371
Physics	-0.99	486	-1.08	613	-0.72	498	-0.66	349
Computer Science	-0.79	158	-0.24	245	-0.67	421	-0.43	198
Engineering	0.85	900	0.20	2384	-0.38	1545	2.41	2080
Mathematics	-0.72	31	-0.40	162	-0.37	117	-0.73	8
Mean	0,00	1390	0,00	1189	0,00	1052	0,00	1744
Asymptotic Sig. (2-tailed)*	0,91		0,97		0,34		0,64	

* Kolmogorov-Smirnov Test for the annual distribution of CI (Large significance values (>0.05) indicate that the observed distribution corresponds to the theoretical distribution)

Note: data processed by authors

draw relationships between the two variables by means of correlation or regression analysis, but to attempt to identify similarities and differences in department practice with respect to the two variables in the years analyzed. The graphs also show the mean number of database accesses per year.

Figure 1 (1997) shows department behavior with respect to the results of scientific endeavor and use of the university library's online databases. The departments with positive CI values, located on the right of the graph, have higher scientific productivity than those located on the left, which have lower, negative CI values. It will be noted that four departments have positive CI scores: law (2.16), engineering (0.85), economics (0.23) and humanities (0.10). With the exception of humanities, these departments were also the ones that used the database service most intensely, in particular law and economics, whose 5,316 and 5,228 accesses, respectively, were considerably higher than the mean value for the year.

A similar pattern emerges in the data for 1998 (see figure 2). The departments with the highest CI values were the same as in the preceding year. Nonetheless, while the highest score (2.07) was again recorded for law, the indexes for both economics (0.80) and humanities (0.30) rose substantially, to levels higher than the index reached by the engineering department (0.20). Likewise, economics (4,272), engineering (2,384) and law (2,352) showed the highest database usage, with clearly larger figures than all other departments.

Humanities continued to be characterized by low database usage, despite its higher index of results of scientific endeavor (0.30) than for other departments, such as physics (1.08), library and information science and documentation (-1.14), and computer science (-0.24), which used the database service more assiduously.

In 1999, as shown in Figure 3, law (2.42), humanities (0.60), and economics (0.15) again had positive CI values, while engineering (-0.38) fell substantially compared to the preceding years, with a slightly lower index value than recorded for mathematics (?0.37). Database usage followed a similar pattern, with the economics (3956), law (1,946), and engineering (1,545) departments reaching values higher than the overall yearly mean. The rest of the departments, grouped on the lower left side of the graph,

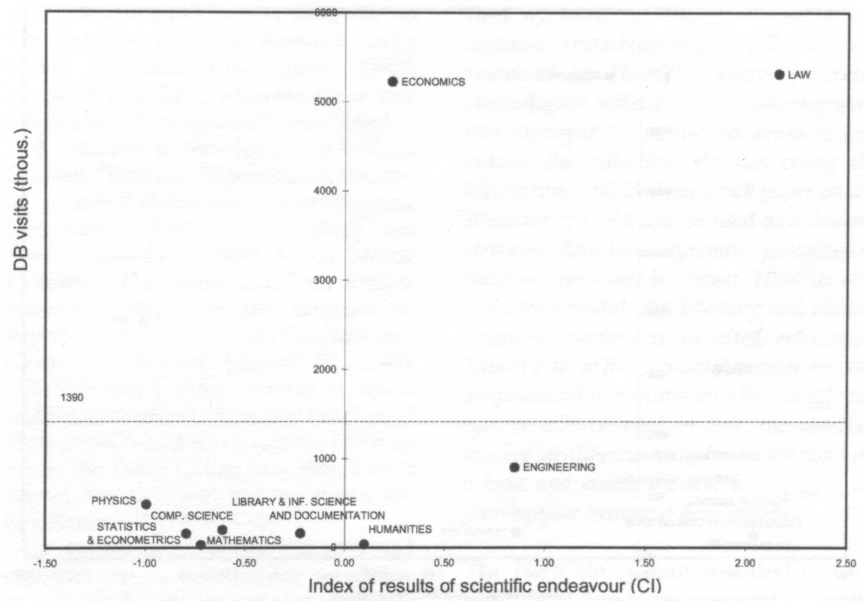


Figure 1. Scientific Process Results Index (CI) versus Database Usage—1997

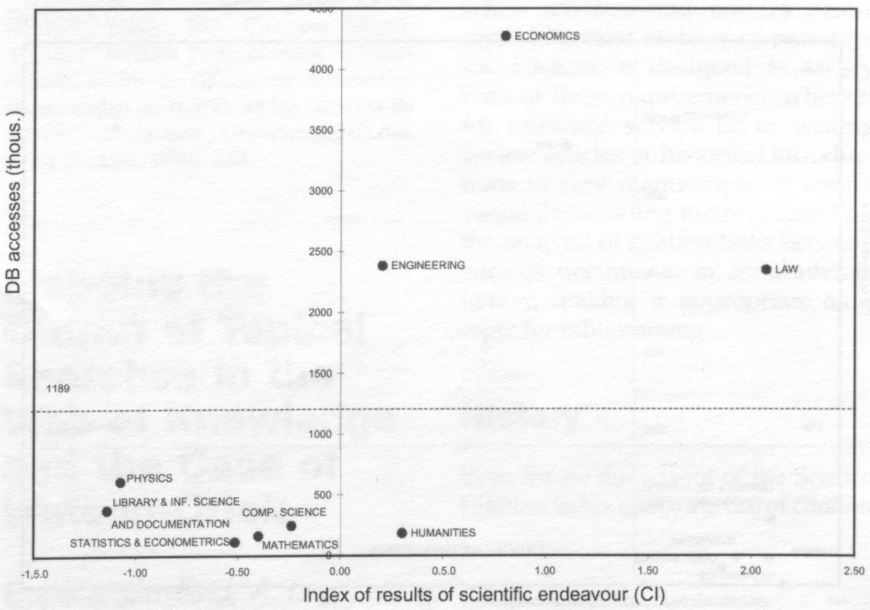


Figure 2. Scientific Process Results Index (CI) versus Database Usage—1998

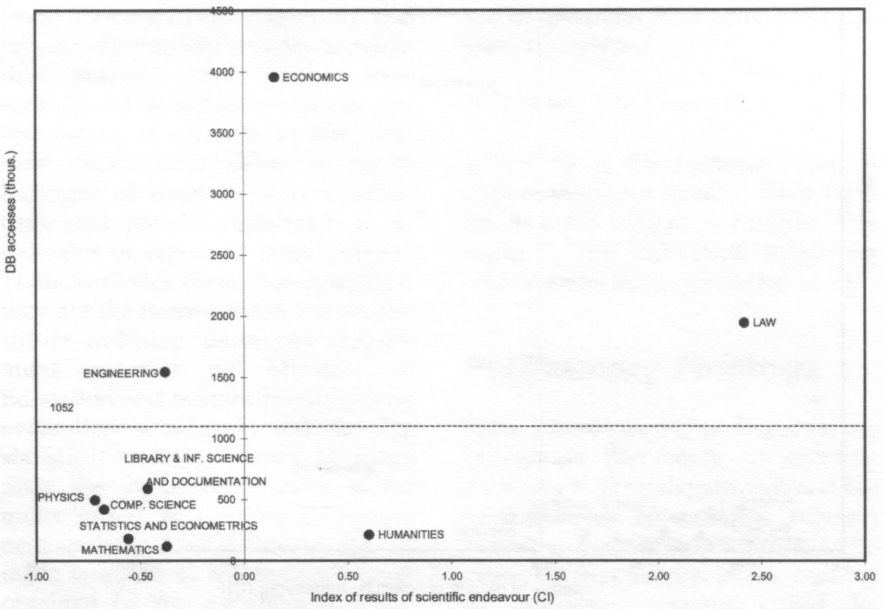


Figure 3. Scientific Process Results Index (CI) versus Database Usage—1999

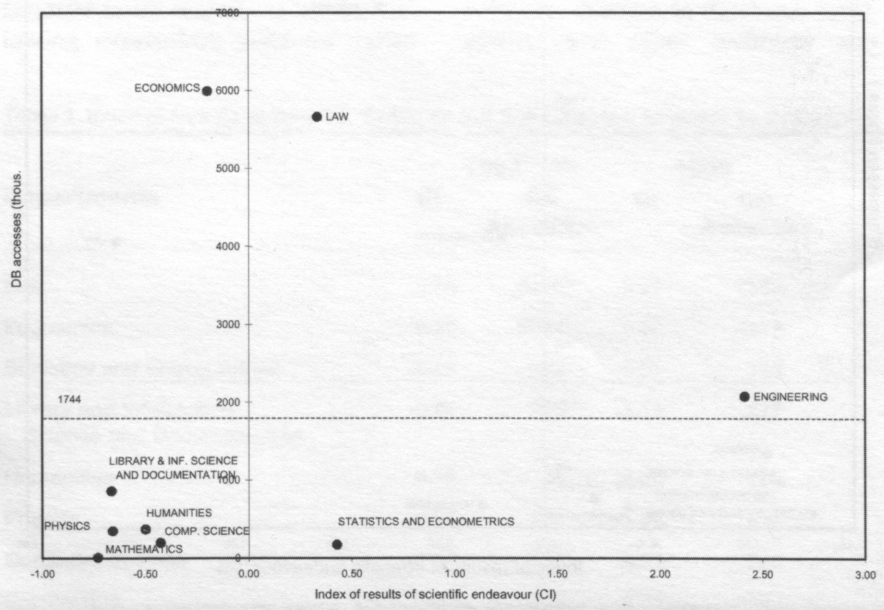


Figure 4. Scientific Process Results Index (CI) versus Database Usage—2000

had very discreet performance in both respects.

Finally, in 2000 (figure 4) a change was recorded in the index of results of scientific endeavor, with statistics and econometrics (0.43), a department that had not exhibited relevant CI values in the preceding years, climbing to a figure higher than recorded for economics (-0.20). The highest index for the year was found for engineering (2.41), a department that, along with economics and law, had a higher access count than the overall yearly mean. No significant changes were detected for the other departments.

Discussion

The findings suggest the existence of a relationship between intensity of research activity and database usage, although such a relationship does not necessarily entail dependence of any nature. The yearly analyses show that the departments with the highest indexes in terms of research results (CI) are also, by and large, the ones that used the resources available through the university library's database service most intensely. This concurs with the findings of other authors in studies on the relationship between the use, in English universities, of the information provided by *Bath Information & Data Services* (BIDS), the leading supplier of access to bibliographic database information in the United Kingdom, and university research activity as recorded in the *Research Assessment Exercise* (RAE).[8]

The exception to the general rule is the department of humanities, the behavior of which can be explained by the fact that such information is used less frequently by humanists, who show a preference for other types of data sources.[9] In contrast is research in other areas of knowledge, where the prevalence of scientific articles—both as a source of information and a vehicle for publishing results—entails intense use of bibliographic databases.[10]

By calculating an aggregate index of the results of scientific endeavor (CI), we have a useful performance indicator to explore possible relationships between the use of the information resources and research outcomes in the university environment. However, as this paper is only a report of

research in its early stages, further studies are required to confirm or reject the findings exposed.

References

1. Carol Shanklin, "A Faculty Perspective: Intrinsic Research Rewards that Make a Successful Faculty Member Tick," in *Evaluating Research Productivity* [online], (Kansas: Merrill Advanced Studies Center, University of Kansas, 2001). Accessed June 2001, http://merrill.ku.edu/publications/2001whitepaper/tableofcontents.html. Halil Dundar and Darrell R. Lewis, "Determinants of Research Productivity in Higher Education," *Research in Higher Education* 39 (1998): 607–30.
2. Carol Hughes, "Faculty Publishing Productivity: The Emerging Role of Network Connectivity," *Campus-Wide Information Systems* 16 (1999): 30–38.
3. Martha Kyrillidou, "From Input and Output Measures to Quality and Outcomes Measures, or, from the User in the Life of the Library to the Library in the Life of the User," *Journal of Academic Librarianship* 28 (2002): 42–46.
4. Valarie Zeithaml, A. Parasuraman, and N. K. Malhotra, *A Conceptual Framework for Understanding e-Services Quality: Implications for Future Research and Managerial Practice* (Cambridge, Mass.: Marketing Science Institute, 2000); Joel A. Cohen, "Computer Mediated Communication and Publication Productivity among Faculty," *Internet Research: Electronic Networking Applications and Policy 6* (1996): 41–63.
5. *Sistema de Información de Actividades de Investigación* (SINAI) online, (Madrid: Oficina de Transferencia de Resultados de Investigación, Universidad Carlos III de Madrid, 1997/1998). Accessed Oct. 8, 2002, http://otri.uc3m.es/pages/OTRI_04.htm.
6. Gobierno de España, Ministerio de la Presidencia, *Boletín Oficial del Estado* [Official State Journal], no. 289, Dec. 3, 1994.
7. George H. Dunteman, *Principal Components Analysis* (SAGE Univ. Paper Series on Quantitative Applications in the Social Sciences, Beverly Hills: SAGE, 1989).
8. Harry East, Badekale Ajibade, and Kathryn Leach, "Charging for Access to University-Wide Networked Database Services in British Academia," *Aslib Proceedings* 50 (1998): 297–307; Harry East and Badekale Ajibade, "Research Assessment Exercise and Usage of BIDS-ISI," in *The Database Resources Research Group* online (London: City Univ., 1997). Accessed July 2002, www.soi.city.ac.uk/informatics/is/drrg/rae97/rae97.html.
9. Stephen E. Wiberley Jr. and William G. Jones, "Time and Technology: A Decade-Long Look at Humanists' Use of Electronic Information Technology," *College and Research Libraries* 61 (2000): 421–31; Marcia J. Bates, "The Getty End-User Online Searching Project in the Humanities: Report No. 6: Overview and Conclusions," *College and Research Libraries* 57 (1996): 514–23; Marcia J. Bates, Deborah N. Wilde, and Susan Siegfried, "Research Practices of Humanities Scholars in an Online Environment: The Getty Online Searching Project Report, No. 3," *Library and Information Science Research* 17 (1995): 5–40.
10. Carlos García Zorita, "La actividad científica de los economistas españoles, en función del ámbito nacional o internacional de sus publicaciones: estudio comparativo basado en un análisis bibliométrico durante el período 1986–1995," (Ph.D. diss., Universidad Carlos III de Madrid, 2000), 405; Carmen Martín Moreno, "Hábitos y necesidades de información de los científicos experimentales encuadrados en el área de las ciencias de la vida," (Ph.D. diss., Universidad Carlos III de Madrid, 1999), 410.

Mapping the Output of Topical Searches in the Web of Knowledge and the Case of Watson-Crick

Eugene Garfield, A. I. Pudovkin, and V. S. Istomin

HistCite™ is a system that generates chronological maps of subject (topical) collections resulting from searches of the Institute for Scientific Information Web of Science (WoS) or Science Citation Index, Social Sciences Citation Index, and Arts and Humanities Citation Index on CD-ROM. WoS export files are created in which all cited references for source documents are captured. These bibliographic collections are processed by HistCite, which generates chronological tables as well as historiographs that highlight the most-cited works in and outside the collection. Articles citing the 1953 primordial Watson-Crick paper on the structure of DNA will be used as a demonstration. Real-time dynamic genealogical historiographs will be shown. HistCite also includes a module for detecting and editing errors or variations in cited references. Export Files of five thousand or more records are processed in minutes on a PC. Ideally the system will be used to help the searcher quickly identify the most significant work on a topic and enable the searcher to trace its year-by-year historical development.

The HistCite system resulted from a long-term-needs assessment of users of bibliographic databases. Librarians and users need to identify the key works on a particular subject, while scholars and editors desire rapid historical reviews of new topics. HistCite is designed to satisfy both of these requirements, whether for reference service or in writing review articles or historical introductions to new manuscripts. It uses a visual data-mining method based on the analysis of citation links between various documents in an academic library, making it appropriate as a topic for bibliomining.

History

Even before the advent of the Science Citation Index (SCI), the use of citation

Eugene Garfield (garfield@codex.cis.upenn.edu) is Chairman Emeritus, Thomson ISI, Philadelphia. **A. I. Pudovkin** (aipud@online.ru) is Chief Scientist at the Institute of Marine Biology, Russian Academy of Sciences, Vladivostok, Russia. **V. S. Istomin** (vi@mail.wsu.edu) is a Systems Analyst at the Center for Teaching, Learning, and Technology, Washington State University.

data was discussed in the 1964 report "The Use of Citation Data in Writing the History of Science."[1] This included a historiograph that sketched the history of DNA from Gregor Mendel in 1865 to Marshall Nirenberg in 1961, through various stages including Avery-McCleod-McCarty in 1944, through Watson-Crick in 1953. Flow charts of the papers were created manually, based solely on the references cited in a set of core source papers identified in a book by Isaac Asimov on the genetic code. This gave rise to speculation on the potential use of citation indexes for historiography.[2] Similar maps were later created by Tony Cawkell.[3]

In an information retrieval course taught at the University of Pennsylvania Moore School of Electrical Engineering, students were required to create similar topical historiographs. It was believed that these historiographs would aid in studying the contemporary history of science. Since history and bibliography were intimately linked, the term "historiobibliography" was coined.[4]

A frequent topic of discussion during the DNA mapping project was the idea of writing computer programs that would create such maps directly from the electronic files of SCI. There was a possibility that this would require random access to ISI's massive files so that cited and citing documents could be retrieved in real time. In the 1960s, however, low-cost gigabyte memories were still a dream. The implementation of real-time mapping had to wait for the time when computer memories were large and cheap enough to handle retrospective files covering many decades of literature. Though the PC had not yet come along, online searches were possible in the 1970s; mapping in real time, however, was still not feasible. Only very recently, when PCs could handle the output of a completely linked large file of thousands of records, did the creation of historiographs in real time become feasible.

There have been many different types of mapping exercises performed on a relatively small scale. In the past, co-citation clustering required mainframe computers and, in most cases, still does.[5] These ideas were later extended to creating small cluster maps online as in the SciMap system developed by Small at ISI. In that system, a starting paper is used to seed the creation of a co-citation cluster map.[6] In spite of the many mapping and visualization reports in the literature, none were applied to the creation of historiographs. Further, none of the many authors on co-citation mapping considered the potential significant relationship between historical displays and the need of reference librarians and users to evaluate the output of literature searches with such sources as SCI, Medline, or Chemical Abstracts.

Until quite recently, the authors thought of creating historiographs primarily by seeding one or two primordial papers. SCI on CD-ROM was used to trace forward in time papers that had cited the starting papers. This is the essence of the now traditional cited-reference search. Since the basic purpose of a historiograph is to display the chronological development of a topic or field year by year from the earliest papers to the most recent, the annual SCI was searched in the same way. Once the initial group of citing papers was identified, further cited-reference searches were done on them, a process sometimes called citation chaining. This process was iterated for as many years of the literature as was necessary, as will be illustrated later in this paper in discussing of the work of Watson and Crick.

HistCite Program

In HistCite, two types of frequency are distinguished—local citation score (LCS) and global citation score (GCS). LCS (or frequency) is the number of times an item is cited in the retrieved collection. The GCS is the global score, the number of cites in the entire SCI/ Social Sciences Citation Index (SSCI). The record for each source document contains both its LCS and GCS. Once HistCite sorts the papers by citation score, the user will select a group cited above an arbitrarily chosen threshold to be mapped. If there are five hundred source papers, then a 5 percent selection threshold would produce twenty-five core papers. These core papers should be of prime interest especially to a searcher who is not familiar with the subject matter. Ordinarily, one would examine this core list first in reviewing a new topic. The coordinates of these papers are used to create a historiograph of the topic that displays the papers and their citation links chronologically.

Identifying Core Literatures

The authors initially assumed that the search would begin with one primordial paper. However, it became apparent that groups of papers could be fed in by one or more authors—and by extension, larger clusters of papers by institution or by key word. Thus, the output of any conventional search or a combination of citation and key word searches could be input to the system. Once the input bibliography is created, the core papers on the topic can be rapidly identified.

Outer References

HistCite also produces frequency-ranked tables of outer references, that is, cited papers and books that fall outside the retrieved original collection. These are works that do not turn up in the original WoS search but, significantly, are cited frequently in the papers that are retrieved. The searcher can examine these candidate references and decide whether to add them to the initial collection.

Watson-Crick 1953 DNA Paper

The following will illustrate the use of HistCite in a truly historical mapping exercise. This year marks the fiftieth anniversary of the Watson-Crick discovery of the double helix structure of DNA.[7] That 1953 paper was used to conduct a cited reference of SCI. The result of mapping the five years from 1953 to 1958 can be seen in figures 1 through 4.

Figure 1 shows the usual HistCite table for the papers that cite the 1953 Watson-Crick paper, with the addition of a few of the key outer references for Avery and other papers.[8]

Figure 2 represents the table for the 975 papers retrieved by virtue of chaining citations to the 210 papers that cited Watson-Crick. In other words, these are second-generation citations to the citing papers in figure 1.

In Figure 3, the year-by-year map of the twenty-two most-cited papers in the chained indexed file can be seen. Notice that there were nine highly cited papers in 1953. And in 1954, there are five. Using the typical reference citation—that is, only author, volume, page, and year—it is not possible to differentiate the month-by-month progression after the Watson-Crick paper of April 1953. However, the HistCite system can take into account the cover dates of the journals involved if they are included.

In a recent interview, James Watson regretted that he had not cited Avery et al. in the 1953 Nature paper.[9] However, this omission made little difference in the HistCite algorithmic mapping exercise. To emphasize that link, a dotted line has been inserted linking node #2 to node #27.

Figure 4 illustrates how the historiograph changes not just year-to-year but month-by-month. Unfortunately, the Institute for Scientific Information Web of Science (WoS) does not contain cover dates until 1985, so it was necessary to manually insert in the export files the cover dates for the few dozen papers involved in this example.

Outer References Missing Links? Journal list All-Author list Citation Matrix Graphs HistCite Guide

Articles from 1953-1958 citing Watson and Crick's 1953 paper, "Molecular Structure of DNA" and selected outer references

Nodes: 210
Sorted by **year, journal, volume, page**.
Page **1**: 1

#	Cited nodes	Nodes / Authors	GCS	LCS
1	0	1 1944 JOURNAL OF EXPERIMENTAL MEDICINE 79():137-157 **AVERY OT; MACLEON CM; MCCARTY M** *Studies on the Chemical Nature of the Substance Inducing Transformation of Pseumococcal Types. Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III*	0	23
2	0	2 1952 JOURNAL OF GENERAL PHYSIOLOGY 36(1):39-56 **HERSHEY AD; CHASE M** *Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage*	747	23
3	2	3 1953 ACTA CRYSTALLOGRAPHICA 6(8-9):673-677 **FRANKLIN RE; GOSLING RG** *The Structure of Sodium Thymonucleate Fibres .1. The Influence of Water Content*	14	11
4	3	4 1953 ACTA CRYSTALLOGRAPHICA 6(8-9):678-685 **FRANKLIN RE; GOSLING RG** *The Structure of Sodium Thymonucleate Fibres .2. The Cylindrically Symmetrical Patterson Function*	10	8
5	1	5 1953 ARCHIVES OF BIOCHEMISTRY AND BIOPHYSICS 46(1):12-17 **SMITH CL** *The Breakdown of Desoxyribonucleic Acid Under Deuteron and Electron Bombardment*	5	1
6	2	6 1953 BIOCHEMICAL JOURNAL 55(5):774-782 **WYATT GR; COHEN SS** *The Bases of the Nucleic Acids of Some Bacterial and Animal Viruses - The Occurrence of 5-Hydroxymethylcytosine*	57	8
7	3	7 1953 COLD SPRING HARBOR SYMPOSIA ON QUANTITATIVE BIOLOGY 18():123-131 **WATSON JD; CRICK FHC** *The Structure of Dna*	61	21
8	1	8 1953 COLD SPRING HARBOR SYMPOSIA ON QUANTITATIVE BIOLOGY 8():133-134 **WYATT GR** *The Quantitative Composition of Deoxypentose Nucleic Acids As Related To the Newly Proposed Structure*	9	4
9	2	9 1953 COLD SPRING HARBOR SYMPOSIA ON QUANTITATIVE BIOLOGY 18():171-183 **LARK KG; ADAMS MH** *The Stability of Phages As a Function of the Ionic*	13	2

Figure 1. HistCite Chronological Listing of Papers Citing Watson-Crick 1953

Conclusion

The authors have described a tool that permits the user to manage the voluminous references produced in a

comprehensive search of the literature. For those who are new to the subject, the mere juxtaposition of the most-cited papers for each five- or ten-year period of the literature will help identify the key literature to be used first. For those who are knowledgeable in the field, the system will help jog the memory to recall the key works that were associated with the development of the field. While the relevance of citing works may be apparent, the collective bibliographic coupling and co-citation of papers in and outside of the basic bibliography should provide a comprehensive structure for completing a synoptic history of the topic.

This tool will prove extremely valuable for reference librarians in academic libraries and those producing bibliographic instruction on an unfamiliar topic. The librarian can use HistCite to produce a graphical view of citation data and can aid a patron in the identification of critical works regarding a topic. Collection development librarians can use the tool to look at the importance of articles from specific journals in order to make purchasing decisions. In addition, librarians doing their own research can use HistCite to understand the history of a research question and to aid in the identification of appropriate work for traditional evidence-based librarianship.

For more information on algorithmic historiography, a paper on this theme has been published in a special issue of the Journal of the American Society for Information Science and Technology on "Visualization of Scientific Paradigms."[10]

To learn more about HistCite and see other examples, visit http://garfield.library.upenn.edu/histcomp.

Outer References Missing Links? Journal list All-Author list Citation Matrix Graphs HistCite Guide

Articles citing Watson and Crick's 1953 paper, "Molecular Structure of DNA", the articles citing them (1953-1958), and selected outer references

Nodes: 975
Sorted by **year, journal, volume, page**.
Page **1**: 1 2

#	Cited nodes	Nodes / Authors	GCS	LCS
1	0	1 1938 JOURNAL OF BIOLOGICAL CHEMISTRY 124():425- **SEVAG MG** *[unknown]*	216	37
2	1	2 1944 JOURNAL OF EXPERIMENTAL MEDICINE 79():137-157 **AVERY OT; MACLEON CM; MCCARTY M** *Studies on the Chemical Nature of the Substance Inducing Transformation of Pseumococcal Types. Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III*	331	43
3	0	3 1945 JOURNAL OF BIOLOGICAL CHEMISTRY 161(1):83-89 **SCHMIDT G; THANNHAUSER SJ** *A Method for the Determination of Desoxyribonucleic Acid, Ribonucleic Acid, and Phosphoproteins in Animal Tissues*	696	34
4	1	4 1945 JOURNAL OF BIOLOGICAL CHEMISTRY 161(1):293-303 **SCHNEIDER WC** *Phosphorus Compounds in Animal Tissues .1. Extraction and Estimation of Desoxypentose Nucleic Acid and of Pentose Nucleic Acid*	952	30
5	2	5 1946 JOURNAL OF GENERAL PHYSIOLOGY 30(2):117-& **MIRSKY AE; POLLISTER AW** *Chromosin, a Desoxyribose Nucleoprotein Complex of the Cell Nucleus*	323	35
6	0	6 1947 JOURNAL OF THE CHEMICAL SOCIETY (SEP):1131-1141 **GULLAND JM; JORDAN DO; TAYLOR HFW** *Deoxypentose Nucleic Acids .2. Electrometric Titration of the Acidic and the Basic Groups of the Deoxypentose Nucleic Acid of Calf Thymus*	70	31
7	3	7 1951 BIOCHEMICAL JOURNAL 48(5):584-590 **WYATT GR** *The Purine and Pyrimidine Composition of Deoxypentose Nucleic Acids*	276	63
8	0	8 1951 JOURNAL OF BIOLOGICAL CHEMISTRY 189(2):597-605 **MARSHAK A; VOGEL HJ** *Microdetermination of Purines and Pyrimidines in Biological Materials*	136	30
9	0	9 1951 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA 37(4):205-211 **PAULING L; COREY RB; BRANSON HR** *The Structure of Proteins - 2 Hydrogen-Bonded Helical Configurations of the Polypeptide Chain*	185	26
10	1	10 1952 BIOCHEMICAL JOURNAL 52(5):558-565 **MARKHAM R; SMITH JD** *The Structure of Ribonucleic Acids .2. The Smaller Products of Ribonuclease Digestion*	104	28
11	0	11 1952 JOURNAL OF GENERAL PHYSIOLOGY 36(1):39-56 **HERSHEY AD; CHASE M** *Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage*	206	

Figure 2. Opening Page of HistCite Collection of "Chained" Citations to Watson-Crick 1953

References

1. E. Garfield, I. H. Sher, and R. J. Torpie, The Use of Citation Data in Writing the History of Science: Report of Research for Air Force Office of Scientific Research Under Contract AF49 (638)-1256 (Philadelphia: The Institute for Scientific Information, 1964). Accessed Oct. 9, 2003, www.garfield.library.upenn.edu/paper/useofcitdatawritinghistofsci.pdf.

2. E. Garfield, "Citation Indexing, Historio-bibliography and the Sociology of Science Biography," Current Contents M25+, 1971. First published in K. E. Davis and W. D. Sweeney, eds., *Proceedings of the Third International Congress of Medical Librarianship* (Amsterdam: Excerpta Medica, 1969), 187–204. Accessed Oct. 9, 2003, www.garfield.library.upenn.edu/essays/V1p158y1962-73.pdf.

3. A. E. Cawkell, "Acoustic Journals and Acoustic Research Articles," *Current*

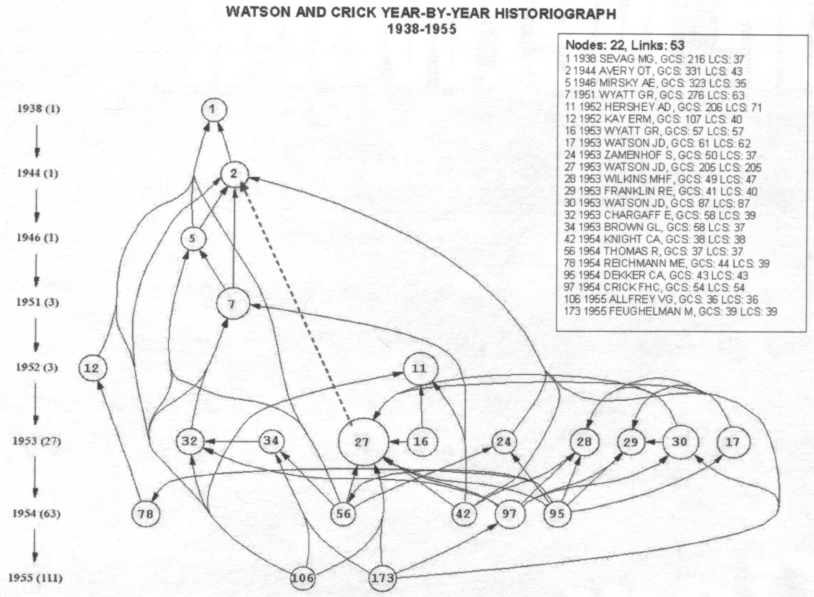


Figure 3. Year-by-Year Map of Watson-Crick 1938–1955

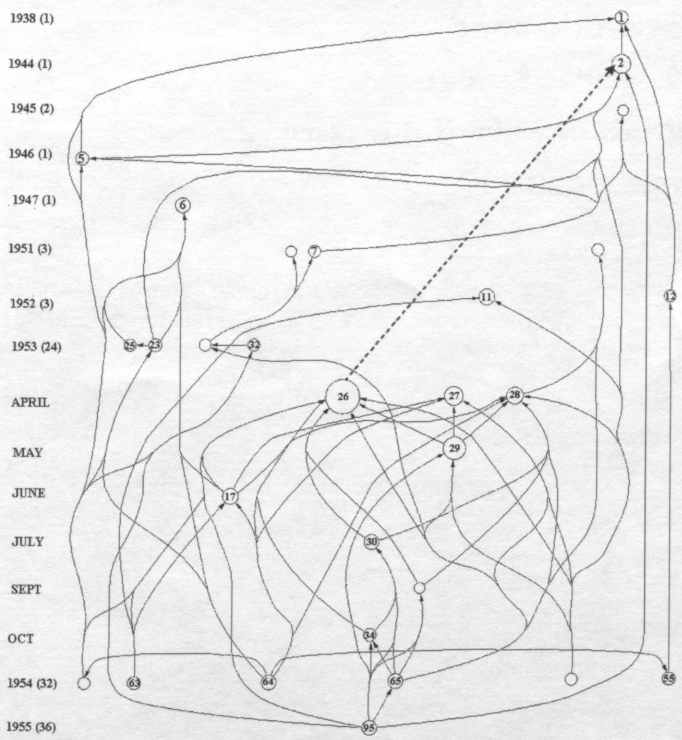


Figure 4. Month-by-Month Historiograph Linking Watson-Crick to Avery

Contents 44 (1989): 4–15. Reprinted in E. Garfield, *Essays of an Information Scientist*, Vol. 12 (Philadelphia: ISI Pr., 1977). Accessed Oct. 9, 2003, www.garfield.library.upenn.edu/essays/v12p301y1989.pdf. A. E. Cawkell, 2000. "Visualizing Citation Connections," in *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*, eds. B. Cronin and H. Barsky Atkins, (Medford, N.J.: Information Today, Inc.), 177–94.

4. E. Garfield, "Citation Indexing, Historio-Bibliography and the Sociology of Science Biography," *Current Contents 15 (1971)*, M25+.

5. E. Garfield, "Contract Research Services at ISI—Citation Analysis for Governmental, Industrial, and Academic Clients," *Current Contents* 23 (1992): 5–13. Accessed Oct. 9, 2003, http://garfield.library.upenn.edu/essays/v15p075y1992-93.pdf. H. Small and E. Garfield, "The Geography of Science: Disciplinary and National Mappings," *Journal of Information Science* 11 (1985): 147–59. Reprinted in *Current Contents* 43 (Oct. 27, 1986): 3–14; *Essays of an Information Scientist*, vol. 9 (Philadelphia: ISI Pr.) 324-35. Accessed Oct. 9, 2003, www.garfield.library.upenn.edu/essays/v9p325y1986.pdf.

6. H. Small, "A Sci-Map Case Study: Building a Map of AIDS Research," *Scientometrics* 30 (1994): 229–41. H. Small, E. Sweeney, and E. Greenlee, 1985. "Clustering the Science Citation Index Using Co-citations: 2. Mapping Science," *Scientometrics* 8 (1985): 321–34.

7. J. D. Watson and F. H. C. Crick, 1953. "Molecular Structure of Nucleic Acids—A Structure for Deoxyribose Nucleic Acid," *Nature* 171, no. 4356 (1953): 737–38.

8. O. T. Avery, C. M. Macleon, and M. McCarty, 1944. "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal types: Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III," *Journal of Experimental Medicine* 79 (1944): 137–57.

9. Anonymous, "Genes, Girls and Honest Jim," *Bio-IT World* 2 no. 4 (2003): 28.

10. E. Garfield, A. I. Pudovkin, and V.S. Istomin, "Why Do We Need Algorithmic Historiography?" *Journal of the American Society for Information Science and Technology* 54, no. 5 (2003): 400–12. Accessed Oct. 9, 2003, http://garfield.library.upenn.edu/papers/jasist54(5)400y2003.pdf.

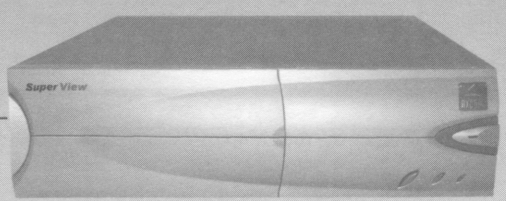
Share Videos & CDs... Over Your Network

Load your VHS and DVD videos or reference CDs on an Axonix server.
Then share them over your network with easy web browser access. That's it. That's all.

Save time. Eliminate the hassle of tracking CDs, discs and tapes. Fast content access via LAN, WAN or the Internet. No tech support or training needed. Easy to use.

Save money. No more costly TVs, VCRs, DVD players, or multiple copies. Scrap old CD-ROM towers and reduce telecom costs. Very affordable. Nothing extra to add or buy.

SuperView™



Video Server

- **Full motion, full screen videos on demand**
- **Store and serve hundreds of hours of video**
- **Uses one-tenth the bandwidth at one-third the price**
- **Serve hundreds of users simultaneously**

SuperCD™



CD-ROM Server

- **Store and serve 1 to 512 CDs/DVDs**
- **Automatic user license monitoring**
- **Thousands of units running 24/7**
- **Award winning network appliance**

Call for a free, no-risk trial

800.866.9797
www.axonix.com

AXONIX®
IT Solutions Since 1983

Information Technology and Libraries Index to Volume 22 (2003)

Compiled by
Charles Anderson

Academic Libraries.
An Architecture for Behavior-Based Library Recommender Systems, Geyer-Schulz, Andreas, Neumann, Andreas, and Thede, Anke, 165–74.
Building a New Infrastructure for Digital Media: Northwestern University Library, Stewart, M. Claire, and Cervone, H. Frank, 69–74.
HERMES: The Hopkins Electronic Resource Management System, Robertson, Nathan D. M., and Cyzyk, Mark, 12–17.
Implementing the SFX Link Server at the University of Iowa (Communication), Soderdahl, Paul A., 117–19.
Study of the Use of the Carlos III University of Madrid Library's Online Database Service in Scientific Endeavor (Communication), Suárez-Balseiro, Carlos A., Iribarren-Maestro, Isabel, and Casado, Elias Sanz, 179–83.
Traces in the Clickstream: Early Work on a Management Information Repository at the University of Pennsylvania (Communication), Zucca, Joe, 175–79.
An Architecture for Behavior-Based Library Recommender Systems, Geyer-Schulz, Andreas, Neumann, Andreas, and Thede, Anke, 165–74.
Art History Education.
Improving Art History Education: Library and Faculty Partnerships in Instructional Technology Development (Communication), Dirst, Tara L., 83–87.
Arts Instruction in the Age of Technology: Providing Library Services to Support Studio and Survey Faculty Who Use Technology for Instruction (Communication), Gerlich, Bella Karr, and Perrier, Amy, 79–83.
Bibliographic Instruction.
Developing the Online Learning Environment: The Pros and Cons of Using WebCT for Library Instruction (Communication), Kraemer, Elizabeth W., 87–92.
Bibliomining.
The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making, Nicholson, Scott, 146–51.
Introduction to This Special Issue on the Bibliomining Process, Nicholson, Scott, 144.
Mining User Communities in Digital Libraries, Papatheodorou, Christos, Kapidakis, Sarantos, Sfakakis, Michalis, and Vassiliou, Alexandra, 152–57.
The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making, Nicholson, Scott, 146–51.
Bills, Linda, Cheng, Rachel J., and Nathanson, Alan J. *Subject Web Page Management without HTML Coding: Two Approaches,* 4–11.
Book Reviews.
Human Needs and the New Computing Technologies, by Shneiderman, Ben, 139–40.
Mastering the Language of Web Design, by Meyer, Eric A., 93–94.
New Tools and Possibilities for Cooperative Resource Description, Calhoun, Karen, and Reimer, John J., eds., 44–47.
Braziunas, Tom, Mundell, Jacqueline, and Celene-Martel, Coryl. *An Organizational Model for Instructional Support at a Community College,* 61–67.
Building a New Infrastructure for Digital Media: Northwestern University Library, Stewart, M. Claire, and Cervone, H. Frank, 69–74.
Building an Internet Gateway, Davies, Ron, 106–109.
Carter, Howard and Rundblad, Kevin. *The Village of ISS: Providing Library-Based Instructional Support,* 54–60.
Casado, Elias Sanz, Suárez-Balseiro, Carlos A., and Iribarren-Maestro, Isabel. *Study of the Use of the Carlos III University of Madrid Library's Online Database Service in Scientific Endeavor* (Communication), 179–83.
Catalogers.
The Impact of Information Technology on Job Requirements and Qualifications for Catalogers, Khurshid, Zahiruddin, 18–21.
Celene-Martel, Coryl, Mundell, Jacqueline, and Braziunas, Tom. *An Organizational Model for Instructional Support at a Community College,* 61–67.
Cervone, H. Frank, and Stewart, M. Claire. *Building a New Infrastructure for Digital Media: Northwestern University Library,* 69–74.
The Changing Role of Libraries in Instructional Support (Editorial), Logue, Susan, 52.
Cheng, Rachel J., Bills, Linda, and Nathanson, Alan J. *Subject Web Page Management without HTML Coding: Two Approaches,* 4–11.
Childers, Scott. *Computer Literacy: Necessity or Buzzword?* 100–104.
Community College Libraries.
An Organizational Model for Instructional Support at a Community College, Mundell, Jacqueline, Celene-Martel, Coryl, and Braziunas, Tom, 61–67.
Computer Literacy: Necessity or Buzzword? Childers, Scott, 100–104.
Cooperative Online Resource Catalog.
New Tools and Possibilities for Cooperative Resource Description (Book Review), Calhoun, Karen, and Reimer, John J., eds., 44–47.
Cyzyk, Mark, and Robertson, Nathan D. M. *HERMES: The Hopkins Electronic Resource Management System,* 12–17.
Databases.
Implementing the SFX Link Server at the University of Iowa (Communication), Soderdahl, Paul A., 117–19.
Data Warehousing.
The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making, Nicholson, Scott, 146–51.
Davies, Ron. *Building an Internet Gateway,* 106–109.
Developing the Online Learning Environment: The Pros and Cons of Using WebCT for Library Instruction (Communication), Kraemer, Elizabeth W., 87–92.

The Development of the NISO Committee AX's OpenUrl Standard (Communication), Hendricks, Arthur, 129–33.
Dewald, Nancy H., and Shank, John D. *Establishing Our Presence in Courseware: Adding Library Services to the Virtual Classroom* (Communication), 38–43.
Digital Divide in India: Need for Correcting Urban Bias (Communication), Parvathamma, N., 35–38.
Digital Libraries.
Mining User Communities in Digital Libraries, Papatheodorou, Christos, Kapidakis, Sarantos, Sfakakis, Michalis, and Vassiliou, Alexandra, 152–57.
Digital *Media.*
Building a New Infrastructure for Digital Media: Northwestern University Library, Stewart, M. Claire, and Cervone, H. Frank, 69–74.
Dirst, Tara L. *Improving Art History Education: Library and Faculty Partnerships in Instructional Technology Development* (Communication), 83–87.
Ensor, Pat. *President's Column*, 50, 60.
Establishing Our Presence in Courseware: Adding Library Services to the Virtual Classroom (Communication), Shank, John D., and Dewald, Nancy H., 38–43.
An Evaluation of Computer-Supported Collaborative Serial Management: A Case Study, Zhang, Ying, and Smulewitz, Gracemary C., 110–15.
Garfield, Eugene, Pudovkin, A. I., and Istomin, V. S. *Mapping the Output of Topical Searches in the Web of Knowledge and the Case of Watson-Crick* (Communication), 183–87.
Gelernter, Judith. *TECHXNY June 2002 and Implications for Libraries* (Communication), 32–35.
Geyer-Schulz, Andreas, Neumann, Andreas, and Thede, Anke. *An Architecture for Behavior-Based Library Recommender Systems*, 165–74.
Hendricks, Arthur. *The Development of the NISO Committee AX's OpenUrl Standard* (Communication), 129–33.
HERMES: The Hopkins Electronic Resource Management System, Robertson, Nathan D. M., and Cyzyk, Mark, 12–17.
A History of Web Portals and Their Development in Libraries (Communication), Zhou, Joe, 119–28.
Human Needs and the New Computing Technologies (Book Review), by Shneiderman, Ben, 139–40.
The Impact of Information Technology on Job Requirements and Qualifications for Catalogers, Khurshid, Zahiruddin, 18–21.
Implementing the SFX Link Server at the University of Iowa (Communication), Soderdahl, Paul A., 117–19.
Improving Art History Education: Library and Faculty Partnerships in Instructional Technology Development (Communication), Dirst, Tara L., 83–87.
India.
Digital Divide in India: Need for Correcting Urban Bias (Communication), Parvathamma, N., 35–38.
Instructional Services.
Arts Instruction in the Age of Technology: Providing Library Services to Support Studio and Survey Faculty Who Use Technology for Instruction (Communication), Gerlich, Bella Karr, and Perrier, Amy, 79–83.
The Changing Role of Libraries in Instructional Support (Editorial), Logue, Susan, 52.
Collaborative Digitization Projects: Opportunities to Enchance Teaching and Learning, Lim, Adriene, 75–78.
Improving Art History Education: Library and Faculty Partnerships in Instructional Technology Development (Communication), Dirst, Tara L., 83–87.
An Organizational Model for Instructional Support at a Community College, Mundell, Jacqueline, Celene-Martel, Coryl, and Braziunas, Tom, 61–67.
The Village of ISS: Providing Library-Based Instructional Support, Carter, Howard and Rundblad, Kevin, 54–60.
Introduction to This Special Issue on the Bibliomining Process, Nicholson, Scott, 144.
Iribarren-Maestro, Isabel, Casado, Elias Sanz, and Suárez-Balseiro, Carlos A. *Study of the Use of the Carlos III University of Madrid Library's Online Database Service in Scientific Endeavor* (Communication), 179–83.
Istomin, V. S., Garfield, Eugene, and Pudovkin, A. I. *Mapping the Output of Topical Searches in the Web of Knowledge and the Case of Watson-Crick* (Communication), 183–87.
Kapidakis, Sarantos, Sfakakis, Michalis, Vassiliou, Alexandra, and Papatheodorou, Christos. *Mining User Communities in Digital Libraries*, 152–57.
Khurshid, Zahiruddin. *The Impact of Information Technology on Job Requirements and Qualifications for Catalogers*, 18–21.
Kraemer, Elizabeth W. *Developing the Online Learning Environment: The Pros and Cons of Using WebCT for Library Instruction* (Communication), 87–92.
Library Committees.
Using Microsoft Share Point Team Services for Library Committee Management (Communication), Rao, Abhjit, 133–37.
Lim, Adriene. *Collaborative Digitization Projects: Opportunities to Enchance Teaching and Learning*, 75–78.
Link Servers.
Implementing the SFX Link Server at the University of Iowa (Communication), Soderdahl, Paul A., 117–19.
Logue, Susan. *The Changing Role of Libraries in Instructional Support* (Editorial), 52.
Mapping the Output of Topical Searches in the Web of Knowledge and the Case of Watson-Crick (Communication), Garfield, Eugene, Pudovkin, A. I., and Istomin, V. S., 183–87.
Marmion, Dan.
When Did the Rules Change (Editorial), 2.

Mastering the Language of Web Design (Book Review), by Meyer, Eric A., 93–94.
Matching Subject Portals with the Research Environment, Wormell, Irene, 158–64.
Mining User Communities in Digital Libraries, Papatheodorou, Christos, Kapidakis, Sarantos, Sfakakis, Michalis, and Vassiliou, Alexandra, 152–57.
Mundell, Jacqueline, Celene-Martel, Coryl, and Braziunas, Tom. *An Organizational Model for Instructional Support at a Community College,* 61–67.
Nathanson, Alan J., Cheng, Rachel J., and Bills, Linda. *Subject Web Page Management without HTML Coding: Two Approaches,* 4–11.
Neumann, Andreas, Thede, Anke, and Geyer-Schulz, Andreas, *An Architecture for Behavior-Based Library Recommender Systems,* 165–74.
New Tools and Possibilities for Cooperative Resource Description (Book Review), Calhoun, Karen, and Reimer, John J., eds., 44–47.
Nicholson, Scott.
The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making, 146–51.
Introduction to This Special Issue on the Bibliomining Process, 144.
Online Public Access Catalogs.
Implementing the SFX Link Server at the University of Iowa (Communication), Soderdahl, Paul A., 117–19.
OpenURL.
The Development of the NISO Committee AX's OpenUrl Standard (Communication), Hendricks, Arthur, 129–33.
An Organizational Model for Instructional Support at a Community College, Mundell, Jacqueline, Celene-Martel, Coryl, and Braziunas, Tom, 61–67.
Papatheodorou, Christos, Kapidakis, Sarantos, Sfakakis, Michalis, and Vassiliou, Alexandra. *Mining User Communities in Digital Libraries,* 152–57.
Parvathamma, N. *Digital Divide in India: Need for Correcting Urban Bias* (Communication), 35–38.
Perrier, Amy, and Gerlich, Bella Karr. *Arts Instruction in the Age of Technology: Providing Library Services to Support Studio and Survey Faculty Who Use Technology for Instruction* (Communication), 79–83.
Portals.
A History of Web Portals and Their Development in Libraries (Communication), Zhou, Joe, 119–28.
Matching Subject Portals with the Research Environment, Wormell, Irene, 158–64.
President's Column, 50, 60, 98.
Pudovkin, A. I., and Istomin, V. S., and Garfield, Eugene. *Mapping the Output of Topical Searches in the Web of Knowledge and the Case of Watson-Crick* (Communication), 183–87.
Rao, Abhjit. *Using Microsoft Share Point Team Services for Library Committee Management* (Communication), 133–37.
Recommender Systems.
An Architecture for Behavior-Based Library Recommender Systems, Geyer-Schulz, Andreas, Neumann, Andreas, and Thede, Anke, 165–74.
Remote Observation Strategies for Usability Testing, Thompson, Susan M., 22–31.
Robertson, Nathan D. M., and Cyzyk, Mark. *HERMES: The Hopkins Electronic Resource Management System,* 12–17.
Rundblad, Kevin, and Carter, Howard. *The Village of ISS: Providing Library-Based Instructional Support,* 54–60.
Serials Management.
An Evaluation of Computer-Supported Collaborative Serial Management: A Case Study, Zhang, Ying, and Smulewitz, Gracemary C., 110–15.
Sfakakis, Michalis, Vassiliou, Alexandra, Papatheodorou, Christos, and Kapidakis, Sarantos. *Mining User Communities in Digital Libraries,* 152–57.
Shank, John D., and Dewald, Nancy H. *Establishing Our Presence in Courseware: Adding Library Services to the Virtual Classroom* (Communication), 38–43.
Smulewitz, Gracemary C., and Zhang, Ying. *An Evaluation of Computer-Supported Collaborative Serial Management: A Case Study,* 110–15.
Soderdahl, Paul A. *Implementing the SFX Link Server at the University of Iowa* (Communication), 117–19.
Standards.
The Development of the NISO Committee AX's OpenUrl Standard (Communication), Hendricks, Arthur, 129–33.
Stewart, M. Claire, and Cervone, H. Frank. *Building a New Infrastructure for Digital Media: Northwestern University Library,* 69–74.
Study of the Use of the Carlos III University of Madrid Library's Online Database Service in Scientific Endeavor (Communication), Suárez-Balseiro, Carlos A., Iribarren-Maestro, Isabel, and Casado, Elias Sanz, 179–83.
Suárez-Balseiro, Carlos A., Iribarren-Maestro, Isabel, and Casado, Elias Sanz. *Study of the Use of the Carlos III University of Madrid Library's Online Database Service in Scientific Endeavor* (Communication), 179–83.
Subject Web Page Management without HTML Coding: Two Approaches, Bills, Linda, Cheng, Rachel J., and Nathanson, Alan J., 4–11.
Technological Change.
Human Needs and the New Computing Technologies (Book Review), by Shneiderman, Ben, 139–40.
TECHXNY June 2002 and Implications for Libraries (Communication), Gelernter, Judith, 32–35.
Thede, Anke, Geyer-Schulz, Andreas, and Neumann, Andreas. *An Architecture for Behavior-Based Library Recommender Systems,* 165–74.

Thompson, Susan M. *Remote Observation Strategies for Usability Testing*, 22–31.
Traces in the Clickstream: Early Work on a Management Information Repository at the University of Pennsylvania (Communication), Zucca, Joe, 175–79.
Usability Testing.
Remote Observation Strategies for Usability Testing, Thompson, Susan M., 22–31.
Using Microsoft Share Point Team Services for Library Committee Management (Communication), Rao, Abhjit, 133–37.
Vassiliou, Alexandra, Papatheodorou, Christos, Kapidakis, Sarantos, and Sfakakis, Michalis. *Mining User Communities in Digital Libraries*, 152–57.
The Village of ISS: Providing Library-Based Instructional Support, Carter, Howard and Rundblad, Kevin, 54–60.
Web.
The Development of the NISO Committee AX's OpenUrl Standard (Communication), Hendricks, Arthur, 129–33.
Web Applications.
Building an Internet Gateway, Davies, Ron, 106–109.
HERMES: The Hopkins Electronic Resource Management System, Robertson, Nathan D. M., and Cyzyk, Mark, 12–17.
Web Pages—Design.
Mastering the Language of Web Design, by Meyer, Eric, 93–94.
Remote Observation Strategies for Usability Testing, Thompson, Susan M., 22–31.
Subject Web Page Management without HTML Coding: Two Approaches, Bills, Linda, Cheng, Rachel J., and Nathanson, Alan J., 4–11.
Web Portals.
A History of Web Portals and Their Development in Libraries (Communication), Zhou, Joe, 119–28.
Matching Subject Portals with the Research Environment, Wormell, Irene, 158–64.
When Did the Rules Change (Editorial), Marmion, Dan, 2.
Wilson, Tom. *President's Column*, 98, 115..
Wormell, Irene. *Matching Subject Portals with the Research Environment*, 158–64.
Zhang, Ying, and Smulewitz, Gracemary C. *An Evaluation of Computer-Supported Collaborative Serial Management: A Case Study*, 110–15.
Zhou, Joe. *A History of Web Portals and Their Development in Libraries* (Communication), 119–28.
Zillner, Tom, ed. *Book Reviews.*
Human Needs and the New Computing Technologies, by Shneiderman, Ben, 139–40.
Mastering the Language of Web Design, by Meyer, Eric, 93–94.
New Tools and Possibilities for Cooperative Resource Description, Calhoun, Karen, and Reimer, John J., eds., 44–47.
Zucca, Joe. *Traces in the Clickstream: Early Work on a Management Information Repository at the University of Pennsylvania* (Communication), 175–79.

Index to Advertisers

Axonix	188	LITA	142, cover 3
EBSCO	151	Sea Change Corp.	164
Harris Infosource	cover 2	SIMA	174
InfoUSA	cover 4	TechBooks	141
Library Technologies	145	NDT Media/Mit Press	173

Register Now for

LITA pre-Midwinter Institutes

San Diego, California

January 9, 2004

We're bringing three of our most popular Regional Institutes to San Diego! Registration for the ALA Midwinter Meeting is not required to attend these institutes. For more information and to register, please visit www.lita.org/mw2004

XML and Libraries

This institute will cover the structure and utility of Extensible Markup Language (XML) with an emphasis on the relevance of XML to libraries. Presenter Art Rhyno is a Systems Librarian at the University of Windsor and holds an adjunct appointment at Memorial University of Newfoundland.

Proxy Web Servers & Authentication

Put web proxy servers to use for you: bandwith conservation, gathering statistics, filtering, and authentication. This institute is a combination of instruction on the theory of proxy servers, their use in library situations, and practical examples of implementations. Presenter Peter Murray is the Director of Library Information Technology Services at the University of Connecticut.

Open Source Software in Libraries

Our newest Regional Institute debuts in San Diego! Participants in this hands-on institute will learn skills enabling them to download and install GNU software and Perl modules, create and download software from a CVS repository for sharing code, evaluate open source software for usefulness and applicability, create and foster a community of developers and users, and conduct usability studies to verify a computer program's functionality. Presenter Eric Lease Morgan is the Head of the Digital Access and Information Architecture Department at the University Libraries of Notre Dame.



Bring a LITA Regional Institute to your area!

LITA Regional Institutes can be licensed or co-sponsored at your library or organization. Visit www.lita.org/institut or call Rob Carlson at 800.545.2433 x 4268 for more information.

"In a library such as ours... ReferenceUSA is among the most in-demand databases because of its scope and accuracy."

John Ganly
Assistant Director
New York Public Library

Our reference information in print, on CD-ROM and via the internet gets used. Why? Because everybody needs our reference products for everyday life. For example:

- CONSUMERS
- PROFESSORS
- SALES PEOPLE
- LAW LIBRARIES
- ENTREPRENEURS
- STUDENTS
- JOB SEEKERS
- NEWS ORGANIZATIONS

*info*USA Inc.'s databases contain company name, address, phone number, SIC/NAICS codes, estimated sales volume, number of employees, credit rating code*, key executive names, and more. Virtually all U.S. businesses - public and private.

Internet Access – *Reference*USA® Enjoy real-time access to the world's finest database of 14 million businesses and 250 million U.S. residents with comprehensive coverage, accuracy and instant access to information. And that's just the beginning! With its power, speed and user-friendly interface, *Reference*USA becomes the most effective, easy-to-use reference tool available. Customized reports, immediate downloads and business profiles are a snap!

American Business Disc® is the ultimate reference tool for Libraries. Imagine having over 14 million businesses on a single CD-ROM! Search by Company Name, Address, SIC Code(s), Yellow Page Heading, Sales Volume, Number of Employees, Geographically by ZIP Code(s), City(ies), State(s), or even the entire U.S.

PowerFinder® contains more than 115 million U.S. businesses and residential listings, 800 numbers, fax numbers Web site URLs and census information. This set of six regional CD-ROMs gives you instant access to virtually every business and nearly every resident in the United States.

American Manufacturers Directory & CD-ROM contain every manufacturer in the U.S. with 20 or more employees, more than 168,000 total listings – all in two volumes. The CD-ROM contains all 645,000 manufacturers.

Big Businesses Directory & CD-ROM include America's largest companies, 193,000 companies that employ 100 or more people. This directory also includes 703,000 top Executives and Directors.

State Business Directories are the most popular directories in every reference department. Before your patrons research any other source, point them to our publications. Four easy-to-use sections include: Businesses by City, Businesses by Yellow Page Category, Major Employers, and Manufacturers by City and Product.

"In a library such as the New York Public Library-Science Industry and Business Library which serves over a thousand users each day, a resource such as ReferenceUSA is among the most in-demand databases because of its scope and accuracy."

John Ganly
Assistant Director
New York Public Library

ReferenceUSA℠
A division of *info*USA

For More Information Call Toll-Free:
(866) 813-3009
or E-mail: reference@infoUSA.com

Friends of Libraries U.S.A.

5711 South 86th Circle • P.O. Box 27347 • Omaha, Nebraska 68127
Phone: (402) 593-4523 • Fax: (402) 596-7688 • E-Mail: reference@infoUSA.com

** Our Credit Rating Codes are indicators of probable ability to pay. They are based on business demographic factors such as number of employees, years in business, industry stability, bill paying history, barriers to entry, and government data. We recommend that these ratings be used primarily as a starting point and should not be the sole factor used in making a credit decision. You must obtain more information from bank and trade references, local credit bureaus, or other sources before extending credit. We will not be liable for any losses resulting from the use of this information.*

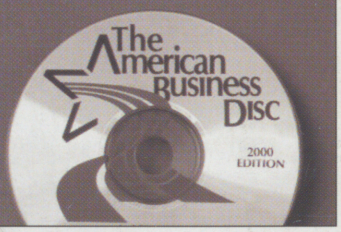
Internet Access

ReferenceUSA-online database



In Print and on CD-ROM

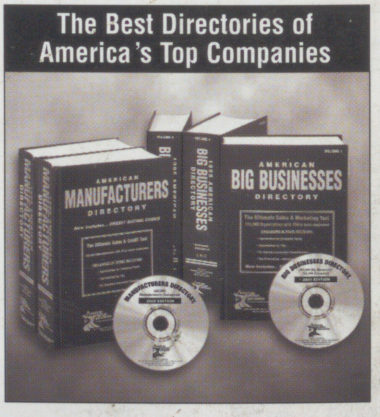
Over 14 Million Businesses on CD-ROM



More Than 115 Million U.S. Business & Residential Listings



The Best Directories of America's Top Companies



Available for All 50 States



#23LJ3(b/w)