Please cite as:
Nicholson, S. (2003). Avoiding the Great Data-Wipe of Ought-Three. *American Libraries, 34*(9), p. 36.

Avoiding the Great Data-Wipe of Ought-Three: Maintaining an Institutional Record for Library Decision-Making in Threatening Times

By Dr. Scott Nicholson, MLIS; Assistant Professor, Syracuse University School of Information Studies

Because of the USA PATRIOT Act and similar legislation that allows the government to track the actions of individuals suspected of terrorist activities, many librarians are concerned about protecting information about library use at any cost. Some propose that the solution is to delete all data from the operational databases whenever possible; in fact, a recent New York Times article discusses daily shredding of library records from the Santa Cruz Public Library System ("Librarians Use Shredder to Show Opposition to New F.B.I. Powers", Apr. 7th, 2003). However, deleting all data associated with library transactions will make data-based evaluation and justification of library services difficult; therefore, libraries must seek a balance between protecting the privacy of patrons and maintaining a history of library transactions.

Why Keep Any Transaction-Level Data?

Many librarians are not concerned about the loss of transaction-level data because they do not use this data in the decision-making process. Traditional evaluation techniques of library services are focused on general aggregates and averages; some suggest that once these aggregates are collected, the underlying transactional data can be purged. However, there are patterns of behavior in different user groups, masked by the aggregate measures, that can be critical in gaining a thorough understanding of library use.

These patterns can be discovered through statistical tools such as data mining, which is the discovery of patterns of use from the data in library systems. These patterns are invaluable for making decisions about library services and providing data-based justifications of these services.

None of these evaluations are possible if librarians delete all of their data-based institutional memories; shredding and deleting data are permanent solutions with disastrous long-term effects. However, there is an option from the corporate world that will allow librarians to save decision-making information from operational systems while still maintaining patron privacy before deleting the records – a data warehouse.

Creating a Data Warehouse

A data warehouse is an external database that contains a cleaned version of the operational data

reformatted for analysis. Data taken directly from operational systems contains many small mistakes and inconsistencies and is not archived appropriately for analysis; therefore, the first step is cleaning and extracting the data from the operational systems. This can be done with the help of systems staff, who can create an external database and import fields that do not contain personally identifiable information.

This extraction and cleaning process is key to protecting patron privacy. As the records are drawn from the operational systems, matches are made from various parts of the system, and the personally identifiable information is thrown away or replaced with codes. This information should never be put into the data warehouse, lest the personal information be backed up, saved, or otherwise archived. After the data warehouse is created, the original data can be deleted. The goal is to *create a data source that contains decision-making information that cannot be used to recreate the original transactional records.*

Use of a Data Warehouse

Creating the data warehouse maintains some of the data-based institutional memories of library use. One immediate benefit of the data warehouse is that the library staff can produce ad-hoc reports much more easily. The data warehouse can also be used to power a management information 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 and monitored. Finally, the warehouse can be used to discover the patterns of behavior to aid in the evaluation and justification of library services. Tools such as Pivot Tables in Excel can be used for basic bibliomining; however, statistical and data mining packages such as SAS, SPSS, Clementine, or the open source Weka suite (http://www.cs.waikato.ac.nz/~ml/weka/) can be used for advanced pattern discovery.

Example of the Data Mining Process

The trustees of a small public library have noticed that circulation was down over the last month and are reviewing the budget for the library. Since the library keeps a data warehouse, the staff can use bibliomining to look for patterns. The director discovers that circulation in all demographic categories remained the same from last month except for children. Upon further exploration of this group, the director discovers a dramatic drop during a single week.

After discovering a pattern, the director learns from the children's librarian that the local school had a book fair during that week. The director presents this data-based evidence to the trustees to explain the drop. In addition, the director has now an appropriate time for shelf reading and weeding in the children's section and can create an automated notification and reporting system if circulation for any subgroup changes significantly from the past. None of this would be possible if all transactional data were purged.

Conclusion

Libraries that delete their institutional memories will lose the ability to make decisions based on documented patterns and evidence from the past. Reacting to privacy threats by discarding this essential decision-making information is a permanent and critical mistake.

To learn more about the process of data mining in libraries, visit http://bibliomining.org.