Library Web Proxy Use Survey Results

Peter E. Murray

[Editor's note: The following article is winner of the first annual LITA/Endeavor Student Writing Award.]

Libraries face many policy and technological difficulties in providing remote access to databases, making effective use of Internet bandwidth, controlling where patrons browse on public computers, and gathering statistics on usage. Some libraries have chosen to employ proxy Web servers to solve these problems. This paper outlines the use of proxy Web servers by libraries to address these areas and documents survey results on their use in libraries.

In its most general definition, a proxy server is "[a]n intermediary server that accepts requests from clients and forwards them to other . . . servers."[1] In the general form of this definition, a proxy server may act as an intermediary for one of many Internet protocols (such as HTTP, FTP, Telnet, NNTP, and others). This paper focuses on the application of proxy Web servers in general, and specifically their use in library networks to solve library-specific problems.

There are four reasons a library may install a proxy Web server: to enable access to resources by patrons outside a library's network, to filter Web requests or responses on public stations, to conserve bandwidth and improve response time, and to gather statistics on Web usage. In order to identify why libraries had installed proxy servers and what proxy server software was used, a survey was prepared and distributed in November 2000. Additional questions were asked about the documentation provided to patrons about how to use a proxy server and any privacy statements about the disposition of proxy server log files. This paper will discuss the information gathered in the survey.

The survey was posted to the following electronic mailing lists: Web4Lib@sunsite.berkeley.edu, PACS-L@listserv.uh.edu, LITA-L@ala1.ala.org, teknoids@listserv.law.cornell.edu, SYSLIB-L@listserv.acsu.buffalo.edu, LIS-LINK@mailbase.ac.uk, and PUBLIB-NET@sunsite.berkeley.edu. E-mail messages were also sent to the attendees of the second and third LITA Regional Institutes on Web Proxy Servers and Authentication. A copy of the survey appears at the end of this paper in appendix A.

Between November 16 and December 22, 2000, eighty-five libraries responded to the call for participation in the survey. Respondents had the option to identify themselves and their institution; that information was used for follow-up information, much of which is discussed in this paper. A breakdown of library types is shown in table 1. Three of the responding libraries were not using proxy servers at this time.

Proxy for Remote Resource Access

By far the most frequent reason for libraries to use Web proxy servers is to enable off-network patrons to access vendor-provided resources. These resources are typically restricted to a particular institution's subscription by one of two methods: vendor-supplied username/password authentication or network address recognition. Although alternate methods exist for the purpose of authenticating access to resources (vendor-supplied scripts and referer-URL authentication, for example), these methods do not enjoy widespread implementation.

The problems of distributing vendor-supplied usernames and passwords to a community of users and keeping them secure are well known; such a method allows access to a resource from nearly any Web browser in the world. A single username and password supplied by the vendor can be distributed to individual patrons or posted on an internal Web site. In addition to the packet sniffing problems discussed in Cole, there is nothing inherent in this scheme that prevents the password from being given to patrons outside the institution's community.[2] The same problem exists for usernames and passwords distributed to individuals, although it is easier to identify abuses with a single user's password and cancel access for a username which has been compromised.

Alternatively, using network address recognition for authentication is very convenient for users on an institution's network because the only requirement for access to the resource is using a machine in the proper IP address range; no password is required. It is harder for unauthorized users to gain access to a resource because an unauthorized individual must be using a computer physically attached to the institution's network. This same physical requirement, though, prevents legitimate remote users from accessing the resource. Since resources are restricted to an institution's network addresses, placing a proxy server within that range of network addresses enables off-network users to appear to come from within the network to database vendors.

Seventy-two libraries responded that they use proxies for remote resource access; the breakdown of software packages is shown in table 2. EZproxy is the most popular, followed by Innovative Interfaces' proprietary Web Access Management. One library uses a combination of Apache and EZproxy and another uses Microsoft Proxy and Netscape Proxy. One library uses more than one proxy

Peter E. Murray (peter@pandc.org) is a graduate student in the MLS program, Simmons College, Boston.

server, but did not elaborate. In addition, a public library was using Web Access Management but is planning to install EZproxy; no proxy server is in place at this time.

Some proxy servers for remote resource access require users to configure their browsers to take advantage of the proxy service. Several libraries supplied URLs to documentation explaining this reconfiguration; these were the best in the author's opinion:

- Central Michigan University, http://ocls.cmich.edu/remoteindex.htm
- University of Waterloo, www.tug.uwaterloo.ca/proxy
- Tarleton State University, www.tarleton.edu/~library/proxy/instructions1.htm

An academic library noted that some patrons must be told to remove the proxy server setting automatically configured by a cable TV-based Internet service provider (ISP) before using the institution's proxy server.

One specialized type of proxy server that doesn't require the user to modify the browser configuration is a "rewriting" proxy Web server. Rewriting proxy servers transform the HTML pages from vendor databases such that URLs on the page are rewritten to point back to the proxy server. Several academic libraries have created their own rewriting proxy servers, often using existing free proxy servers as the basis. One example is the library at the University of Calgary as reported by Eric Tull. Another example, although not mentioned in the survey, is Brown University's implementation.[3] EZproxy is an example of a commercial rewriting proxy server.

One academic library makes its authenticating proxy server available to other campus departments besides the library, but at this time only the main library and the law school library are using the service. The same library is considering expanding the proxy server use beyond remote resource access to bandwidth conservation.

The library at SUNY–Oswego has set up its EZproxy server to allow access to IP-restricted resources on campus Web servers in addition to vendor provided databases. The types of resources made available in this fashion are campus network maps, faculty committee documents, and software for which a site-license for home access has been negotiated. The systems librarian and campus webmaster seek out other campus Web information to make available using this mechanism.

Table 1. Breakdown of Respondents to Survey

Library Type	Count
Academic	68
Public	10
Other/None	7

Table 2. Software Packages Used for Remote Resource Access

Software	Count
EZproxy	29
Web Access Management	14
Squid	9
Apache	6
Microsoft Proxy	4
Netscape Proxy	3
Homegrown	2
Delegate	1
Netware BorderManager	1
Other—More than one	3

Proxy for Filtering

Since proxy servers are in the middle of the transaction between the client and the server, the proxy can examine the HTTP request from the client or response from the server. Based on programming or configuration parameters, the proxy can modify the client's request before delivering it to the server or modify the server's response before returning it to the client. Proxies performing this kind of action are referred to as "filtering" proxies. The modification can affect headers in the HTTP transaction or change the HTML files or other data returned by the Web server.

Eighteen libraries use Web proxies for filtering Internet stations. Software packages used are shown in table 3. There are many reasons for a library to install a filtering proxy. Most libraries use filtering proxies for "allow lists" that permit access to only specific Web sites and "deny lists" that prevent access to specified Web sites. Seven libraries use proxies for deny lists, three for allow lists, and seven use proxies in their libraries for both reasons.

Dan Lester from Boise State University (Idaho) included in his survey response details about how his library uses WinProxy to deny access to Web-based e-mail, gaming, and chat sites.[4] In addition, Lester edits a list of known Web sites with Web-based e-mail, gaming, and chat functions; libraries are encouraged to submit additions and corrections to the list.[5]

Table 3. Software Packages Used for Filtering

Software	Count
Microsoft Proxy	5
Squid	3
WinProxy	3
Apache	1
Bess	1
Netscape Proxy	1
Novell BorderManager	1
WebManager	1
Other—More than one	2

A number of libraries are using plug-ins to Microsoft's Proxy Server to do various forms of filtering. One public library is using the CyberPatrol plug-in to filter content on library stations in addition to using a proxy server. Another library is using a plug-in called Websense to provide optional filtering of sexual materials from patrons. A multi-type library consortium is using the SmartFilter plug-in for Microsoft Proxy Server.

The University of Waterloo (Canada) Library forces public library stations to use a proxy server. A router between the public library network and the campus network restricts HTTP requests to just the proxy server (in addition to other network restrictions). Stations must therefore use the proxy server to access Web resources. The proxy server includes allow/deny directives denying access to Web-based e-mail services.

A public library uses a proxy to filter advertisements out of responses sent back from servers. Other uses for filtering proxy servers not reported in survey responses are to scan files for viruses before they are received, or prevent certain file types (movies, audio files, executables, and others) from being downloaded.

Proxy for Bandwidth Conservation

Bandwidth conservation is typically the reason that organizations other than libraries install a proxy server. The goals are twofold: to reduce the amount of traffic crossing an Internet connection, and to reduce the amount of time it takes for a Web browser to receive content. A caching proxy server does this by storing Web requests and responses for use by subsequent requests. The caching proxy is a server on the local network and browsers are configured to contact the proxy server for every Web request. A browser's first request for an entity (HTML page, graphic, and so on) may take slightly longer to be received because of the added processing required by the proxy server. However, subsequent requests for the same entity by the same browser or other browsers using the same caching proxy will be served faster because the proxy server on the local network can respond more quickly to subsequent requests without traversing the Internet connection.

Nineteen libraries use proxy servers for bandwidth conservation; the proxy servers used by libraries are listed in table 4.

In a related response, one library reported that it employs a proxy server to reduce the load on an old, proprietary Web server that cannot be replaced for several months. Because requests come through the proxy server first, the proxy server rather than the old Web server can handle requests for static content such as graphics and HTML files that don't regularly change. This implementation of a proxy server is called a "reverse" proxy server.[6]

A number of the responses to this question did not specify traditional software proxies, but rather interception proxies. An interception proxy requires no changes to Web clients; it operates instead at a network infrastructure level. Network routers and switches redirect HTTP requests to the interception proxy transparently where the proxy returns the response out of its cache or contacts the Web server for response on behalf of the client. Of the responses to this question, CacheFlow 5000, Cisco's cache engines, Cobalt, Novell BorderManager, and Novell Internet Router are interception proxies. (Cobalt and Novell Internet Router can also be noninterception, traditional proxies.)

Table 4. Software Packages Used for Bandwidth Conservation

Software	Count
Microsoft Proxy	6
Squid	3
Netscape Proxy	2
WinProxy	2
CacheFlow 5000	1
Cisco cache engine	1
Cobalt	1
Novell BorderManager	1
Novell Internet Router	1
Other—More than one	1

Proxy for Gathering Statistics

Another side effect of the interaction among the client, proxy, and Web server is that the proxy server will contain log entries for all of the accesses by the client. By configuring OPAC stations to use a proxy server for requests to vendor databases, a library can get a rough gauge of database usage by examining the log files of the proxy server. Thirteen libraries use proxy servers to gather statistics on Web requests; the proxy servers used are listed in table 5.

The same proxy server used for filtering or bandwidth conservation can be used for gathering statistics. The log files of a proxy server for remote resource access, when correlated with information about what percentage of resource accesses was aided by that proxy server, can also be used to report on resource access.

In the survey, libraries were asked to identify what applications they used to create statistical reports from the proxy server logs. The applications listed were (each application was mentioned once unless otherwise noted):

- WebTrends (4)
- Analog (2)
- Program developed in-house (2)
- HttpAnalyze
- Websense
- Software built into the Novell Internet Caching System proxy
- Excel manipulation
- MS Access for most; WebTrends LogAnalyzer for some

One academic library uses a homegrown counter on links to databases; tracking the number of times the link is accessed gives the library an idea of how often databases are used. Another academic library periodically uses Squid (a full-featured proxy cache) for in-house activity views. Only a limited, random number of sessions are examined.

Table 5. Software Packages Used for Statistics

Software	Count
Microsoft Proxy	5
EZproxy	2
Squid	2
Netscape Proxy	1
Novell Internet Router	1
WinProxy	1
Other—More than one	1

The survey response for a multi-type library consortium included a comment that the institution's proxy Web server does not collect statistics. There are special programs running on the system to delete personal data collected by the proxy server that cannot be disabled in order to protect users' privacy.

Other Results

Respondents were asked if the library publishes a privacy policy regarding the use of proxy server log files. One library includes a statement regarding data collection and use on the proxy server login page:

> This information is collected under the Freedom of Information and Protection of Privacy Act. It is required to verify the identification of the researcher and to authorize access to the database. If you have any questions about the collection or use of this information, please contact the Public Services Systems Librarian.[7]

Boise State University used proxy logs in the arrest of a patron who was viewing child pornography. Four days of logs were given to law enforcement personnel. There was no issue of needing a court order to get the data as it was the library that filed the complaint.

An academic library is installing a Virtual Private Network (VPN) for off-campus clients on DSL and cable modem connections to access resources restricted by IP address. VPNs extend the institution's IP addresses to machines outside the local area network by tunneling network traffic through the general Internet. As such, VPNs work at a network infrastructure layer below that of a Web proxy server, but can accomplish the same result as a Web proxy server for remote resource access.

Conclusions

It comes as no surprise that proxy servers are most often used for remote resource access. Attendees at the four LITA Proxy Web Servers and Authentication workshops stated that learning about remote resource access is their primary reason for attending. In addition, the most common reason attendees have installed proxy servers prior to attending the workshop is to provide remote resource access.

One of the surprising outcomes from the survey is the use of interception proxies by libraries and institutions. Almost one quarter of the responses to the "Proxies for Bandwidth Conservation" question came from libraries using interception proxies. Although public libraries only made up 10 percent of the survey responses, two of four interception proxy installations are in public libraries. The great benefit of this type of proxy server is that it performs the proxying function without requiring modifications to the browser configuration. Consequently, an institution can use one of these proxy servers for filtering, bandwidth conservation, or statistics without a visible impact on the user. Interception proxies cannot be used to enable remote access to databases.

Interception proxies have caused problems for libraries in the past, however, especially when installed by ISPs. The interception proxy changes the IP address of the client making the request to the IP address of the interception proxy. As a result, the database vendor detects the resource request as coming from an IP address outside the range of the institution's IP addresses, and the database vendor will deny access. The interception proxy can typically be bypassed for specified Web servers, but the library must submit a list of database vendor server addresses to the ISP for inclusion in the interception proxy's exception table.

Another surprising result was the lack of privacy statements for the log files of proxy servers. The proxy server's log files are particularly sensitive because the proxy will record all accesses by a client. It is possible to reconstruct the actions and perhaps even the individual searches of a user by analyzing the log files. As users become more sensitive of their personal information being misused in e-commerce transactions, patrons may begin to question the security of their information in the library.

View of the Future

Proxy Web servers are beginning to gain acceptance in library networks. Although the proxy Web function was included in the first Web server and specified in the first version of the HTTP protocol specification, widespread use of proxy Web servers for library issues is only now being seen. Some libraries have used proxy Web servers to solve remote access problems, but there are other ways they can be exploited. What institution doesn't want to offer access to subscription databases to patrons without regard to where the patron is physically located? Or fulfill a policy directive to control the types of material accessed at all or a subset of public workstations? Or extend the life of an expensive connection to the Internet or a network connection between branches by reducing repetitive network traffic? Or be assured that the money spent for subscription databases is effectively used? Proxy servers, ranging from freely available software packages to vendor-supplied turnkey systems, can solve the technical and policy requirements of libraries.

Although proxy Web servers provide a means to address a number of important library issues, in time one hopes that better alternatives will evolve to meet some library needs. Remote resource access is the most popular use of proxies in libraries today, but it represents a cumbersome and inefficient way to solve the remote resource access problem. These proxies can be complicated to set up, both for the user and the library, and cause content for the remote resource user to cross an institution's Internet connection twice. Proxies for statistics give the library just a crude measurement of the use of databases, representing the number of actual HTTP transactions to the database vendor and not the number of searches or records retrieved by patron search sessions. The adoption of standards for interinstitution access control and for the gathering of common statistics should reduce the reliance of proxy servers for these uses. Proxy servers, however, will likely remain a useful way to reduce bandwidth consumption and implement filtering requirements for some time to come.

References

1. Ari Luotonen, *Web Proxy Servers* (Upper Saddle River, N.J.: Prentice Hall, 1998): 4.
2. Timothy W. Cole, "Using Bluestorm for Web User Authentication and Access Control of Library Resources," *Library Hi Tech* 14, no. 1–2 (1997): 62–63.
3. Richard Goerwitz, Pass-through Proxying as a Solution to the Off-Campus Web-Access Problem. Accessed Mar. 28, 2001, www.brown.edu/Facilities/CIS/Network_Services/libproxy.
4. Dan Lester, Blocking Chat, E-Mail, and Game Playing at Library Internet Workstations. Accessed Mar. 28, 2001, www.riverofdata.com/tools/blocking.htm.
5. Dan Lester, Proxy Server Blacklist for Chat, Web E-Mail, and Game Playing Sites. Accessed Mar. 28, 2001, www.riverofdata.com/tools/blacklist.htm.
6. I. Cooper, I. Melve, and G. Tomlinson, Internet Web Replication and Caching Taxonomy. Accessed Mar. 28, 2001, www.rfc-editor.org/rfc/rfc3040.txt.
7. Eric Tull to Peter Murray, personal communication, Nov. 16, 2000.

Appendix A. Web Proxy Use Survey

Introduction

The purpose of this survey is to gather information about the use of Web proxy servers in libraries. Responses to this survey may be used in future presentations and publications. This survey is being conducted in conjunction with the LITA Regional Institute "Proxy Web Servers and Authentication."

The survey consists of a group of questions regarding your library's use or planned use of proxy servers to solve one or more of these problems:

1. bandwidth conservation;
2. gathering statistics;
3. filtering; and
4. remote resource access.

For each area, you will be asked which proxy server you are using or plan to use to solve a particular problem along with followup questions specific to each problem. You can skip an entire major section if it does not apply to your institution.

Proxies for Bandwidth Conservation

Proxies for bandwidth conservation are used to reduce latency (the average time it takes for Web pages to display due to network delays) and network traffic on your network segments. If you do not use a proxy server for bandwidth conservation, you can skip to the next section.

What proxy server does your institution use for bandwidth conservation?
___ Apache
___ Squid
___ Delegate
___ Microsoft Proxy Server
___ Netscape Proxy Server
___ WinProxy
___ WebManager (Sagebrush)
___ Homegrown software
___ Other (please specify): ____________________
___ More than one software package
___ No response

Proxies for Statistics

By forcing all Web requests for Web resources through a proxy server, an institution can use the log files from the proxy server to gather statistics on what resources are used. If you do not use a proxy server for statistics, you can skip to the next section.

What proxy server does your institution use for statistics?
___ Apache
___ Squid
___ Delegate
___ Microsoft Proxy Server
___ Netscape Proxy Server
___ WinProxy
___ Obvia
___ WebManager (Sagebrush)
___ Homegrown software
___ Other (please specify): ____________________
___ More than one software package
___ No response

What statistics program do you use to process the log files? ____________________

Proxies for Filtering

Some proxy servers can be configured to allow access to only specific Web sites (also known as "allow lists") or deny access for specified Web sites (also known as "deny lists"). If you do not use a proxy server for filtering, you can skip to the next section.

What proxy server does your institution use for filtering?
___ Apache
___ Squid
___ Delegate
___ Microsoft Proxy Server
___ Netscape Proxy Server
___ WinProxy
___ WebManager (Sagebrush)
___ Homegrown software
___ Other (please specify): ____________________
___ More than one software package
___ No response

For what purpose do you use a filtering proxy server?
___ Allow lists
___ Deny lists
___ Both
___ Neither
___ No response

Does your institution use proxy servers for other types of filtering (such as removing cookies, blocking advertisements, and virus scanning). If so, please describe: ______
____
____
____

Proxies for Remote Resource Access

Some proxy servers can be used to provide remote access to vendor databases from computers outside your institution's network. If you do not use a proxy server for remote resource access, you can skip to the next section.

What proxy server does your institution use for remote resource access?
___ Apache
___ Squid
___ Delegate
___ Microsoft Proxy Server
___ Netscape Proxy Server
___ WinProxy
___ Ezproxy
___ Obvia
___ WebManager (Sagebrush)
___ Remote Patron Authentication from epixtech
___ Web Access Management (WAM) from Innovative Interfaces
___ Homegrown software
___ Other (please specify): ______________________
___ More than one software package
___ No response

If your proxy server requires users to make changes to their browser configuration, do you provide instructions on your Web site?
___ Yes; please list URL: ______________________
___ No
___ No response

Other Uses of Proxy Servers

Is your library using proxy servers for another reason? If so, please describe: ______________________
____

Does your library publish a statement regarding the type and amount of information collected by the proxy server, and the use and disposition of proxy server log files?
___ Yes; please list URL if published on your Web site:
____
___ No
___ No response

Demographics

Type of library
___ Academic
___ Public
___ School
___ Corporate
___ Special
___ Other (please specify): ______________________
___ No response

Where did you hear about this survey?
___ Posting on Web4Lib
___ Posting on PACS-L
___ Posting on LITA-L
___ Posting on Teknoids
___ Posting on SYSLIB-L
___ Posting on LIS-LINK
___ Posting on PUBLIB-NET
___ E-mail received because I attended a LITA Regional Institute
___ Word-of-mouth
___ Nick Moore's column in *Library Review*
___ Other (please specify): ______________________
___ No response

Institution name: ______________________

Your name: ______________________

Your e-mail address: ______________________

Your name, institution, and e-mail address are optional. This information will be used for followup to survey answers, and will not be published or disclosed to third parties without your consent.