LIBRARY WEB SITE ANALYSIS

A library's Web site functions as the front door of the library's presence to its remote users and is the delivery vehicle for much of its content and many of its services. Especially for remote users, the library's Web site reflects to a large extent the library's identity. Given the strategic reliance that libraries place on their Web site and the amount of resources channeled into its development, the measurement and analysis of its use is an important and well-justified activity.

The development and maintenance of a library Web site represents a huge investment of staff resources. Although small libraries may have a single Webmaster who develops their Web presence, larger libraries generally distribute the responsibility among many people. Tasks include graphic design, information architecture and organization, content development, standards compliance, HTML coding, and server management. The huge number of staff time hours devoted to the Web site further reinforces the need to carefully study how the Web site is used.

As the library analyzes its own Web site, including any locally provided databases or content sources, it gains general insight on what's involved in measuring Web-based resources. This perspective helps librarians develop realistic expectations of what statistics the library requires from external providers of electronic content.

Goals and benefits

Some broad goals that can be achieved by analyzing the use of the library's Web site include the following topics.

Improved usability

The ability to measure the overall use of the Web site and to provide detailed information about the use of each resource within the Web site can be used to make continued improvements. The creation of a library Web site is a complex, time-consuming process. The content provided, the organization of information, and the graphical presentation are each key components of the Web site design. Ongoing measurement of use patterns can be enormously helpful to validating the functionality of the Web site and in improving its usability. Low levels of use of any given resource can indicate either a lack of interest in the content itself or the level of difficulty in finding it. Results from use analysis can indicate whether fine-tuning of the site is needed or if a complete redesign is in order.

Understanding who uses the Web site

One of the most important measures of Web site use involves characterizing the locations of visitors. Not all use of the library's Web site happens remotely. Patrons and staff may heavily use the library's Web site within the building. Some basic information about the use of the Web site relates to breaking down any statistics into categories that correspond to different origins of use:

- **In-house use by patrons.** Libraries almost always provide computer workstations inside the library for access to the OPAC and electronic resources. A significant portion of access to the library's Web site may occur through these in-library public workstations. The library should be able to identify this type of use through the IP addresses that correspond to these workstations and generate use reports accordingly.

- **In-house use by library staff.** Library staff use the Web site in different ways from patrons, so libraries may want to count staff use separately. The number of accesses involved in developing, testing, and managing the Web site is generally considered as overhead and not as actual use. IP addresses can identify accesses originating from this category of users.

- **Organizational use.** Understanding the volume of access that originates from outside the library's own building but still within the organization's network is helpful. Those who access the Web site from within a campus network fall into this category, to use an academic library example. For public libraries, this category may not apply since most of their users rely on commercial Internet service providers (ISPs) and not organizationally provided networks.

- **Remote use.** This category of access includes those who access the library's Web site from beyond the bounds of the library building or its immediate organizational context—access from home computers, remote offices, and other distant locations.

- **Search-engine robots.** When analyzing the activity of the library's Web site, take into consideration the hits attributed to search engine robots. Although libraries enjoy having their Web sites included in the major Internet search engines, a significant number of the page requests originate from the crawlers associated with this model of information discovery. Each search engine employs software robots that systematically visit all known Web servers and retrieve every page available. Fortunately, well-established ground rules exist on how this activity can be done, which ensure the Web servers are not overwhelmed at any given time. These robot exclusion practices allow a Web site manager to place a file called robots.txt in its server's main directory to specify which pages, if any, a crawler may access.

  From a technical viewpoint, page requests from robots look no different from those initiated by Web browsers operated by humans. But pages accessed by robots should be considered differently. Robotic crawlers can account for a large percentage of pages delivered by a server. For servers with low or moderate activity levels, the percentage of page requests attributable to Web crawlers can be significant—possibly as much as 10% or 20% of the total. For high-volume servers, the percentage is less. Web crawlers can affect the counts for any given page. Web crawlers request with equal vigor those pages that contain little interest to human readers. Thus pages that users rarely view have an inordinately high instance of use by the crawlers.

- **Link-checking agents.** Avoid having broken links on the library's Web site. Broken links reflect poorly on the library and frustrate Web site users. Many Web site managers employ automatic link checking software to discover any broken links within their realm of control. Like search engine robots, link-checking agents are automated computer-initiated processes that do not represent actual human activity. Since link-checking inflates the number of page request entries in the server logs, also consider those numbers when analyzing server use.

Categorizing remote users

You can further characterize remote users according to the domain names associated with their IP addresses. The type of top-level domain provides general categories, for example:

.edu = institutions of higher education
.com = commercial businesses
.org = nonprofit organizations (usually)
.gov = government agencies (mostly United States)

Some domain names describe the geographic location of the user. Each country has a designated top-level domain. Within the United States, the domain name also may include a state designation.

The geographical designations in domain names are, however, becoming less common. Especially within the United States, organizations frequently choose not to use geographically identifiable domain names. A typical example is a city using a domain name that reads www.nashville.gov instead of www.nashville.tn.us.

With the continuous growth of the Web, the demand for new domain names is high. Finding available names that are meaningful to the organization is becoming difficult. New top-level names are being proposed, and there is lots of unofficial use of new ones already. The popular .tv domain was originally assigned to the small nation of Tuvalu in the Southwest Pacific region but has been licensed by that country for use by sites associated with the television industry. This arrangement has proven to be a major source of revenue for its citizens.

The domain names and IP addresses associated with individual users tend not to be well-differentiated. A huge portion of home users access the Internet through addresses and domain names assigned by their ISP. The typical home user is recorded in a Web server's logs as coming from a .com or .net domain.

Overall, Web servers do only a mediocre job of describing their use. Differentiating the software robots from human users, much less understanding important demographic details, is difficult. Only a moderate likelihood exists of knowing the geographical location of a given use reported in a Web server access log. Seldom can you distinguish those people who are part of the library's designated community of users from those completely unaffiliated. No demographic information about age, sex, education, or other characteristics of the users can be discerned.

The goal is to measure how many times a Web page is viewed by a human user, subtracting all the activity of software robots and other anomalies that do not reflect real use. To be able to identify users who are part of the organization's target audience would be good. Although libraries enjoy serving a wide audience, most have a well-defined set of users that represent their patrons. Discerning whether a given instance of use is attributed to a member of the library's direct clientele using the system remotely or if that user is completely unaffiliated is usually impossible. As long as the Web preserves its anonymous and unauthenticated character, these uncertainties will likely remain.

Basic terminology

Consider some definitions:

Requests. Often called *hits,* this number represents each time each file is requested from a server. If a page includes images, each of these associated files counts as a new request. Requests measure overall server activity, regardless of how the requests relate to user behavior.

Page views. Given that the act of viewing a complete page by a user may incorporate many individual file requests, counting each complete page view corresponds more closely to the actions of the user.

Sessions. When a person visits a Web site, he or she likely navigates through many pages of the site. But given the stateless nature of the Web, the concept of a session is an artificial construct. Sessions are generally defined as a series of page views originating from a single IP address that take place within an identifiable continuous time frame.

Entry page. This page marks the beginning of a visitor session, identifying the specific page the visitor first viewed on entering the site. For most Web servers, the most common entry page for visitor sessions is its default page. In some cases, however, popular pages within the site stand out as common entry points. Links to internal pages from external sites or search engines can lead visitors to specific pages within the site rather than the main top-level page.

Exit point. Marking the end of a session, this page indicates the point at which a visitor left your site for another, waited idle for a considerable time period, or closed the browser.

Stateless versus statefull. In its simplest form, the Web is a stateless system. Each request for a page from a browser to a Web server is a complete and independent transaction. No persistent connection is maintained. This system contrasts other network communication systems, such as telnet, which maintains a continuous connection between the user's software and the server. A system is statefull when it keeps a steady stream of communication between clients and servers. Stateless systems treat each request independently without tracking any given user session.

Although stateless systems offer an advantage in communications efficiency, they also present challenges. Creating the effect of a continuous session is desirable for people as they visit your Web site. You may have them log in to gain access to information held in a personalized account, to gain access to restricted resources, or to preserve their preferred settings. Since the Web is a stateless system, tracking the actions of a given user as he or she navigates through your Web site is difficult. User sessions can usually be identified by grouping requests from a given IP address during a limited time span.

Cookies. In reaction to its stateless nature, many Web-based systems use client-side cookies to make the discrete tasks associated with a person using a Web site cohere into a continuous session. Cookies are small bits of data a Web site can store on the user's Web browser. A session cookie can be set, for example, when a user enters a Web site but is removed when the user leaves the site or exits the browser. As the user goes from one page to another on the Web site, the presence of an identifier in the cookie can be used to tie the session together. A persistent cookie also can be used to maintain information between sessions.

Methods for measuring Web site use

Hit counters

In the early days of the Web, hit counters were popular. These counters were typically placed on the bottom a Web page, announcing the number of times the page has been viewed since its original creation. The counter increments each time the page is requested. These viewable hit counters have fortunately fallen from fashion.

By current standards, counting hits is considered the most unsophisticated and least helpful approach to measuring Web site use. Problems with simple hit counting include:

- All page requests are treated as equal.
- Page requests from robots are generally not excluded.
- Pages delivered from cache are not counted.
- Hit counts reflect cumulative use over time, not for any given time period.

Hit counters usually rely on a script that reads and increments a stored numerical value each time the server delivers the page. Counters can be established for the overall Web site or for any individual page.

In addition to the hit counters that work through a script installed locally on your Web server, many free or commercial services exist that provide a use counter on your Web pages without any local programming. These counters operate by sending information to their own server each time a page on your server is accessed. Web editors simply add a snippet of HTML code according to the service provider's specification to each of their pages.

Although these services may offer a level of convenience for those that do not have the ability to run scripts on their Web servers, they do introduce many concerns. The use of many third-party Web site counters can degrade performance since each page access on your Web site also involves contact with the service provider's site. But the main concern is that this arrangement allows a third party to capture information about the use of your server. A library cannot be certain how this detailed use information might be treated on the service provider's server. These companies may not have the same concern for privacy and confidentiality as that held by the library community.

Analysis of Web server log files

Analysis of the access logs maintained by the server is a more sophisticated approach to measuring Web site use. Each time a Web server receives a request for a page, the Web server makes a record of that access in its log files. Especially on busy servers, these log files can be voluminous, but they are essential to measuring and interpreting how the server is used.

The data held in Web server logs underlies almost all Web site analysis. Web server logs are the basis of the understanding the use of a library's Web site, its Web-based online catalog, and any external Web-based resources.

Given the broad use of this information, looking in some detail at the information they record is valuable. Although some variation exists in how Web servers record information, most follow the same conventions. Most

Web servers follow the “Common Logfile Format,” which makes log analysis software able to understand the log files of those servers without detailed configuration.

The following figure shows a few entries from a typical Web server log.

```
129.59.150.100 - - [12/Apr/2002:10:25:28 -0500] “GET /
librarylogo.gif HTTP/1.1” 200
    5663
129.59.150.100 - - [12/Apr/2002:10:27:19 -0500] “GET /
HTTP/1.1” 200 276
129.59.150.100 - - [12/Apr/2002:10:27:19 -0500] “GET /
librarylogo.gif HTTP/1.1” 304 -
129.59.150.100 - - [12/Apr/2002:10:27:54 -0500] “GET /
HTTP/1.1” 200 268
129.59.150.100 - - [12/Apr/2002:10:27:54 -0500] “GET /
librarylogo.gif HTTP/1.1” 304 -
129.59.150.100 - - [12/Apr/2002:10:39:37 -0500] “GET /
diglib HTTP/1.1” 404 287
129.59.151.140 - - [12/Apr/2002:10:39:50 -0500] “GET /
diglib/login.pl HTTP/1.1” 500
     616
129.59.151.140 - - [12/Apr/2002:10:48:03 -0500] “GET /
diglib/login.pl HTTP/1.1” 200
    1100
129.59.151.140 - - [12/Apr/2002:10:51:09 -0500] “GET /
diglib/login.pl HTTP/1.1” 200
    1074
129.59.151.140 - - [12/Apr/2002:10:51:21 -0500] “POST /
authenticate.pl HTTP/1.1” 404
     296
129.59.151.140 - - [12/Apr/2002:10:51:52 -0500] “GET /
diglib/login.pl HTTP/1.1” 200
      1081
129.59.151.140 - - [12/Apr/2002:10:52:02 -0500] “POST /
diglib/authenticate.pl
     HTTP/1.1” 302 345
```

Some of the basic data elements include:

Host. The address of the computer that initiated the page request is the host. The address can be recorded in its basic numerical form of a raw IP address (such as 129.59.150.100) or the server can attempt to resolve the address to its registered, fully qualified domain name (breeding.library.Vanderbilt.edu).

```
129.59.151.140 - - [12/Apr/2002:10:39:50 -0500] “GET /
diglib/login.pl HTTP/1.1” 200
     616
```

The domain name is more valuable to understanding server use than the raw IP address but procuring it takes additional processing time. Before recording the domain name, the server must issue a request to the DNS (Domain Name System). For most Web servers, this additional processing overhead poses no significant problem. Configuration options in the server determine whether IP addresses should be resolved to their domain names, through a process called a reverse DNS lookup. Some server administrators prefer to record only the IP address in real time and resolve the addresses when they analyze the log files. Most Web log analysis software can resolve the host names if the logs contain only raw IP addresses.

As noted above, since domain names carry information about the type of organization (.com, .edu, .org), or its country or origin, resolving IP addresses to domain names is essential if you want to gather this type of demographic data. Some IP addresses are not registered in DNS, making discerning the origin of the request difficult.

Date. Each request includes a date and time stamp. This information is valuable for analyzing patterns of use by time of day, day of week, or by month or season. Most servers record the date in the form: [day/month/year:hour:minute:second zone].

```
129.59.151.140 - - [12/Apr/2002:10:39:50 -0500] “GET /
diglib/login.pl HTTP/1.1” 200
     616
```

Request. The request usually consists of a directive, such as GET or POST followed by the name of a file, representing the Web page requested. The server records all files it delivers, including graphics files, style sheets, and other components in addition to the basic HTML page. All scripts executed on the server also are recorded in the logs.

```
129.59.151.140 - - [12/Apr/2002:10:39:50 -0500] “GET /
diglib/login.pl HTTP/1.1” 200
     616
```

Status. This component of the log file indicates how the server was able to respond to the request. Not all page requests on a Web server are successful. Identifying any requests with abnormal status codes can help ensure the Web server operates reliably with no pages containing broken links.

```
129.59.151.140 - - [12/Apr/2002:10:39:50 -0500] “GET /
diglib/login.pl HTTP/1.1” 200
     616
```

A standard set of three-digit result codes describes the relative success of a page request:

Source: Status codes in the HTTP protocol. W3C World Wide Web Consortium: www.w3.org/Protocols/HTTP/HTRESP.html

Code	**Meaning**
2XX	Success. All the status codes in the 200 series indicate the request was successful.
200	OK. The item requested was delivered successfully. In a healthy Web server, most log entries should show this result code.
201	Created. This response code can be issued in response to a POST command, indicating that the script was successful, and that a document was created that is available by a specified URL.
202	Accepted. The request was accepted for processing by the server but was not completed.
203	Partial information. This status code indicates that the page returned is not necessarily complete and definitive but may be from a private copy or cache.
204	No response required. The execution of the script was successful, but no visible information displayed.

3XX Redirection. These result codes indicate some action other than simple page delivery needs to be taken to satisfy the request. In most cases the action is taken automatically.

301 Moved. The page request has permanently moved, and the browser should display the page from the URI specified in the response. From the user's perspective, this redirection takes place transparently.

304 Page not modified. If a page appears in the client's local cache, the local cache may ask the Web server to deliver a new copy of the page only if the page has been modified after a designated timeand date. If the page has not been modified, the server responds with this status code, indicating that the local copy can be used safely. Using a cached copy can reduce network server load and increase perceived performance.

4XX All result codes in the 400 series indicate a problem with the browser or the structure of the request.

400 Bad request. The request contains a syntax error or has some other problem that prevents the server from understanding or satisfying the request.

401 Unauthorized. The requested resource is restricted and requires authentication before proceeding. The typical response would be for the client to retry the request with the proper authentication.

402 Payment required. The item requested requires payment for access, but the URL presented did not include the required "ChargeTo" header necessary.

403 Forbidden. The page requested is restricted and requires authentication for access. In most cases this code indicates that the process of authentication was attempted but not successfully completed.

404 Not found. No document on the server matches the URL specified in the request. Most often seen when pages have been removed from the server with no redirection specified. Web site managers should monitor the Web server's logs to reduce future occurrences of these errors.

5XX Server error. All codes in the 500 series indicate some problem with the Web server itself.

500 Internal error. If a significant number of requests in the access log show this result code, carefully review the configuration of the server.

502 Timed out. The server did not respond within the time allotted.

Bytes. The size of the file delivered by the server is recorded. This value, in bytes, helps server administrators gauge the total load. From the server's perspective, delivering huge files represents more network and internal server bandwidth than working with smaller files.

```
129.59.151.140 - - [12/Apr/2002:10:39:50 -0500] "GET /
diglib/login.pl HTTP/1.1" 500
     616
```

Although these standard elements are captured in a Web server log, other items of information also can be recorded.

Referrer. Knowing how users arrive at your Web site is interesting. One of the data elements available is the referrer, which is the URL the user clicked to arrive at your site. The Referrer can either be written into a separate log file or incorporated into the server's main access log. The following example shows the page ltg.html was accessed through a link on page http://staffweb.library.vanderbilt.edu/breeding.

```
http://staffweb.library.vanderbilt.edu/breeding/ -> /
breeding/ltg.html
```

Analysis of the server's HTTP referrer information can provide clues to how users find your site and on how they move from page to page within your site.

User-agent. As they make requests, Web browsers identify themselves with a signature that identifies their brand and version. Given the differences in the way the various browsers render HTML, Web designers need to track which browsers frequent their site. Like the referrer, the user-agent information is optional and may be added to the standard access log or recorded into its own log. The following user-agent string represents Microsoft Internet Explorer, Ver. 6 operating on Windows XP.

```
Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
```

Error log. Web servers maintain an error log that records all failed or invalid requests. Although much of the information in the error log is also written to the primary access log, the error log is a convenient way to discover problems with the Web site or with external servers that have outdated versions of links to pages on your server. The error log also may contain errors generated by scripts that are not written to the access log.

Log rotation and retention

Web server logs can grow quickly and become huge. Periodically, these log files need to be rotated, which involves starting a new file, while keeping the previous one for analysis. Most Web server software can perform log rotation automatically on any set interval. Web servers with high loads likely are set for daily log rotation; for others, weekly or monthly rotation is adequate. Log files consume a significant amount of disk space but should be kept as long as possible for analysis in the future. Before deleting any logs from the server, copy them onto CD-ROM, DVD, or magnetic tape for future reference.

Time spent on the server

The Web as a stateless system does not automatically track use as an end-to-end session with a specific beginning and end time. One of the challenges in analyzing server logs involves forming an impression of a continuous user session from a set of independent data elements. Relationships among log entries must be inferred, given the absence of absolute connections. User sessions must be reconstructed through the user logs. Only by correlating IP addresses with the timestamps is tracking how long any given person spent on your site possible. The time difference from one request to another is usually the only measure of how much time was devoted to any given page.

A **timestamp** is the current time of an event that is recorded by a computer.

Limitations of Web server log files

Remember that the server logs do not necessarily represent intended actions by the user of a Web site, but the logs are a technical representation of tasks performed by the server. Lots of conditions exist where machine-generated activity is recorded, where unintentional user behavior results in superfluous log entries, and where the log files fail to capture use.

Several artifacts of user behavior can skew statistics. If a user reloads a page with the Refresh button of the Web browser, this action registers on the Web server as another page request, even though it does not represent a new instance of use. In some circumstances, using the forward and back buttons of the browser to navigate through the site records as new entries pages already viewed. Both of these circumstances cause access statistics to be artificially high.

Caching

Under some conditions, a user can view a Web page without it being recorded in the server's log files. To speed perceived performance and decrease the load on networks and servers, pages retrieved from the server are stored in a temporary area, called a cache. When requesting pages, the browser checks the cache to see if the page can be obtained locally without traversing the network or bothering the server. The basic concept revolves around asking for any given page or supporting components once, going back to the server only for new pages or for pages that have been updated since they were first delivered. Most Web servers use the same graphics for all their pages. These supporting graphic files need be loaded only once. Subsequent pages viewed can take advantage of the graphic files available in the cache. Since the graphics files tend to be larger than the HTML pages that call them, the use of a cache greatly improves speed and efficiency.

The most common type of cache is managed within the Web browser. The browser has to procure a copy of each page at least once, but subsequent views of the page can take advantage of the cache. Depending on configuration settings, the browser retrieves a new copy of a page with each session, or pages may be loaded from a cached copy for longer periods of time. The use of the browser's cache makes a tremendous difference in the speed in which pages load, since loading files from the computer's hard disk can be an order of magnitude faster than retrieving them through the Internet. This difference is even more dramatic for computers that connect to the Internet through slow dial-up connections.

Some organizations set up a cache that is shared among a group of users, meaning that if any user has requested a page, others are able to fetch it from the cache rather than going to the original Web server. Such caching is usually transparent to the users and can make a significant difference in the overall load of an organization's network. Some ISPs create a cache that is shared by thousands of users.

Even though these shared caches must be accessed through the network, the part of the network where the cache resides is close to the user and therefore faster to access than most Web servers. Although the improvements in speed made possible through a shared network cache are not necessarily dramatic, they can significantly reduce the overall level of network traffic between an ISP and the rest of the Internet.

Sometimes while working with a cached environment, though, the patron doesn't end up viewing the current version of each page unless he or she clicks the Reload button.

The main negative effect of caching from the measurement and assessment perspective is that for a certain number of pages viewed, no record exists of the use in the server's logs. This effect means the server logs can often underrepresent the resources viewed by users.

Given the limitations inherent in relying on the Web server logs alone, libraries may want to consider other strategies to gain understanding about how their patrons make use of their library Web site. The library may want to add scripts to some of their pages that provide an additional layer of statistical monitoring. The implementation of a MyLibrary-style environment not only offers a set of personalized services, but it also provides the means to collect additional information on the way library resources are accessed by particular classes of users. The library also may consider using a quick survey form to gather demographic information about their virtual user base and to procure feedback on issues relating to the library's Web site, its services, or other issues of concern.

Web server log analysis

Many software programs are available for the analysis of Web server log files. You don't need to be an expert statistician or a programmer to transform the raw data of a log file into a meaningful representation of Web server use. Web server log analysis software ranges from simple shareware programs that can be obtained without cost to commercial products that have sophisticated features and require payment of license fees.

The following products and services are examples of what is currently available for analyzing the use of a Web server. Though the list is not comprehensive, it does present a range of the offerings, including high-end commercial systems as well as open source software that can be obtained for little or no cost.

www.webtrends.com

NetIQ offers many products designed to provide sophisticated analysis of an organization's Web site. The products are scaled according to the complexity of the Web site, ranging from ones that can handle a large cluster of Web servers to single-server products.

- **WebTrends Log Analyzer**, the company's basic product, provides log file analysis for a single Web server. The software produces many reports in HTML format, summarizing Web server use organized to several broad sections: general statistics, resources accessed, advertising, visitors and demographics, activity statistics, technical statistics, referrers, and keywords. Each section includes a table that presents the data in summarized form, a graphical representation, followed by tables with detailed data. The product works with Web server log files in any form. The software runs on Windows NT/2000 and retails at $499. This version of the product is appropriate for libraries with only one or two Web servers.
- **WebTrends Analysis Suite** offers all the reporting features of WebTrends Log Analyzer but with expanded and advanced capabilities appropriate for more complex Web environments. Going beyond after-the-fact log file analysis, the system includes many real-time monitoring

capabilities. Many features provide technical diagnostics that can be used to tune Web server performance and monitoring services to ensure the availability and stability of each system in a multiserver Web cluster. This version also includes modules for the analysis and monitoring of streaming media. The WebTrends Analysis Suite includes an optional server plug-in that can be loaded on the Web server for advanced monitoring. The pricing for the WebTrends Analysis Suite Standard Edition starts at $999 and the Advanced Edition starts at $2,499. The server plug-in is $1,499. The Standard Edition is priced for use with a single server, but that server can host up to 100 domains. The Advanced Edition can support multiple servers through the purchase of additional server plug-ins.

- **WebTrends Live** offers a similar set of features as the WebTrends Analysis Suite through a remotely hosted service. This version of the system operates through information passed in real time from the Web server to the WebTrends Live service instead of using log files. The service is priced according to the size and complexity of the organization's Web server environment. The e-Business version, scaled for a single server with less than 50,000 page views per month, costs about $35 per month, and the Enterprise edition, which supports e-commerce-related features, costs $2,000 per month.

Webalizer is a free, open-source Web log analysis program. The program operates under the Linux operating system on Intel-based servers. It produces reports in tabular format, supplemented with graphs for the broader categories. Report sections include: general monthly statistics (hits, files, pages, visits, file size), daily usage graphs, hourly statistics, and tables of the top resources accessed by URL, file size, entry pages, and exit pages, as well as tables that list the statistics for each resource. A section describes the top referrers that led visitors to the site, including tables that show the keywords used in the search engines. The types of reports available compete well with its commercial competitors, and the presentation of the information is understandable and attractive.

www.mrunix.net/webalizer

FlashStats from Maximized Software Products is a low-cost log file analysis program that generates 12 reports: summary report, top URLs requested, top referrers, search phrases, most common browsers, bad URLs, bad referrers, user domain analysis, types of domains, daily totals, hits per day of week, and hits per hour. Most of the reports are in tabular format, with the daily totals, hits per hour, and hits per day of week presented graphically. The software operates on Windows, Macintosh, and many Unix platforms, and supports major Web server log formats. The program is priced at $99 for the standard edition, which supports up to 25 Web sites, or the ISP edition, which has no limitations on the number of sites.

www.maximized.com/
products/flashstats

Web-Stat Traffic Analysis is a remotely hosted hit-counter system that provides general Web server traffic analysis rather than detailed log file analysis. The system works by adding a snippet of HTML code to the main page on the Web server. The page then displays a counter that increments each time the page is requested by a user, sending information to the Web-Stat server, where it is recorded in a log file associated with that server. The Web site manager can then log into the Web-Stat server to view reports about the server's use. Web-Stat can generate reports for total traffic in the last 30 days, returning traffic in the last 30 days, first-time traffic in the last 30 days, total traffic on the site since its creation, traffic per month for all previous months, returning traffic per month, first time traffic per month, traffic per week, traffic per hour of day, referrers to the site, a search engine tracker

www.web-stat.com

including search strings used, country of origin, and type of browsers used. Although this service does provide some general measurements on the overall load of the Web server, it does not provide detailed page-by-page use analysis, track visitor sessions or offer many of the other features commonly found in Web server log analysis software. Communicating with the Web-Stat server can impede the performance of the Web server. Even at the 0.22 seconds claimed by the vendor, this time is much slower for a direct page load than a typical Web server. The cost for using Web-Stat is $5 per month.

www.deepmatrix.com/
livestats6_corp

LiveStats from DeepMatrix takes a different approach from the majority of the log file analysis applications. LiveStats reads the Web server's log files and transfers all the data into a relational database. For clustered Web servers, up to 10 server logs can be imported into the database. Once the data have been processed into the database, reports can be generated on demand or viewed through a Web browser. Many reports can be produced, including the standard use statistics, as well as sitemaps, forecasts to predict future use based on current trends, and many other features geared to the e-commerce Web site. Since the data are held in the database and not limited to a given month's Web server logs, reports can easily incorporate historical trends. The LiveStats software includes three main components, the Data Collector, the Reporting Server, and the SQL database, and can be purchased for $695. It operates on the Windows NT/ 2000, and comes bundled with the MySQL relational database. The system also operates with Microsoft SQL Server but must be licensed separately from Microsoft.

www.boutell.com/wusage

WUSAGE 8 from Boutell.Com, Inc., is a low-cost Web server log analysis program. It produces reports from a calendar menu of daily, monthly, and weekly selections. For each of these periods, reports can be generated for all pages served, page views, home page views, unique IP addresses, unique cookies used at least twice, and unique visitors. Each monthly report includes a narrative executive summary and the usual options of statistics: general totals, access per hour, access per day, top documents, top entry pages, top exit pages, documents by directory, a navigation report, top visitor sites, top visitor domains, top users, top proxy sites, top Web browsers, top operating systems, top referring URLs, top referring sites, top search servers, top search keywords, CGI parameters and values, documents not found, and access by result code. WUSAGE costs $25 for a single server license, $75 for a five-domain license, and an unlimited Web hosting license for ISPs for $295.

www.analog.cx

Analog, created by programmer Steven Turner, is a free, basic Web server log analyzer program. The program was one of the earliest to perform Web log analysis and continues to be popular. Although its reports are plain, they generate quickly. Available reports include a general summary, monthly summary, daily summary, hourly summary, listing of source domains, visits by organizations, search word, operating system statistics, status code, requests by file size, file type report, directory summaries, and totals of requests by page.

Almost any of these products are adequate for providing the basic level of analysis needed by the typical library. Many of these systems are designed to accommodate a complex e-commerce environment, where tracking customer transactions, banner ad impressions, and other features not common in library Web sites is important. Some of the systems target ISPs that may host hundreds of Web sites. Compared with the commercial arena, the needs that libraries have in monitoring their Web servers are relatively modest.

A typical strategy for a library is to start out with one of the free or no-cost Web analysis applications to see if the program is adequate before investing in a more expensive commercial application. Unless you are in the e-commerce or ISP arenas, the commercial packages offer only a few relevant features not available in the low or no-cost programs.