Goddard, Lisa. "Getting to the Source: a Survey of Quantitative Data Sources Available to the Everyday Librarian: Part I: Web Server Log Analysis" *Evidence Based Library and Information Practice* [Online], 2 14 Mar 2007
Final version: http://ejournals.library.ualberta.ca/index.php/EBLIP/article/view/196/240

Getting to the Source: a Survey of Quantitative Data Sources Available to the Everyday Librarian: Part 1: Web Server Log Analysis [Post Print Version]

Lisa Goddard
Emerging Services Librarian/Division Head for Systems
Memorial University of Newfoundland Libraries
St. John's, Newfoundland, Canada
Email: lgoddard@mun.ca

Received : 01 December 2006 Accepted : 07 February 2007

Introduction

So you're a librarian who has decided to embrace evidence based decision-making. Your reasons might be to justify a purchase decision, to measure your organization's effectiveness, to identify more efficient ways of organizing your library's limited resources, or to provide improved services to your patrons. Finding a place to start this undertaking may seem to be an overwhelming task. However, there is an enormous amount of data about patron activity available for your review and analysis logged on servers in your library every day. Each of these servers produces reams of transaction information in the form of text files and databases, as do those of your remote resource providers. Every link you follow, every search strategy you develop, every online form you complete, and every byte you download is recorded and stored among billions of lines of similar data on servers all over the world. In addition to all this data, computing systems operate on high capacity networks and utilise powerful computers capable of correlating, processing, and organising much of this data into neatly formatted reports for human consumption.

There are several requirements to harnessing this data, and one of the first is simply knowing what kinds of information may be available and how they can be accessed. Server logs are usually protected behind firewalls and strict authentication mechanisms, and reporting tools are often hidden in restricted administration interfaces. This article is intended to help identify some of the data sources and tools that can aid in planning and decision-making.

Web Server Log Analysis

All Web-based electronic resources and services are made available through Web servers. In this context 'Web server' refers to a software application such as Microsoft Internet Information Server or Apache HTTP server. The Web server operates by accepting HTTP requests from a browser (typically in the form of a URL) and providing an HTTP response (typically an HTML document) to the requester. Each Web server application has the ability to log each of the requests it receives, and the responses that were made to each request. These Web server logs are usually stored on the server as plain text files. Web server logs provide a rich source of quantitative data for any librarian who wishes to gather information about usage of electronic services or resources.

Because almost all electronic resources produce Web server logs, it is useful to look at the elements of these logs in greater detail. The first section of this article will outline the kinds of information that can be gleaned from Web server logs, discuss several Web log analysis tools to help librarians aggregate Web log data for user-friendly reports, and provide some caveats about interpreting Web server log data.

What's in a Web Server Log?

Each time a Web page, image, or object on a site is accessed through a browser, a record of the transaction is written to the log. The server administrator determines the amount of information written to the log and will sometimes choose to log less information about each request in order to prevent log files from growing too large. A library's server logs may have more or less data, depending on the log format chosen by the administrator.

One of the critical aspects of Web log analysis is to ensure that logs are collecting the kind of data that will be required to make evidence based decisions at a later time. If data is collected in order to understand patron needs and to create a more usable and visible site, then it is necessary to determine which statistics can be used to measure those needs. You must know ahead of time what kinds of questions you hope to answer with Web log analysis, and carefully choose the log elements to provide sufficient and accurate measures.

Web Log Data Elements

There are nine types of basic Web log data that may be useful to librarians when conducting quantitative analysis of electronic service and resource usage:

- Date/Time
- Requested Item
- URL Query
- Referrer
- Client IP Address
- Host Name
- Unique Session ID
- Client Side User Agent
- Client Side Username

This section will describe each type and will provide examples of each from raw log data.

Date/Time: the year, month, day, hour, minute, and second that a request is made.

```
2006-11-04 21:47:33
[16/May/2006:00:18:41 -0230]
```

The date/time stamp is a basic element included in all standard log formats. It helps to determine the times of year, days of the week, or hours of the day that an electronic resource or service is most heavily used. This element will identify increasing or declining patterns of use over time for the entire site or for a specific page, resource, or object. The date/time stamp is also used in any report that attempts to gauge the length of a patron session or the length of time a patron spent looking at a particular Web page.

Requested Item: the URL of the requested item.

/guides/howto/index.php
/qeii/cns/photos/cnsphoto0108002.jpg

Identifying the requested item's URL allows one to see which resources, objects, or services are being accessed most often by patrons. It is used to generate top *n* analysis and to provide counts of hits on specific pages or objects.

URL Query: the portion of the request that appears after the question-mark (?) in a dynamic URL (Uniform Resource Locator).

/eindex/DBSearchResults.asp?subhead=Environmental%20Science

/eindex/alphaSearchResults.asp?SearchText=W

/viewnews.php?item=265

/query.html?col=spidert&la=en&qt=+Dictionary+of+newfoundland+English

/cgi-bin/docitemview.exe?CISOROOT=/Newfoundlandquarterly&CISOPTR=602

/?genre=article&isbn=&issn=00332917&title=Psychological+Medicine&volume=36&issue=8&date=20060801&atitle=The+temporal+relationship+of+the+onsets+of+alcohol+dependence+and+major+depression%3a+Using+a+genetically+informative+study+design.&aulast=Kuo%2c+PoHsiu&spage=1153&sid=EBSCO:PsycINFO

If users have the option to search against a back-end database, the query portion of the URL contains the search term entered or the search option selected from a menu. This is important if an e-index or e-journal title search box is present on the page. It would also be helpful to analyse queries entered into the library's 'site search' engine. Resolvers rely on the OpenURL standard and full citation information for a requested resource in the query portion of each URL generated.

Referrer: the site and page that referred a visitor to the site.

http://www.library.mun.ca/guides/howto/write_book_review.php

http://images.google.com/imgres?imgurl=http://www.library.mun.ca/hsl/bates/Ch17p490b.jpg&imgrefurl=http://www.library.mun.ca/hsl/docs/Bates.php&h=370&w=400&sz=38&tbnid=OqAdKkWsz3muXM:&tbnh=115&tbnw=124&hl=en&prev=/images%3Fq%3Dmusculoskeletal%2Bsystem%26svnum%3D10%26hl%3Den%26lr%3D&frame=small

http://www.google.com/search?hl=en&lr=&q=sample+annotated+bibliographies&btnG=Search

http://Web.ebscohost.com/ehost/results?vid=42&hid=123&sid=60138425-1342-4032-811c-7121f3daf2e1%40sessionmgr102

It is useful to know the referring page to determine the methods by which users discover resources on the Web site -- whether they are following links from within the site, from other sites, or from search engine results. Referrer data is also useful to measure the impact of Web site metadata initiatives and search engine optimization (SEO) projects.

Client IP Address: the unique IP address of the computer making the request.

134.153.184.170
66.249.72.4

The client IP address is included by default in all standard log formats. The IP address helps to identify each individual computer using the Web site, and it can be used to track repeat visitors. Along with date/time stamps, the IP address can help to determine the path a user has taken through a site and the resources viewed during that visit. Some of the difficulties involved in using IP addresses to track unique user sessions are discussed below ("Interpreting General Summary Statistics: Terms and Definitions").

Host name: the computer host name and domain to which an IP address belongs.

beluga.library.mun.ca
wiley-411-2130.roadrunner.nf.net
crawl-66-249-64-54.googlebot.com

If a Web server is configured to perform reverse Domain Name Server (DNS) lookups, it will automatically translate each requesting IP address into a full DNS host name in the log. If, for example, the IP address '134.153.184.70' appeared in the log file, the Web server could perform a reverse DNS look-up on the fly. The server could also record the host name of the requesting computer, which in this case would be '<proxy1.library.mun.ca>.'

Knowing the domain suffix of each user's Internet service provider (ISP) makes it possible to run reports that may provide information about a client's organisational membership and geography based on the domain to which the IP is registered. Reverse DNS look-ups can slow down the performance of a busy Web server. The server administrator will need to determine whether or not it is advisable to log this information. Many Web log analysis packages will perform DNS look-ups against IP addresses when running their reports, permitting domain name analysis even, if this information does not exist in the log file.

Unique Session ID: Uniquely identifies each client session.

PHPSESSID=8aa4a615a20382e917731ffc8c6e6bd5

IP addresses are an unreliable means to use to identify individual visitors. Dynamic scripting languages such as PHP can be set to automatically generate a unique ID for a user session which can be propagated in each URL selected by the user during that session. This provides an easy way to track a user's path through the site during a session. Unique session IDs are useful for click-path analysis, for generating statistics about the duration of each user session, and for counting the number of resources viewed during each session. Cookies are another method frequently used for tracking individual sessions. More information on the use of cookies is presented below ("Interpreting General Summary Statistics: Terms and Definitions").

Client Side User Agent: Records information about the client's browser type, version, operating system, and language

Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+InfoPath.1)

Mozilla/5.0+(Windows;+U;+Windows+NT+5.1;+enUS;+rv:1.8.0.1)+Gecko/20060111+Firefox/1.5.0.1

Mozilla/5.0+(X11;+U;+Linux+i686;+en-US;+rv:1.7.13)+Gecko/20060418+Fedora/1.0.8-1.1.fc4+Firefox/1.0.8

Mozilla/5.0+(Macintosh;+U;+PPC+Mac+OS+X;+en)+AppleWebKit/312.8+(KHTML,+like+Gecko)+Safari/312

The client side user agent provides information on which operating systems, browser types, and versions used by patrons. This information can be helpful when making decisions about site functionality and design. It also indicates whether patrons frequently access the site using small screen devices such as mobile phones or handheld computers.

Client Side Username: Authenticated username entered into a name and password dialog when some portion of the site is restricted, such as

lgoddard

Sites that require authentication can also log usernames, so resource requests can be tracked according to the username of the individual making a specific request. If a site does not require a login, then the username field in the log will contain a dash (-) indicating that username information is not available. Most libraries ask users to login only to those services that require personal identification, (e.g., personal account information or licensed resources). As a result, this field is rarely available as a general statistical measure.

What's Not in a Web Server Log?

Although Web server log analysis can provide a great deal of quantitative information about Web site usage, there are also some very real limitations to the available data.

○ **External Links** - Web logs track requests only for resources that reside on the Web server. Many library Web sites contain links that point to external sites, such as those of e-journal and e-index providers, online reference sources, and useful sites available elsewhere on the Web. When a patron clicks one of these links, that selection is not recorded in your Web log, but in the log of the target server. As a result, Web log analysis cannot determine which external links or resources patrons are accessing from the library's Web site. There are other strategies to help track use of external links on your Web site. Several are examined later in this paper ("Where Did They Go?").

○ **User Profile Information** – Unless a user login is required, Web logs contain no information that allows the identification of a particular individual who has visited the site. Web logs cannot analyze usage according to personal characteristics such as age, gender, or affiliation.

○ **Qualitative Data** – It is difficult to draw conclusions about a patron's reasons for visiting a site. The log data will not help determine whether the resources found there met his or her information needs or how the data was used.

○ **Cached Pages** – Most Web browsers have their own cache where they store and serve frequently viewed pages. If the browser replies to a user request by returning a page cached on the user's hard drive, the Web server log will have no record for that request. This makes it difficult to obtain accurate data about the number of times pages are being viewed by patrons.

Tools and Guidelines for the Analysis of Web Server Logs

Web logs contain a great deal of useful information about resource use, but this information is presented in enourmous plain text files with thousands of entries for any given day. Every single object returned is recorded as a separate hit. For example, a single Web page with nine individual embedded images will generate ten entries in the Web log – one entry for the page, and an additional entry for each of the images. . 'Server-side includes' are a way of pulling images, side-bars, menus and other common elements into pages as they load. If 'server-side includes*'* are used in web site design then each of these items also generates a log entry when a page is viewed A single user session can generate hundreds or thousands of individual lines in a log file. Web logs can contain millions of lines and achieve file sizes well over 500 MB.

Very large text files are extremely difficult to manipulate. It is often impossible to read log files in text editors (e.g., Notepad, WordPad, or MS Word), as these applications are not optimised for huge files. Readers with programming experience may be able to write scripts to extract pertinent data elements from log files. Perl, a general-purpose programming language originally developed for text manipulation, is excellent for parsing and manipulating log files. However, there are many commercial and freeshare software packages available to extract information from Web server logs for those without this expertise.

A Web log analyser is a piece of software that parses the information from Web logs, and uses it to generate different types of reports that may be delivered as HTML pages and charts, in text files, or in Excel spreadsheets, depending on user preference and the software package. Two of the more popular commercial log analysis packages are WebTrends (<http://www.Webtrends.com/>) and ClickTracks (<http://www.clicktracks.com/>).

There are also several excellent open source solutions released under the GNU General Public License and freely available from the Internet, including

- Analog <http://www.analog.cx/>
- Webalizer <http://www.mrunix.net/webalizer/>
- AWStats <http://awstats.sourceforge.net/>

One of the drawbacks to Web log analyzer software is that it requires access to logs from the server. This can be difficult if your library is a branch of a large library or multi-departmental system with central IT support providing Web hosting. Librarians wishing to perform detailed or custom Web log analysis should consult with their Web site hosts about log formats and the availability of log data. Some hosts will offer Web-based reporting systems. Work with the server administrator to be sure that the reports will be configured in a format useful to answer your questions

At Memorial University of Newfoundland (MUN) the Web server is housed in the library, and the commercial WebTrends software package is used to generate Web page statistical reports. The following report examples are all generated from WebTrends, although the features shown here are common to many Web log analysis applications. Note that the data contained in the following tables and reports may have been truncated or altered. The data provided does not necessarily represent actual activity for any MUN library service and is included only for the purpose of example.

Interpreting General Summary Statistics: Terms and Definitions

It is not difficult to generate numbers about Web site usage to add to annual reports or to requests for increased funding. It is somewhat more difficult to create actionable metrics for the library Web site. A ‘metric’ is any type of measurement used to gauge some quantifiable component of an organization’s performance. A ‘key performance indicator’ is a metric tied to an objective set by the organisation. Web log analysers provide numbers – lots of them. They do not, however, indicate which of those numbers is meaningful in a particular environment, nor do they help to set objectives for the improvement of electronic services. Understanding the way in which this data is generated can help to develop key performance indicators centered on the goals of your own organisation.

Elements in the General Statistics Report

	Hits	**Entire Site (Successful)**	**19,364,751**
		Average per Day	**212,799**
		Home Page	**0**
	Page Views	**Page Views**	**832,949**
		Average per Day	**9,153**
		Average per Unique Visitor	**4**
		Document Views	**305,827**
	Visits	**Visits**	**530,214**
		Average per Day	**5,826**
		Average Visit Length	**0:18:04**
		Median Visit Length	**0:00:01**
		International Visits	**56.14%**
		Visits of Unknown Origin	**4.00%**
		Visits from United States	**39.84%**
		Visits Referred by Search Engines	**128,744**
	Visitors	**Unique Visitors**	**178,065**
		Visitors Who Visited Once	**139,072**
		Visitors Who Visited More Than Once	**38,993**

Table 1-1 - General Statistics Report (Jan. – Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

Notice that the report in Table 1-1 contains several levels of information about site activity. The first section, 'Hits,' measures every single resource requested from the server. This includes requests for images, menus, and style sheets that may be embedded in any given page. This number is usually artificially inflated and should not be taken as an indicator of pages viewed. The element labelled 'Home Page' inaccurately reports no activity, because the server administrator failed to include the URL of the library home page when configuring the reporting software.

The second section contains two separate measures: 'Page Views' and 'Document Views.' These elements are a much closer approximation of actual resource usage on the site. In this case 'Page Views' counts hits on all documents, including forms and dynamic pages. Supporting graphics and other non-page files are not counted. WebTrends considers any page retreived with either a POST command or a GET command with a question-mark ('?') to be a dynamic page. The 'Document Views' element is defined according to locally-configurable criteria established by the server administrator. Typically 'Document Views' counts hits on complete HTML pages, PDF documents, and other static content retrieved with a GET command. It does not include form submissions, nor does it count dynamic pages that require user input (e.g., a search query) to construct results from a back-end database. Both 'Page' and 'Document' views are susceptible to undercounting, because many Web browsers cache frequently viewed pages. These pages are then served back to the user from his or her own hard drive, rather than being requested from the server, so no evidence of cached views exists in the server log.

The third section indicates statistics for 'Visits.' In this report group the log analyser attempts to isolate specific session information. A visit may be comprised of a single page view or of many pages viewed by an individual from a particular IP address without an idle time of more than 30 minutes occuring between views. The idle time can be configured according to preference.

The fourth section, 'Visitors,' records the number of single or repeat visits according to the IP address of the visiting computer. Be careful when using an IP address as a measure of individual visitors to your site. If, for example, there a large number of public stations from which people can access your Web-based resources, these station IPs will appear frequently in the logs. Statistical analysis based on IP address will show these as repeat visitors, even though a different person may have been using that station for each session.

The opposite problem is true of users who access the site from a public ISP such as AOL or RoadRunner. In most cases home broadband providers do not assign a static IP to each computer on the network, but assign an IP from a pool each time a connection is made. In some cases the ISP will channel many users through a proxy server, and the IP appearing in your log will therefore be that of the proxy, and not of the end user computer. This means that a user may visit from home on a regular basis, but cannot be identified as a repeat visitor, because the IP trace left in the server log will be different each time that person visits your site.

Many Web site administrators implement tracking cookies to solve this latter problem. A cookie is a small file written to a patron's computer during a visit to a Web site. Web sites sometimes use cookies to personalize the information seen by a user during a given Internet session, or during subsequent sessions. In order to track repeat visitors who do not have static IPs, a persistent cookie can be served from the site. A persistent cookie is one that never expires and stays in the patron's cookie folder until he or she chooses to delete it. Each time that patron visits the site, the value held in the cookie is incremented to indicate a repeat visit. The cookie information is sent back to the server, and it can be logged to help identify unique visitors who may not have a unique or stable IP address.

Cookies are often perceived as a privacy threat, and browsers can be configured to not accept cookies of various types. First party cookies that do not collect personal information (such as those left by the library server when a patron views the site) are accepted by most Web browsers. Cookie information can be read only by the server that sets the cookie, so there is no risk that a cookie from the library Web server can be read by other Web sites.

Who are They? Geographic Location and Organisational Membership of Site Visitors

The information about 'International visits' in the third section of Table 1-1 should be examined more closely, as this particular report has been configured to assume that visits from countries other than the United States are international. The administrator should configure these reports to indicate the country where the library is situated. Further

geographic breakdowns are available to identify activity from specific countries, states, or provinces.

Most Active Countries

	Countries	Visits
1	Canada	245,429
2	United States	211,256
3	Australia	7,998
4	United Kingdom	5,684
5	Europe	5,446
6	France	3,214
7	China	2,138
8	Netherlands	1,988
9	United Kingdom	1,985
10	Germany	1,930

North American States &Provinces

	State	Visits
1	Newfoundland	167,543
2	Ontario	56,215
3	Virginia	38,073
4	California	14,269
5	New Brunswick	3,152
6	Nova Scotia	2,981
7	Alberta	2,943
8	Massachusetts	2,915
9	New York	2,868
10	British Columbia	2,590

Table 1-2 - Most Active Countries and North American States and Provinces Reports (Jan. – Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

Table 1-2 demonstrates more detailed reports where we see a high level of activity from various Canadian provinces, as one might expect at a Canadian library. How do we account for the large amount of traffic out of Viriginia? The 'Most Active Cities' report in Table 1-3 shows that the usage is predominantly from Herndon, Virginia.

Most Active Cities

	City, State, Country	Visits
1	St Johns, Newfoundland, Canada	162,458
2	Ottawa, Ontario, Canada	45,841
3	Herndon, Virginia, United States	35,868
4	Mtn View, California, United States	4,533
5	Toronto, Ontario, Canada	4,530
6	Marina Del Rey, California, United States	4,349
7	Milton, Australia	4,068
8	Halifax, Nova Scotia, Canada	2,221
9	St John, New Brunswick, Canada	1,783
10	Calgary, Alberta, Canada	1,652

Table 1-3 - Most Active Cities Report (Jan. – Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

The high-speed Internet provider RoadRunner has its company headquarters in Herndon, Virginia. These log hits occur because Internet traffic from across North America is channeled through Herndon. Other reports can help to futher untangle the provenance of site visitors.

Most Active Organizations

	Organizations	Hits	% of Total Hits	Visits
1	library.mun.ca	4,872,809	25.16%	44,563
2	aliant.net	1,972,426	10.18%	24,617
3	nf.net	1,483,644	7.66%	14,016
4	med.mun.ca	619,718	3.20%	15,022
5	pcglabs.mun.ca	547,373	2.82%	9,813
6	wsr.mun.ca	321,116	1.65%	7,949
7	googlebot.com	293,935	1.51%	16,737
8	comcast.net	238,617	1.23%	6,557
9	rogers.com	235,181	1.21%	5,096
10	wst.mun.ca	168,942	0.87%	4,165

Table 1-4 - Most Active Organizations Report (Jan.- Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

By the translation of the IP addresses of visitors into DNS names, log analysers can help group them according to the domain extensions of their ISPs (Table 1-4). In this case, it is apparent that many of the visitors to the library's Web site come from within the 'mun.ca' domain, and are, therefore, members of the university community. More granular information is also available, as demonstrated by Table 1-5, where details of site activity are listed according to the host name of each frequent visitor:

Top Visitors

	Visitor	Visits	Hits	% of Total Hits
1	vhost.ucs.mun.ca	2,134	6,362	0.03%
2	crawler.bloglines.com	2,120	4,187	0.02%
3	altair.ucs.mun.ca	1,474	1,596	0.00%
4	msnbot.msn.com	1,074	33,482	0.17%
5	bentley.mha.mun.ca	945	1,203	0.00%
6	wiley-411-2130.roadrunner.nf.net	843	1,165	0.00%
7	egspd42141.ask.com	705	15,315	0.07%
8	Hslcircbehind.med.mun.ca	696	54,905	0.28%
9	bastion.hccsj.nf.ca	629	90,674	0.46%
10	med-sur1831b.med.mun.ca	596	1,040	0.00%
11	public01.med.mun.ca	589	14,360	0.07%
12	lib-mason.library.mun.ca	583	18,291	0.09%

Table 1-5 - Top Visitors Report (Jan. – Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

Another issue becomes clear from this report: many of the site's most frequent visitors are not people at all, but are Web bots and spiders that crawl the site to provide indexing information to Web search engines. Examples from Table 1-5 include "crawler.bloglines.com" and "msnbot.msn.com". While these visitors are a welcome means to increase site visibility in search engines, these hits should not be included in use data reports. Spiders often repeat visits to the same sites, and they look at many pages very quickly, generating a huge number of hits on each Web site. The report on visits and

hits from spiders (Table 1-6) can help to isolate spider activity, so one can be sure not to include this data in patron statistics.

Visiting Spiders

	Spider	Visits	Hits	% of Total Hits
1	Mozilla/5.0 (compatible; Yahoo! Slurp)	48,073	52,454	12.01%
2	Googlebot	15,822	101,427	23.24%
3	Mozilla/5.0 (compatible; Yahoo! Slurp China)	1,013	1,044	0.23%
4	Mozilla/5.0 (compatible; Googlebot/2.1l)	894	177,967	40.77%
5	Gigabot	777	2,455	0.56%
6	Yahoo-MMCrawler	240	2,901	0.66%
7	http:	205	502	0.11%
8	Baiduspider (http:	187	188	0.04%
9	gsa-crawler (Enterprise; GIX-03519)	141	286	0.06%
10	Mozilla/4.0 (compatible; MSIE Crawler)	94	1,917	0.43%
	…	…	…	…
	Total For Spiders Above	68,454	435,090	99.69%

Table 1-6 - Visiting Spiders Report (Jan. – Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

When Do They Come? Time Dimensions of Web Server Activity

Elements in the Activity by Time Summary Report:

Summary of Activity for Report Period	
Average Number of Visits per Day on Weekdays	6,388
Average Number of Hits per Day on Weekdays	239,871
Average Number of Visits per Weekend	8,843
Average Number of Hits per Weekend	290,239
Most Active Day of the Week	Tue
Least Active Day of the Week	Sat
Most Active Date	16-Mar-06
Number of Hits on Most Active Date	426,614
Least Active Date	1-Jan-06
Number of Hits on Least Active Date	57,945
Most Active Hour of the Day	13:00-13:59
Least Active Hour of the Day	04:00-04:59

Table 1-7 - Activity by Time Summary Report (Jan. – Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

This report (Table 1-7) aggregates activity according to the Date/Time stamp in the Web log. This information may be used to inform decisions on chat and e-mail reference service availability, library and service desk hours of operation, opportunities for server maintenance, and windows for running processor intensive tasks, such as back-ups or large analysis reports like those shown here. Each of the temporal elements can be examined to show activity summaries by week, day of the week, and hour of the day.

Date/Time analysis, combined with session information based on a user's IP address, SessionID, or cookie information, can provide data on the length of user visits and the duration of page views. Many hits will be less than a minute in length, if bot activity has not been filtered out of the report, or if many public or staff stations are set to show the library home page as their browser default. Short session times are also normal when patrons are using one specific page on the site to gain access to external resources (e.g., electronic indexes and journals). If the Web site, library catalogue, metasearch interface, and resolver interface reside on different Web servers, then patrons using the Web site as an entry point to other tools will also show very short visit durations, even though they are continuing to use library resources during their session.

If the library houses a large number of public or kiosk type stations, there may be many extremely long sessions. Stations used by many individuals can skew use data, as several consecutive patrons can use the same machine to view the Web site. Unless an idle-time period of 30 minutes occurs between these sessions, the report is configured to assume that that all of the activity comprises a single visit.

Another metric used by commercial Web sites is the number of pages viewed per session, which indicates whether visiting patrons usually look at only a single page on the site or follow a path through many pages. Unlike commercial enterprises, which usually want to keep people on their own sites as long as possible, library Web sites are often designed to function as portals to other sites and services. For libraries short session durations and low numbers of page views per sessions may be common and are not necessarily undesirable. Metrics designed around user session time and the number of pages viewed in a session are, therefore, difficult to interpret. Did the user spend a lot of time on the site because he couldn't find the correct information, or because he found a useful link to another site? Did he view many pages because navigation of the site was difficult for him, or because there were many pages relevant to his information need?

What do they want? Resources Accessed, Search Strategies

There are data sources that help assess the reasons that patrons are visiting the library Web site, including reports on the resources or groups of resources accessed most often. Top *n* analysis can be run against the whole site, or against resources grouped together in specific directories.

Table 1-8 represents usage on the directory containing the library's user guides.

Top Pages: Guides				
Pages		**Views**	**% Total Views**	**Avg. Time Viewed**
1	**/guides/howto/annotated_bibl.php**	**1,054**	**12.10%**	**00:01:53**
2	**/guides/howto/ write_book_review.php**	**450**	**5.17%**	**00:02:35**
3	**/guides/howto/turabian.php**	**312**	**3.58%**	**00:04:17**
4	**/guides/howto/apa.php**	**229**	**2.63%**	**00:02:38**
5	**/guides/howto/mla.php**	**134**	**1.53%**	**00:04:58**
6	**/guides/howto/evaluation.php**	**126**	**1.44%**	**00:02:00**
7	**/guides/howto/tips.php**	**155**	**1.78%**	**00:02:38**

8	/guides/howto/offcampus.php	188	2.15%	00:03:07
9	/guides/howto/primary.php	98	1.12%	00:02:52
10	/guides/howto/reserve.php	130	1.49%	00:01:51

Table 1-8 - Top *n* Pages Report - Guides (Jan. – Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

This report helps to identify those guides that are in heavy demand and the average amount of time spent viewing each page. The 'time viewed' variable may help to determine whether or not a patron has found the resource useful. In the case of guides, it is likely that a view of 5 seconds means that the document did not meet the user's information need, but a view of 30 seconds could indicate that the patron found the document useful enough to print for reading at a later date. Of course, just because a page is open in the user's browser, it is still not possible to determine if she is engaged with the information on the page.

In addition to specific page view data, it is possible to generate a top directories report, which provides a useful snapshot of the areas of the site accessed most frequently. The first entry in Table 1-9 indicates hits on the root directory, the main directory where the homepage for the site is located. The root directory is represented by a slash (/).

Top Directories

	Path to Directory	Visits	Hits
1	/	272,117	918,972
2	/guides/howto	48,991	135,371
3	/eindex	45,288	296,900
4	/hsl/bates	40,551	339,321
5	/eindex/images	37,000	559,921
6	/qeii/cns	32,233	131,857
7	/hsl/images	32,186	126,310
8	/hsl/docs	29,334	61,378
9	/hsl	27,517	112,816
10	/swgc/music	22,366	180,713

Table 1-9 Top Directories Report (Jan – March, 2006). Generated from WebTrends Analysis Suite v. 7.

Just knowing that patrons are viewing pages or collections doesn't provide enough information to determine whether or not the information that they find there has met their needs. One clear indicator that a patron found a resource useful is the number of times that a document is downloaded, meaning that it has been saved to the patron's local computer. Presumably a patron who has bothered to download a document from your Web site has found a resource that she considers relevant and useful.

Most Downloaded Files

	File	No. Downloads
1	http://www.library.mun.ca/guides/howto/apa.pdf	2,199
2	http://www.library.mun.ca/guides/howto/mla.pdf	1,709
3	http://www.library.mun.ca/guides/howto/annotated_bibl.pdf	1,570
4	http://www.library.mun.ca/qeii/holidays.pdf	945
5	http://www.library.mun.ca/guides/howto/music_citations.pdf	717
6	http://www.library.mun.ca/qeii/cns/waterpoweredsawmills.pdf	880
7	http://www.library.mun.ca/guides/howto/turabian.pdf	419
8	http://www.library.mun.ca/guides/howto/offcampus.pdf	394
9	http://www.library.mun.ca/hsl/ guides/CINAHLsearchguide.pdf	424
10	http://www.library.mun.ca/qeii/maps/1-138b.pdf	472

Table 1-10 - Most Downloaded Files Report (Jan. – Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

Frequently accessed files may give an indication of a patron's information needs, but another important source of data is contained in search strategies, where the patron's own words express his information needs. When a search engine refers a patron to a site, the query portion of the referrer URL contains the words that were entered in the patron's original search. This data can be aggregated in a number of ways, including a basic report that shows the search phrases that most commonly lead a user to the library site.

Top Search Phrases

Phrases	Phrases found	% of Total
how to write an annotated bibliography	1,926	1.49%
how to write a book review	1,711	1.32%
newfoundland map	1,042	0.80%
mun library	937	0.72%
how to write citations	651	0.50%
how to write bibliographies	642	0.49%
queen Elizabeth ii	539	0.41%
free scores	376	0.29%
newfoundland newspapers	333	0.25%
male physical exam	296	0.22%
ship drawings	248	0.19%

Table 1-11 - Top Search Phrases Report (Jan .– Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

Referrals from search engines on the Web, however, do not necessarily indicate how patrons are using the Web site. Search engine users did not choose the library Web page as a starting point for their research, even if that site meets their information need. Strategies entered into a 'site search' box on the library Web page provide more information about the kinds of information being sought by patrons, and may also help to identify pages that need to be more visible on the site. Table 1-12 is an example of a Perl report, with aggregated search strategies from the library Web site.

Inktomi Search Results

Search Results for January - March 2006

Created on
6/16/2006

Search Criteria	SearchResult
17.03.001	71
APA	16
jstor	15
reserves	15
wireless	15
pubmed	14
annotated bibliography	13
cisti	13
MLA	12
hsc	12
first space	11
reserve	11
z39.50	10
apa	9
evidence based practice	9
nl railway	9
uMemorial University Libraries	9

Table 1-12 -Top Search Our Site Requests (Jan. – Mar., 2006). Generated from Web server logs using a custom Perl script.

Any dynamic search box on a Web site that searches a local backend database will also provide query strings in the Web server log. Table 1-13 is a Perl-based report that indicates the most frequent selections from the subject and A-Z search menus on the library's electronic index site.

Eindex Search Results

Report for the Month of January - March 2006 **Date Printed:6/20/2006**

subhead=nursing	4,881
searchtext=w	1,466
subhead=business	1,314
searchtext=p	1,300
subhead=education	1,249
subhead=psychology	1,150
subhead=biology	908
subhead=english+language+and+literature	743
searchtext=j	671
subhead=history	574
subhead=sociology	510
searchtext=c	472
subhead=folklore	418
subhead=health+sciences	402
searchtext=g	395
subhead=kinesiology	379
subhead=newfoundland+and+labrador	356
searchtext=e	338
subhead=medicine	303
subhead=biochemistry	296
subhead=geography	279
searchtext=a	273
subhead=english%20language%20and%20literature	265

Table 1-13 - Top Electronic Index Subject Selections Report. (Jan – Mar. 2006). Generated from Web server logs using a custom Perl script.

Where Did They Come From? Referrers

If the server administrator has configured a Web server to log referring URLs, it is possible to see which sites are driving traffic to the library Web site. Referrer data can be loosely grouped according to the referring site (Table 1.14).

Top Referring Sites

	Site	Visits
1	No Referrer	230,388
2	http://www.library.mun.ca	86,325
3	http://www.mun.ca	57,071
4	http://www.google.com	38,622
5	http://www.google.ca	23,225
6	http://images.google.com	7,694
7	http://www.swgc.mun.ca	6,853
8	http://profile.myspace.com	6,771
9	http://www.google.co.uk	4,158
10	http://qe2a-proxy.mun.ca	3,903
11	http://images.google.ca	2,185
12	http://thecommons.mun.ca	1,734

Table 1-14 - Top Referring Sites Report (Jan. - Mar., 2006). Generated from WebTrends Analysis Suite v. 7.

This report helps to determine the number of patrons who came to the site through search engines (e.g., Google in line 4), those who were following links found on other Web sites that may be part of the institutional Web presence (e.g., <http://thecommons.mun.ca> in line 12), and those who came to the site through pages hosted outside of your own institution (e.g., MySpace in line 8). If few visitors arrive via major search engines, then it is possible to experiment with the use of different 'meta-tag' keywords and descriptions in page headers to try to increase the site's visibility in search engines.

It is also possible to generate reports that show the full address of the referring pages. As patrons follow click-paths through a Web site, the referrer information indicates which pages are generating a lot of internal clicks to other resources on the Web site.

Table 1-14 indicates a large number of visits with 'no referrer.' A referrer is only registered in the server log when a user clicks through to the site using a link. Referrers are not registered when a patron types the Web site URL into a browser, when the page is loaded as the browser home page, or if a patron uses a bookmark to access the site. In some cases the referrer information is also lost if the link from another Web page opens in a new window.

Where Did They Go?

Reports on top exit pages will provide clues as to why patrons left the site. Did a patron leave because she couldn't find what she was looking for, or did she leave because she found a link to a resource which met her information need? Ideally, the top exit pages on

your site are those which provide links to recommended services and resources such as the catalogue entry page, lists of useful external resources, or e-journal and e-index search pages.

One of the most frustrating aspects of Web log analysis is the inability to tell where a patron has gone once he has left the site. Library Web sites are often designed to help users find information, resources, and services that exist in other places. It is difficult to determine whether this objective has been met, if one cannot see the resource selections that users make from the library Web site. Additionally, libraries invest significant amounts of money to license electronic reference materials including indexes, journals and other content that does not reside on library servers. Measuring the use of this material is important for on-going cost-benefit evaluation of subscribed electronic content.

Although standard Web log analysis cannot provide information about the external links selected by patrons, there is a way to track this information through the use of small programs that log and redirect requests for external links. These programs can be written in languages such as PHP, ASP, or CGI/Perl. External link-tracking is implemented by changing all of the URLs on a Web site to point to the redirect program on the local server, rather than pointing the user directly to a remote resource. Following are examples of two ways to use HREFs to redirect users to local servers:

- HREF pointing to an external site:
  - **<A HREF="http://agricola.nal.usda.gov/">AGRICOLA Citation Index<A>**
- HREF pointing to a CGI/Perl script named 'getit' on your own server.
  - **<A HREF="/cgi/getit/http://agricola.nal.usda.gov/">AGRICOLA Citation Index</A>**

When the latter link is invoked, the URL of the remote resource is sent to the CGI program. The CGI can log information about the request, including the date/time of the request, the page from which a user clicked the link, and the URL of the remote resource. The CGI then redirects the user to the requested URL. The CGI may store logged information in a delimited text file, or in an SQL-compliant database. Similar CGIs can be added to URLs in MARC records to help track resource usage from the library catalogue.

There are two drawbacks to tracking external links selected from the library Web site. The first is that all URLs to external resources will have to be re-written to point to your CGI, instead of directly to the URL. The second is that the CGI adds a fractional delay in the time that it will take patrons to see the external page which has been requested. Performance hits are best measured in a specific environment, as they have to do with the hardware resources available on the Web server, the amount of information logged by the CGI, and the amount of traffic on a site. If a link checker is used to find broken URLs, then ensure that the CGI returns appropriate error codes when it cannot contact a site.

Page Tagging and Traffic Analysis

If Web site logs are not easily accessible, but you have access to the source code, e.g., 'hosted solution,' then traffic analysis programs may be the solution. These include:

- Google Analytics <http://www.google.com/analytics/>
- OneStat <http://www.onestat.com/>
- StatCounter <http://www.statcounter.com/>

Each of these applications relies on client-side data collection through JavaScript and cookies. These programs require that a small piece of JavaScript code is added to each page that will be tracked. In Web analytics literature the process of adding this JavaScript to Web pages is called 'page tagging,' (not to be confused with 'tagging' offered by social bookmarking sites such as <del.icio.us>). The Google Analytics 'tag,' for example, looks like this:

```
<script src="http://www.google-analytics.com/urchin.js" type="text/javascript"></script>
<script type="text/javascript">
      _uacct = "<tracking number goes here>";
      urchinTracker();
</script>
```

Once the JavaScript has been added to the page header, activity can be tracked by signing into a Web-based interface that allows the generation of reports, charts, and graphs that contain information about site traffic. The type and level of analysis possible will depend on the chosen vendor solution, but most of the types of reports that are generated through Web log analysis are also available from page tags.

Analysis performed through page tagging is close to real-time, so a snapshot of site activity is quickly available. It is possible to monitor take-up of a new service on the day it is launched, or to watch the effects of a publicity campaign or a policy change from its inception.

One of the main advantages of page tagging is that the JavaScript code runs every time the page is loaded, whether from the server or from a browser cache. This means that usage information is updated even when a user looks at a cached page. In some cases tagging solutions can collect data that cannot be found in Web logs. Tags store data in cookies that can be configured to log additional information, such as JavaScript and Flash events triggered as a user navigates through the site. Another advantage is that the software used to view and analyze traffic is often hosted on a third-party server, so technical expertise is not required in the administration of these reports.

Some of the drawbacks of page tagging include the fact that the JavaScript header has to be added to every page that is tracked. 'Server-side includes' are a means by which a common piece of code can be inserted into Web pages as they load. If a site has server-side includes that put header information on pages, then it may be a simple matter of adding the JavaScript to the single header file read into all pages. If a site does not have

includes for header information, then some manual labour will be required to add this information to each page of the Web site. The execution of JavaScript code is performed by a users' web browser ,' and although it may not be detectable to the casual user, Javascript creates a slight lag in page load time. Some users may have JavaScript disabled in their browsers, in which case the appropriate information cannot be collected. 'Page-tag vendors' may also rely on third-party cookies that are frequently blocked in modern browsers. Third party cookies are those which do not originate from the server on which the web page is located, but are pulled in from a web server at a different location. This renders the cookie information available to the third part server, and so can constitute as a privacy threat. It is preferable to select a vendor that supports first-party cookies.

Due to storage restrictions on the remote server, data may not be available on the vendor site indefinitely. Data may also disappear if the page tag subscription with that vendor is terminated. Ideally the data store can be periodically exported to a local server for long-term storage and additional analysis.

Other concerns for libraries include privacy issues. In most tagging solutions, Web site traffic data is held on the server of the page tag vendor. It is important to understand the vendor's privacy policy up front, including whether or not the data may be used for purposes other than your own. Privacy laws are applied according to the country in which the vendor server is located.

Part I of this article has introduced concepts which will be important for any librarian who wishes to engage in webserver log analysis. These examples have been developed around the library's main website, however there are many other library resources which produce server logs. In Part II the author provides an overview of server log data from library-specific applications including proxy servers, link resolvers, and Integrated Library System (ILS) servers.

Resources

Beitzel, Steven, Eric Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. "Hourly Analysis of a Very Large Topically Categorized Web Query Log." Proceedings of SIGIR (Special Interest Group in Information Retrieval), July 25–29, 2004, Sheffield, South Yorkshire, UK. 321-8.

Breeding, Marshall. "Analyzing Web Server Logs to Improve a Site's Usage," Computers in Libraries. 25.9 (Oct. 2005):26, 28-9.

Carter, David S. "Web Server Transaction Logs: Dave's ULA Project." 7 Mar. 1996. 25 Feb. 2007 <http://www-personal.umich.edu/~superman/AP/>.

Clifton, Brian. Whitepapers: Web Traffic Data Sources & Vendor Comparison. Omega Digital Media Ltd. 7 Dec. 2006. 25 Feb. 2007 <http://www.ga-experts.co.uk/Web-data-sources.pdf>.

Coombs, Karen A. “Using Web Server Logs to Track Users Through the Electronic Forest,” Computers in Libraries 25.1 (Jan. 2005): 16-20.

Cram, Jeff. “Building a Web Site for Analytics,” Digital Web Magazine 16 Oct. 2006. 25 Feb. 2007 <http://www.digital-Web.com/articles/building_a_Web site_for_analytics/>.

Davis, Philip M. “Information-Seeking Behavior of Chemists: A Transaction Log Analysis of Referral URLs.” Journal of the Association for Information and Library Science and Technology (2004) 55.4: 326-32.

Eisenberg, Bryan. “Accurate Analytics Require Cookies,” ClickZ Networks 5 Mar. 2004. 25 Feb. 2007 <http://www.clickz.com/showPage.html?page=3319891>.

Flaherty, P. “Transaction Logging Systems: A Descriptive Summary,” Library Hi Tech 11.2 (1993): 67-78.

Jasra, Manoj. “Web Analytics Comparison – Google vs. VisiStat.” Web Analytics Association. 16 Aug. 2006. 9 Nov. 2006 <http://www.Webanalyticsassociation.org/en/art/?119>.

Kurth, Martin. “The Limits and Limitations of Transaction Log Studies,” Library Hi Tech 42 (1993): 98-104.

Menasalvas, Ernestina, Socorro Millán, José M. Peña, Michael Hadjimichael, Oscar Marbán. Subsessions: A Granular Approach to Click Path Analysis,” International Journal of Intelligent Systems 19(2004): 619-37.

Sandor, B. “Applying the Results of Transaction Log Analysis,” Library Hi Tech 11 (1993): 87-97.

Schwartz, Randal L. “Clicking-Through Tracking in Perl,” New Architect (May 1998). 25 Feb. 2007 <http://www.Webtechniques.com/archives/1998/05/perl/>.

Sterne, Jim. Web Metrics: Proven Methods for Measuring Web Site Success. NY: Wiley, 2002.

Web Analytics Tutorial. Summary.Net (18 Apr. 2002). 25 Feb 2007 <http://www.summary.net/manual/tutorial/toc.html>.

Software:

WebTrends 7.0, c2004. WebTrends, Inc., Portland, OR. 12 February 2007 <http://www.webtrends.com/>.

ClickTracks Optimizer 3.0 c2005. ClickTracks Analytics, Inc. Santa Cruz, CA 12 February 2007.

<http://www.clicktracks.com/>.

Analog 6.0 c.2004. Written by Stephen Turner**.** Released under the GNU General Public License. February 12 2007.
<http://www.analog.cx/>

Webalizer 2.01 c.1997-1999. Written by Bradford L. Barrett . Released under the GNU General Public License. Februrary 12 2007.
<http://www.mrunix.net/webalizer/>

AWStats 6.6 c 2006. Written by Laurent Destailleur. Released under the GNU General Public License. February 12 2007.
<http://awstats.sourceforge.net/>

Google Analytics c.2006. Google Inc. Mountain View, CA . February 12 2006.
<http://www.google.com/analytics/>

OneStat Enterprise 3.0 c2006. Onestat.com, Netherlands. Febrarury 12 2006.
<http://www.onestat.com/>

StatCounter c2005. Statcounter.com, Dublin, Ireland. February 12, 2006.
<http://www.statcounter.com/>