7-14-2005

Using Server Logfiles to Improve Website Design

Elaine Maytag Nowick
University of Nebraska-Lincoln, enowick@unl.edu

***Library Philosophy and Practice* Vol. 4, No. 1 (Fall 2001)**
(www.uidaho.edu/~mbolin/lppv4n1.htm)
ISSN 1522-0222

Using Server Logfiles to Improve Website Design

Elaine Nowick
Associate Professor
Branch Services
University of Nebraska--Lincoln Libraries
P.O. Box 880410
Lincoln, NE 68588-4100

Introduction

To provide appropriate web-based information in an accessible form, librarians need to know how users visualize the organization of information within the system and how they search for the information they need. The Internet has been compared to a very large library before classification schemes were devised. The amount of information is incredible, but trying to find relevant answers to questions is a frustrating experience. The Agricultural Networked Information Center (AgNIC) is an effort by a coalition of land-grant universities and the National Agricultural Library (NAL) to provide a user-friendly guide to reliable information on the Internet on all aspects of agriculture. The University of Nebraska-Lincoln (UNL) joined AgNIC in September of 1996, providing a website serving as a clearinghouse to Internet sites useful for plant scientists and extension researchers. The UNL AgNIC Plant Science website is located at: www.unl.edu/agnicpls/agnic.html. The website consists of a main page listing topics within the field of plant science, such as crop statistical information, genetics information, and information on plant pests. These topics are then linked to subpages with lists of relevant websites. Short descriptions are included with each website listed. Another important feature of AgNIC allows users to access a reference service.

In developing UNL's plant science website, feedback and advice on website design was sought from other members of the AgNIC coalition and from plant scientists at UNL. Since this website is publicly accessible, most users are from outside of the UNL community. One way to learn more about these users is from the server logfiles.

Logfiles provide a record of communications between the server and computers requesting information residing on the server. When a user accesses a website, the browser program on the user's computer communicates with the server computer that hosts the website. The server sends the requested file to the user's computer and it then appears as a web page on the user's screen. The information coding each linked page is sent as a separate file. When the user downloads a file, the host server records the address of the computer to which the file was sent, the date and time of the request, and the name of the file sent. The logfile can be analyzed to tell something about the user, what subpages were viewed and in which order, which in turn indicates the navigation patterns of the user. Logfile analysis has been suggested as a way to understand user behavior in hypertext systems (Barab, 1996), to improve OPAC use efficiency (Connaway, et al. 1995), and to improve website design (Suleman, et al., 2000).

The purpose of this study was to identify ways that the design of the AgNIC Plant Science website could be improved by analyzing server logfiles to determine the types of users who accessed the site, the kinds of information they were seeking, and how they navigated within the site.

Methods

Logfiles for the UNL AgNIC Plant Science website were analyzed for two time periods: October 2, 1996-November 1, 1996 and December 11,1996-January 28, 1997 (Time 1) when the site first was published and July 30, 1997-October 2, 1997 (Time 2). These time periods were used because logfiles were available. The total number of users, the number of users per day, and the average number of users for each day of the week were examined for the two time periods.

IP address	date/time	file sent
xxx-xxxxxxx.unl.edu	--[15/Aug/1997:09:46:47-0500]	“GET/agnicpls/agnic.html HTTP/1.0”200 219

Figure 1. Sample Logfile Entry

A sample logfile entry is shown in Figure 1. The requesting computer is identified by an IP address. There may be one or more sets of numbers or letters at the beginning of the IP address that identify the specific computer to which the file will be sent. In the sample entry these numbers are represented by x’s. The next set of letters (unl) indicates the institution that has registered the IP address, in this case the University of Nebraska-Lincoln. The last set of letters (edu) indicates the domain of the institution. The Internet is divided into domains or types of institutions: *edu* (educational), *gov* (government), *mil* (military), *net* (network), and *com* (commercial). For locations outside of the United Sates, the domain name will identify the country of location. The locations and registering organizations of American IP addresses were identified through InterNIC (www.internic.net). European addresses were checked through the European registration service (www.ripe.net) and Asian addresses through APNIC (www.apnic.net). The *net* and *com* domains overlap user types for the purposes of this study. For example, users from America Online (aol.com) are more like users of Navix (navix.net, a local Internet provider) than like users from a large seed company who would also have a *com* domain. For this reason, users were divided into types based on the registering institution rather than the domain name.

The number of times a file was downloaded, the numbers of files viewed at each IP address, and the combinations of files downloaded were analyzed. Accesses from the same address on separate occasions were also identified. Accesses from IP addresses belonging to “crawlers” were eliminated from the analysis, as were in-house accesses and those from

NAL. Accesses by America Online (AOL) and CompuServe users were not included in the location data analysis since the location of these users is always indicated as the location of the AOL or CompuServe proxy server.

Results

Number of Users

The total number of users and the number of users per day are shown in Table 1. As shown in this table, the average numbers of users per day more than doubled during the second time period. This increase is probably due to both an increase in the numbers of users on the Internet and to publicity for the website and the AgNIC project.

	Time 1	Time 2
Total number of observers	757	1385
Days observed	80	65
Average number of users per day	9.46	21.3

Table 1. Number of users, number of days included, and average number of users per day for the two time periods analyzed

Types of Users

The types of users for the two time periods are shown in Figure 1. The types of users were not divided by domain, but by the type of institution registering the IP address in order to have a clearer idea of the type of information sought. During the first time period, a quarter (25.1%) of the users were from IP addresses located at colleges or universities within the US. Another 8% were from universities in other countries. One percent were from K-12 schools. A total of 34% of the users were from educational institutions. Federal, state and local government users, including government research agencies, accounted for 8.1% of the total. Other research organizations made of 3.1% and, business users made up 4.1% of the total. The remaining users accessed the site through Internet providers, 3.2% from AOL subscribers, 0.5% from CompuServe, and 4.2% from local Internet providers or nets.

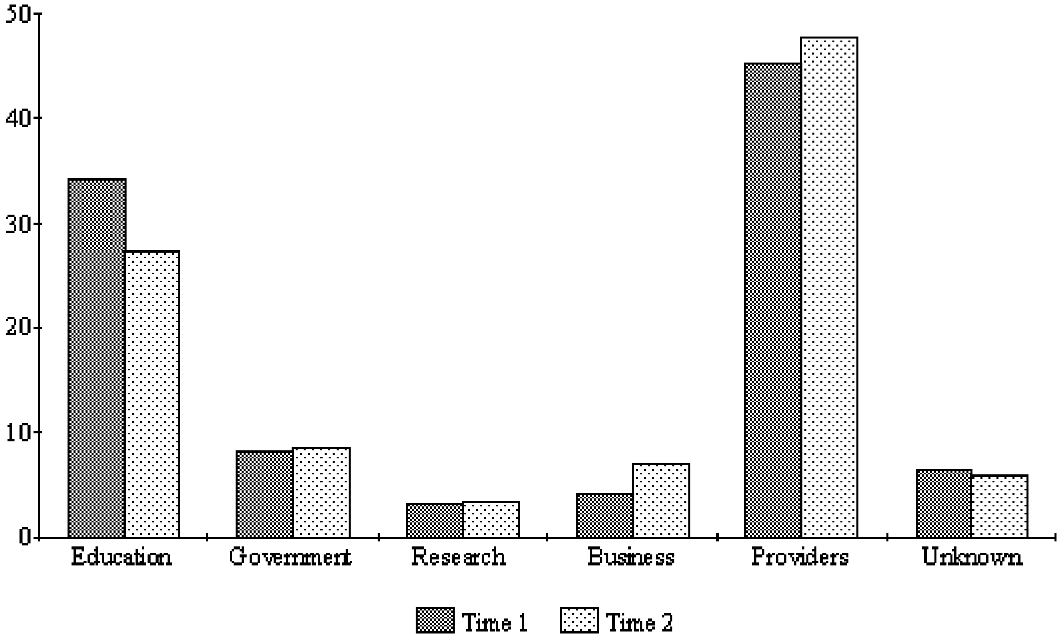


Figure 1. Types of Users Accessing Web Site (research users include non-university or non-government research institutes, providers include Internet providers such as AOL and local nets or providers).

During the second time period, the numbers of users from educational institutions increased in absolute numbers but fell in percentage to 27.3% of the total users. Government users increased slightly to 8.6%. Private companies and other research institutions were at 10.6% and Internet providers made up 47.6% of the users.

Locations of Users

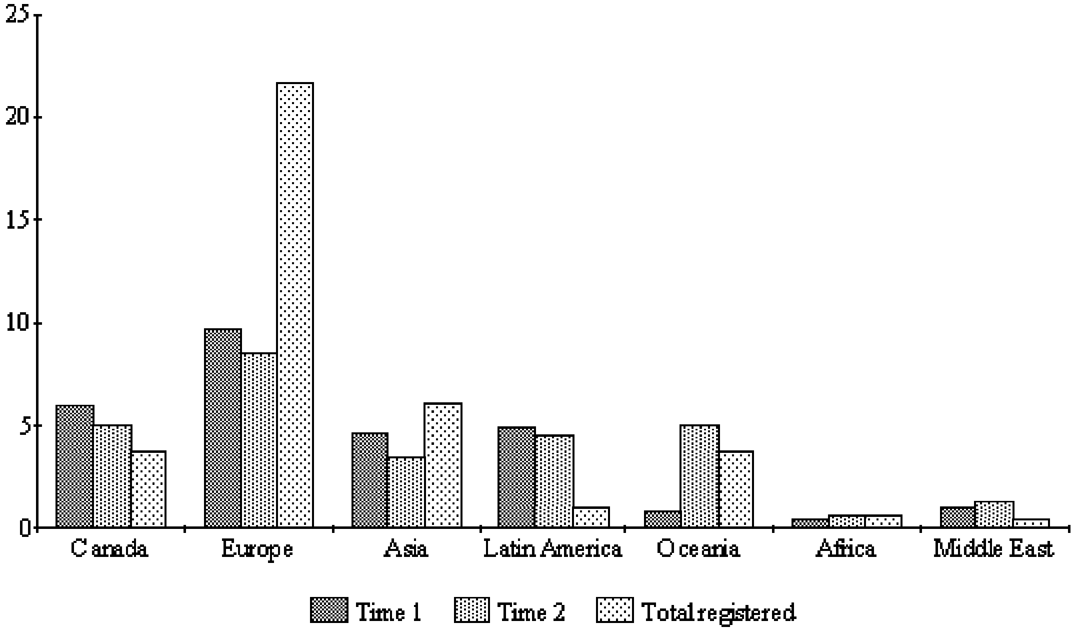


Figure 2. Distribution of users outside of the US. The total registered refers to the total number of IP addresses registered in a region.

During the first time period analyzed, 27.7% of the users were located outside of the United States. During the second time period the percentage of foreign users was 28.3%.

The percentages of IP addresses registered in regions of the world give an indication of Internet access in those regions. Slightly fewer of our users than would be expected, based on the number of IP address registrations, were from locations outside of the United States. Relatively more users were from Canada, Latin America, Australia and Oceania, and the Middle East, while fewer were from Europe and Asia than expected for the first time period studied (Chi-square=8.84 .025<P<.05, dof=3). In calculating the Chi-square value, data from Latin America, Oceania, Africa, and the Middle East were combined because of the low percentages (Snedecor and Cochran, 1967). The trend for more users from these areas was seen in both time periods studied and was more highly significant during the second time period (Chi-square=15.23, .005<P<.01, dof=3).

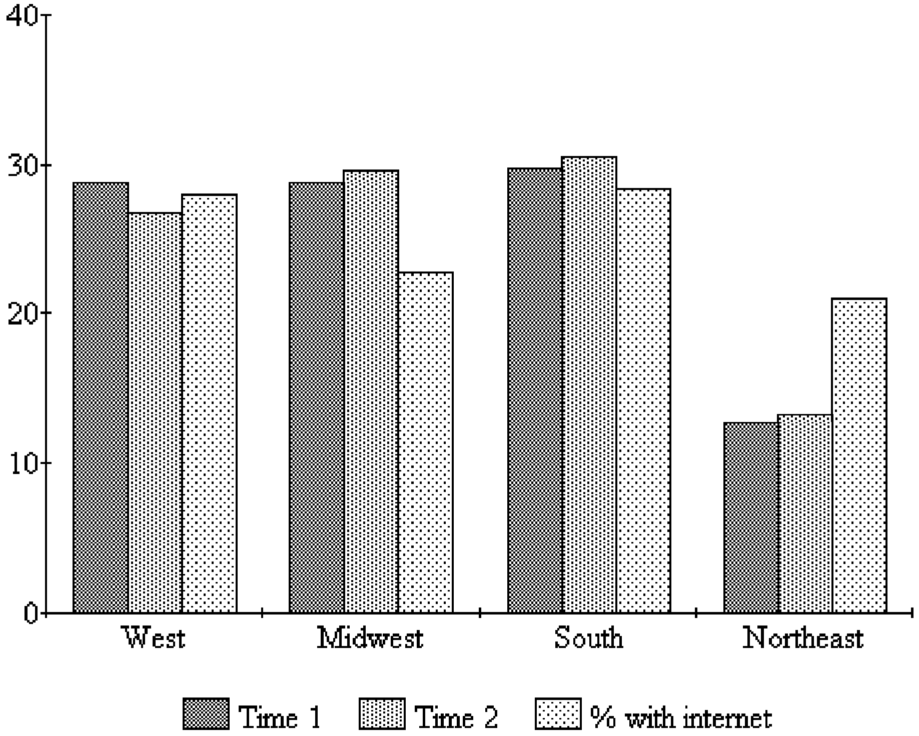


Figure 3. Distributions of users within the US. The percent with Internet refers to the numbers of adults in the region who have access to the Internet.

The distribution of users within the US is shown in Figure 3. In comparison to the percentage of adults with Internet access by region, there were somewhat more users of the site from the Midwest and fewer from the Northeast (*Statistical Abstract of the United States*, 1997). This trend was not statistically significant for either time period. For the first time period the Chi-square value was 4.9 ($.1<P<.25$, dof=3). For the second it was 5.2 ($.1<P<.25$, dof=3).

Accesses by Day of the Week

Another indication of the type of users is the variation in numbers of users accessing the site by day of the week (Figure 4). The average number of users was down over the weekend for both time periods.

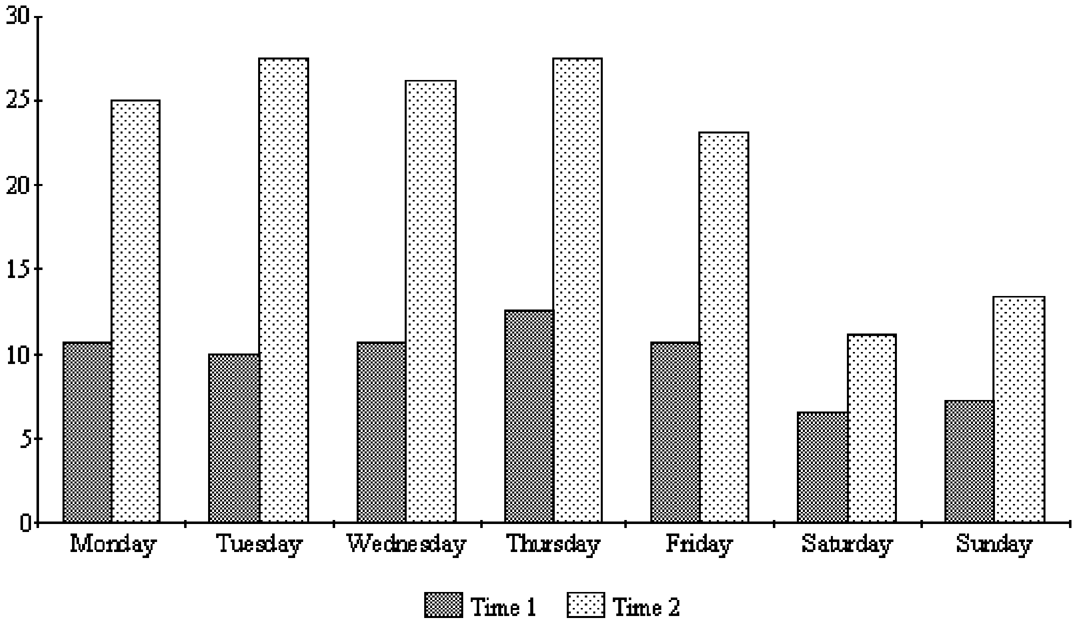


Figure 4. Average number of users by day of week.

How the Website Was Used

Repeat Uses

During the first time period there were twenty-nine (3.8%) repeat requests from the same IP address. During the second time period there were seventy-eight (5.6%) repeat requests. There were eleven accesses from the same IP address during both time periods. From the IP address it appeared that some of these computers were located in public areas in libraries or computer labs and it is possible that more than one person used the same computer. For this reason, these numbers are the upper limit of the number of repeat users.

Number of Files Viewed

A total of 1,385 users from the second time period were analyzed in more detail. They downloaded a total of 2,710 files. Sixteen users had failed connections. The average user viewed 1.96 files. (Figure 5). Twenty-four percent or 334 users viewed three or more files.

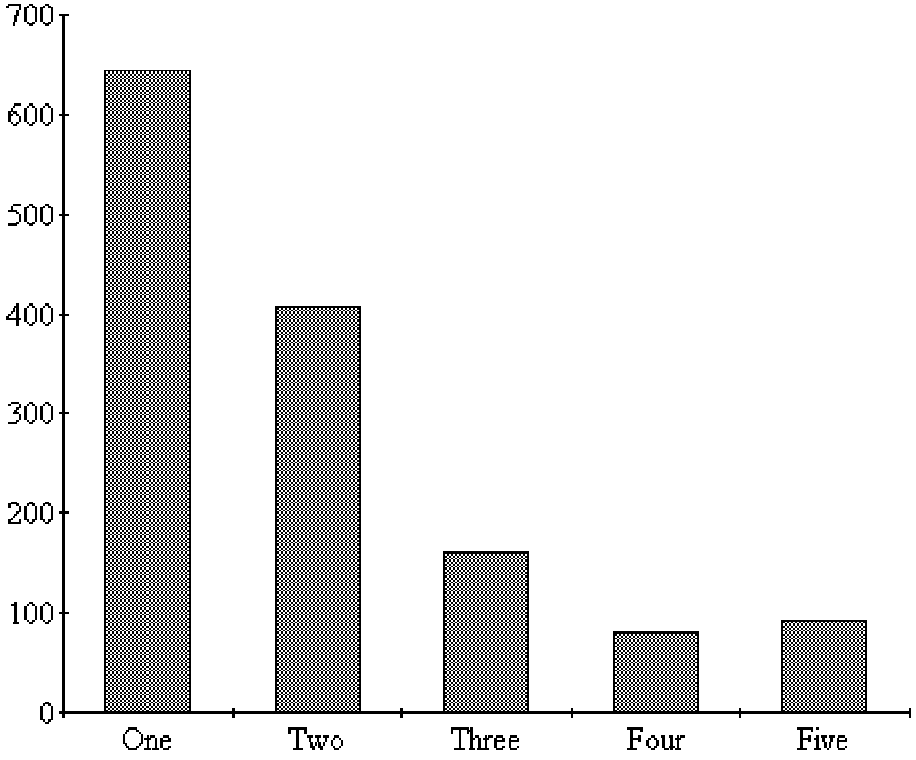


Figure 5 Number of files viewed by users

Eight hundred and ninety-nine users (65%) started at the main page, while 486 users first accessed one of the subpages. Links to the main page were located on all of the subpages and ninety of the 486 users started at a subpage and followed the link to the main page. Almost 30% of the users never viewed the main page. The agform file is an online form that users can fill out to send a reference question. Sixty-two users (4.5%) requested the reference service which was only available through the main page. Most (89.8%) of the users downloaded the page graphics. Some users may have had the graphics capability of their browsers turned off. Many of these users were from overseas and they may have done this to make download time faster. From the main page the users are given an option to go to a text only version. Only thirty-seven (2.6%) users requested the text- only version, indicating that there is not a great need to maintain the text-only page. Sixteen users had failed connections.

Type of Information Sought

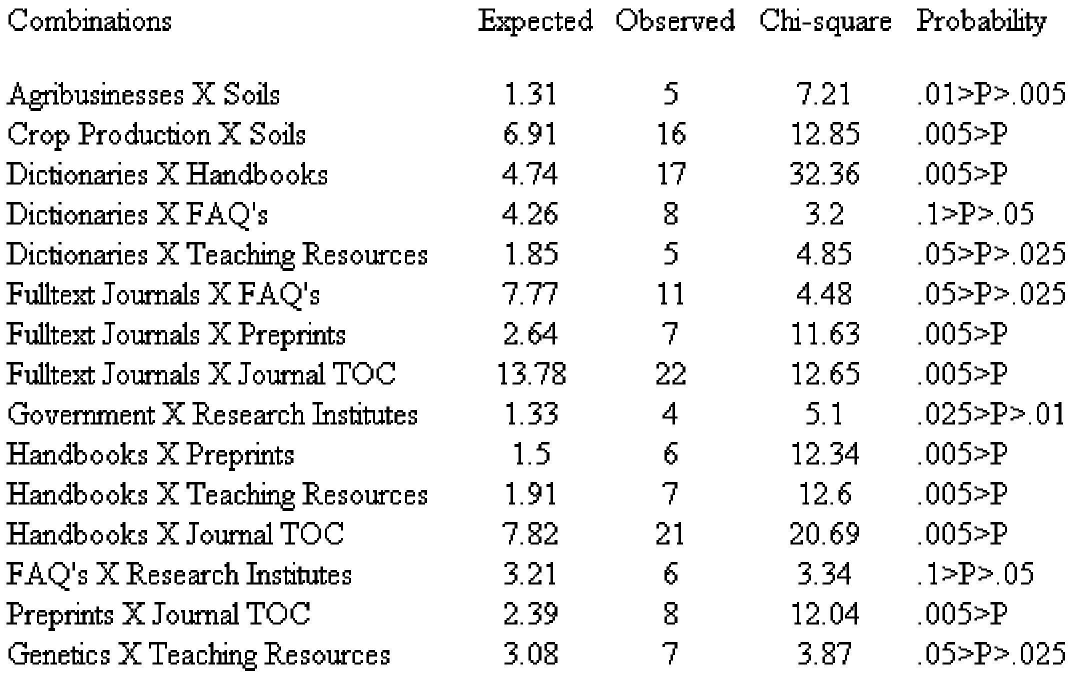
File	Times viewed
Main Page	989
Plant Pests	205
Crop Production	205
Statistical Information	167
Genetics	144
Table of Contents	137
Fulltext Journals	113
Handbooks	85
Dictionaries	82
FAQ's	75
Agform	62
Research Institutes	61
Agribusinesses	50
Soils	49
Other Webguides	43
Obsolete Files	38
Text Version	37
Teachers' Resources	34
Government	32
Associations	27
Weather	27
Preprints	27
Directories	21
Failed Connections	16

Table 2. Number of times files were viewed by users

The two files viewed most often, after the main page, were those listing links to information on plant pests and to information on crop production. Nearly half of the users of the plant pests file came directly to that file rather than through the main page. Those users who viewed the crop production file, however, were much more likely to have come through the main page. Other files, frequently accessed directly, contained links to crop statistical information, genetics and taxonomic information, research institutions and agribusinesses.

The most popular files were those linking to sites with discipline oriented information. Sites with reference information were the second most popular group. These included online dictionaries, handbooks, and journal tables of contents. The least popular files were lists of web sites for professional associations and agribusinesses, and those with links to information somewhat peripheral to plant sciences such as soils or weather information. New versions of several files had been written. Some of the older versions of these files were kept on the server, but no longer linked to the main page. These obsolete files were still viewed thirty-eight times.

There was no apparent order in which users viewed the files, but certain combinations of files were viewed more often than would be expected by chance (Snedecor and Cochran, 1967). Table 3 lists the file combinations that were viewed most often.



Combinations	Expected	Observed	Chi-square	Probability
Agribusinesses X Soils	1.31	5	7.21	.01>P>.005
Crop Production X Soils	6.91	16	12.85	.005>P
Dictionaries X Handbooks	4.74	17	32.36	.005>P
Dictionaries X FAQ's	4.26	8	3.2	.1>P>.05
Dictionaries X Teaching Resources	1.85	5	4.85	.05>P>.025
Fulltext Journals X FAQ's	7.77	11	4.48	.05>P>.025
Fulltext Journals X Preprints	2.64	7	11.63	.005>P
Fulltext Journals X Journal TOC	13.78	22	12.65	.005>P
Government X Research Institutes	1.33	4	5.1	.025>P>.01
Handbooks X Preprints	1.5	6	12.34	.005>P
Handbooks X Teaching Resources	1.91	7	12.6	.005>P
Handbooks X Journal TOC	7.82	21	20.69	.005>P
FAQ's X Research Institutes	3.21	6	3.34	.1>P>.05
Preprints X Journal TOC	2.39	8	12.04	.005>P
Genetics X Teaching Resources	3.08	7	3.87	.05>P>.025

Table 3. Observed and expected numbers of times combinations of files were viewed, Chi-square values for the differences between the observed and expected values, and levels of significance for 1 degree of freedom.

Discussion

Type of Users

The AgNIC Plant Science website was originally designed to be used by researchers and other plant professionals. The number of users coming from educational institutions and other research facilities would indicate that these types of users were finding the website.

Another indication that many users were visiting the website seeking information for class, research, or professional purposes is indicated by the decrease in usage over the weekend. Nearly half of the users came through an Internet provider. While some of these users may have been students or professionals working at home, it is likely that many were non-professionals. During the second time period the number of users coming through Internet providers increased somewhat while educational uses decreased slightly as a percentage of the total users. During this time, the Internet was moving away from its roots as a military and research tool and was attracting more business use. More people were also using the Internet from home. There are contextual clues available for traditional information in print media, such as the type of library at which it is found, the size of books, the types and numbers of pictures contained. On the Internet, contextual clues to the type of information on a web page are often missing. Advertising sites are mixed in with consumer information, websites for children, research data, and other information types. There is a need for better indexing of websites to indicate the intellectual level of the material and ways for users to be able to limit their searches to the type of information they need.

Location of Users

Because agricultural information is often dependent on local climatology, the location of users is an important factor. While there was a slight trend for users of the website to be located in the Midwest, there were no significant differences in the numbers of people accessing the site from different regions of the country. Descriptions of the linked websites have been updated to contain more information on geographical origins and limitations of the information they contain. Many others were accessing the site from around the world. There was proportionately more usage from the developing world than from Europe although the numbers of such users were low. These users are less likely to have convenient access to libraries and fewer websites available from their own countries. Material specifically geared to their needs would be a good addition to the website.

About half of the users either accessed the main page and one file or came directly into the subpage and viewed only that subpage. These users were either directed to a subpage by a search engine that had indexed the subpages at the site or were repeat users who had bookmarked the site. The site was designed with the assumption that users would come first to the main page. Users who do not view the main page miss key information, such as the availability of an online reference service. An improvement to the site would be to include links to key information about the AgNIC plant science website on each subpage.

Users who accessed obsolete files must also have accessed the site through a search engine or through a bookmark from a previous visit. Including a forwarding URL on obsolete files will redirect users who do not use the main page to the replacement page. Redirecting them will be more helpful to the users than deleting the page so that they receive an error message when attempting to view the site.

Conclusions

By studying the logfiles for types of users, locations, and files accessed, websites can be designed that better meet user needs. Changes in website organization that will facilitate navigation can also be deduced. Better descriptions of information are suggested by the patterns of usage as well. Because of the time and effort involved in studying logfiles in detail, the limitations on the kind of information contained in the logfiles, and the indirect kind of information that logfiles contain, logfiles should be considered as only one tool in library website development. They do, however, offer a way to learn something about the anonymous users of a website.

References

Barab, Sasha A, et al. 1996. "Understanding Kiosk Navigation: Using Logfiles to Capture Hypermedia Searches." *Instructional Science* 24:377-395.

Connaway, Lynn Silipgni, John M. Budd, and Thomas R. Kochtanek. 1995. "An Investigation of the Use of an Online Catalog: Characteristics and Transaction Log Analysis." *Library Resources and Technical Services* 39(2):142-152.

Snedecor, George W. and William G. Cochran. 1967. *Statistical Methods*, 6th ed. Ames, Iowa: The Iowa State University Press.

Suleman, Hussein, Edward A. Fox, and Marc Abrams. 2000. "Building Quality into a Digital Library." *Proceedings of the ACM International Conference on Digital Libraries*. p. 228-229. New York, NY: ACM.

Statistical Abstract of the United States, 1997: The National Data Book. 117th ed. Washington, DC: U.S. Dept of Commerce, Economic and Statistics Administration, Bureau of the Census.