Enhancing the Effectiveness of World Wide Web Sites

by

Norbert J. Goetzinger III

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degrees of Bachelor of Science in Computer Science and Engineering and Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

May 1996

Author_____
Department of ~~Electrical~~ Engineering and Computer Science
May 28, 1996

Certified By_____
Fernando J. Corbató
Thesis Supervisor

Accepted By_____
F.R. Morgenthaler
Chairman, Department Committee on Graduate Theses

Enhancing the Effectiveness of World Wide Web Sites

by

Norbert J. Goetzinger III

Submitted to the Department of Electrical Engineering and Computer Science
May 28, 1996
in Partial Fulfillment of the Requirements for the Degrees of Bachelor of Science in Computer Science and Engineering and Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

This thesis develops a methodology for gathering and analyzing data concerning the traffic through a site on the World Wide Web. Methods are then proposed for improving the accessibility and usability of information in Web sites using the data and analysis acquired. This thesis also provides a guide to finding resources for information about Web site design and sets out some basic guidelines for Web site design.

Thesis Supervisor: Fernando J. Corbató
Title: Ford Professor of Engineering

Table of Contents

1.0 Introduction

1.1 Background

The Internet (in particular, the most organized section of the Internet, the World Wide Web) is one of the fastest-growing, most exciting technologies in our world today. It is rapidly expanding into one of the most useful tools available for the transfer of information, yet very little is known about how information is actually located and accessed by the average user.

One of the characteristics of the Web is the fact that it is accessible at any time, from any place, to any number of users, with no third-party intervention necessary. It is this quality that makes the Web such a useful tool for the dissemination of information; an organization, or even an individual, need only post information on the Web once to reach many different people with no incremental effort. It is not necessary to send a separate packet to each person the organization wishes to reach, as in the case of a mass-mailing, nor is it necessary to staff a desk or telephone line during the hours the organization wishes to have information available, as in the case of a reception desk or toll-free hotline.

These qualities make a Web site an ideal tool for the rapid and inexpensive distribution of information. However, these qualities also make it very difficult to gather accurate information about just how many people access a given site, and how well the information contained within was delivered. It is easy to know how many people are reached with a mass-mailing; it is not much less difficult to keep track of how many calls a day come in to a hotline. In the case of a hotline or reception desk, the personal interaction also gives an organization feedback as to how well the information was received by the public; a large number of disgruntled or confused calls gives the organization an idea that something about their method of communication needs to be changed. How, though, can these types of information be gathered when the method for information dissemination is a Web site? This is the question that will be focused upon.

Why would any of these things be useful? Site designers may wish to increase the attractiveness of their sites for many reasons. Commercial sites are obviously interested in reaching as many potential customers as possible and getting those potential customers interested in a product or service. Some companies may wish to attract users to their sites in the interest of improving public

relations. Personal sites may wish to reach more users for reasons as varied as the people who design them, which could range from a desire to provide help and information to others, to vanity, to a desire to reach out to others with similar interests. Whatever their reasons for doing so, site designers have decided to create a publicly-accessible document and place it on the World Wide Web -- this paper offers advice on how to do it well.

1.2 Purpose

There are only a few tools currently available for trapping information about site accesses on the Web. These tools are limited in capability, and the data they gather requires a certain amount of interpretation before it can be fully utilized. At present, no methodologies for the analysis of this type of data have been developed. The guidelines and advice currently available concerning site design and traffic management are based primarily on rough estimates, simple counts of accesses, and anecdotal evidence. While these things do provide a rough guide, and do give an idea of at least the magnitude of raw accesses, they leave much to be desired in the way of traffic analysis and management. It is this paucity of accurate information and advice concerning traffic through Web sites that this paper will attempt to address.

Once data is gathered about traffic through a Web site and analyzed, this information should be used to improve the Web site: to correct any aspects of the site which are inhibiting the flow of traffic, to better tailor the site to handle the traffic it receives, and to modify the site to attract more traffic to under-utilized areas. This paper will propose some responses to common problems and shortcomings exposed by the analysis of site traffic, as well as some recommendations for improving the accessibility and usability of Web sites.

1.3 Objectives

The first objective of this paper is to enumerate and discuss various methods of data acquisition. Each method will be explained, the requirements for its use listed, and the advantages and disadvantages of its use discussed. There will also be a brief discussion of the type of data gathered by each method: its usefulness, its accuracy, its applications, and alternative methods for gathering that type of data. The methods listed will be primarily commonly known methods, such as counters, but there will also be some discussion of designing a Web site so that the structure allows each method to be more efficiently used, and so that the information

gathered through traditional methods can be used to more accurately describe the traffic through a Web site.

The second objective of this paper is to develop a methodology for analyzing data gathered by utilizing the methods enumerated. This methodology will allow site designers to analyze the data that has been gathered, and to use this analysis to create an accurate model describing the traffic through the Web site. This objective is more interesting and ambitious than the first -- a listing of data acquisition methods can be found in many books on Web site design. It is this proposal for a method of interpreting gathered data and using it to create a model for the traffic through a Web site which will be the most challenging objective of this paper.

A third objective of this paper will be the discussion of methods for increasing accessibility of a Web site. This will include methods for attracting accesses by modifying the Web site, as well as methods which involve no change to the Web site in question but rather involve assuring that the site in question can be located by the users that make up the target audience of the site.

The final objective of this paper will be to propose Web site design

responses to the results of the traffic analysis. An analysis of traffic flow would be a merely academic pursuit if the information were not used to improve the Web site being analyzed, so this paper will discuss some possible results of analysis, and some appropriate design responses to these discoveries. There will also be a section discussing some general design suggestions and guidelines for Web sites, as well as some common problems encountered when designing Web sites.

2.0 Data Acquisition

2.1 Server-Level Scripts

"Server-side includes" are a feature of HTML (Hypertext Markup Language, the language used to create Web sites) that allow a server site to be self-modifying. They are used by setting tags in the HTML code that makes up the page, just like the tags that specify images or links. However, when these tags are looked up, they run CGI scripts on the server. These scripts can be very simple, such as the scripts that implement counters. These scripts merely increment themselves by one when called, and optionally use the new value to access and return the appropriate images of digits. More complex

scripts exist to record information about the server initiating the access and time of day. The primary difficulty with server-side includes is that, in order to use them, the site designer must have server access to the machine on which the site is located. While this is not a problem for most corporate sites, individuals who are using corporate, academic, or commercially rented sites will need to check with the administrators of the server on which their sites reside to find out whether CGI scripts can be supported.

The first, and most common, server-side include is the counter. Counters keep track of how many times they are accessed. These are useful for tracking gross number of hits to a site, and can be used to infer much more information than that when properly placed. For instance, a multi-page site with counters on each page can give an idea of the relative popularity of the pages. If the site is hierarchical, comparing the counters on the pages from each level can give an idea of how many users bothered to go past the first page, and how many went how deeply into the site. Counters have some drawbacks, however. If the counter is slow to run or load for some reason, then the user accessing the page may leave before the counter has incremented, and then that access will not be counted. Another fault is that a counter tells only that it has been accessed -- nothing else.

Another type of server-side include records the time and date of an access. These can be useful for tracking many different things when used properly. A log of time-stamps of accesses can be used to discover peak access times and days of the week for a server. When used in this fashion, they can give a much more accurate idea of necessary server capacity to support a site than counters alone. When used in conjunction with a properly constructed multi-level site and server-side includes to trap server and hostname of accessing users, time-stamps can give a good idea of the amount of time a user spends in a site, which is a good way to gauge the level of interest in the site and the amount of information that is actually being communicated.

Some other types of includes will record the server and hostname of the user accessing the page. The logs produced by these are useful for deciding whether the accesses to a site are all from different people, or whether there are a large number of repeat visitors, and how many times they return. When used in conjunction with time-stamp logs, these become very useful in re-creating the actual path a user takes through the site. This is, perhaps, the most interesting and useful information that can be gathered by server-side includes

-- a timed re-creation of the traffic through the site.

2.2 Structure

The includes described in the last section become useful only when used in conjunction with a properly-constructed site. A single counter, time-stamp log, and server/host log on a single page will suffice to record some basic information, but it is only when a site is broken into a multi-page hierarchical format with these logs on each page that these logs reach their full informative potential. It is in a multi-layered site that these logs may be compared to give "second-order" information, such as time spent and path taken through the site by a particular visitor.

When using a multi-level structure, another factor becomes important: linking to the subordinate pages. A first-time user may enter the site from the top-level page, but if the site is a complex one, containing a variety of information, it is not uncommon for users to merely link to the subordinate pages in which they are interested, bypassing the path through the site. Users who find the site via a Web search engine may also find subordinate pages directly. This is not necessarily a bad thing, since it is generally bad practice to irritate visitors if you want them to return (which forcing them to

repeatedly follow an intricate path to reach the page they want will most likely do). However, it does affect the analysis of much data.

One other design criterium that will affect data acquisition is the content of the pages making up the levels of the site. If the layers at the top levels contain nothing but lists of links to subordinate pages, the users will spend very little time on those pages. This is undesirable for a number of reasons. First, if the user is too quick to pass through a page, the server-side includes may not have time to record their information. Second, quick pass-throughs defeat the purpose of the multi-level logs because the information on time spent on the page becomes simply a measure of how fast the user reads and decides between the options rather than a measure of how much time the user spent reading the information on a page. If all the information on a site is contained in the lowest-level pages, it will not be possible to measure how much time is spent reading the information, because there are no lower-level pages with logs with which to compare times and get a difference.

3.0 Data Analysis

3.1 Counters

3.1.1 Gross Access Count

The first, and most obvious, type of data that can be gathered from counters is overall number of accesses. A counter located on the primary page of a Web site will give a fair idea of number of accesses made. This information can be used to make judgments about necessary server capacity for support of the site, and some general judgments as to whether goals are being reached for site accessibility and popularity. However, a single counter that merely counts the gross number of hits a site receives has limited usefulness at best, even for these judgments, for a number of reasons.

First, if the site in question is multi-paged, then users linking to any page other than the introduction page will not be counted. Second, any links to subordinate pages will go uncounted. Finally, there is no way to determine how many of the users who visited actually stayed and followed other links. For these reasons, there should be a counter on every page within a site in order to most accurately track

the accesses to that site. (This is also true of the other types of includes -- there should always be one on every page for the most accurate modeling of traffic).

Some other problems with using just counters to judge necessary server capacity are those discussed before: the counters do not tell when they were accessed, so the traffic could be constant or could happen primarily at certain times of day. A site that receives 10,000 hits a month at a constant rate would need much less capacity than one that receives 10,000 hits a month, with 80% of the hits occurring between 9:00 AM and 5:00 PM on weekdays, so for the best estimate of bandwidth required, time-stamps are necessary.

If it is determined that a site is not receiving the number of hits desired, there are a number of options for increasing number of accesses. The discussion of these methods is involved enough that an entire section is devoted to it further on in this paper.

3.1.2 Depth of Access

When counters are placed on every page in a layered multi-page site, they begin to become more useful. It is now possible to compare the counters at the different levels in order to see how many users are

following internal links within the site. This can give an idea whether users are finding the site interesting. If the primary page of a site has 10,000 hits, and the total number of hits on all of the subordinate pages is only a small fraction of that , then the users that come to the top page are apparently not finding enough of interest to make them want to follow a link to another part of the site.

There are several possible reasons for such behavior. One reason could be that users visiting the first page decide immediately that the site is not what they thought it would be or is uninteresting, and so they would leave immediately. This could be caused by simple poor site design, a lack of content, or by search keywords that do not accurately reflect the content of the page.

Another possible reason for this type of access behavior is that the users may find everything they need on the first page and then they have no reason to search any further. This reflects poor decisions in structuring the site; if all information needs to be in the primary page, there is little reason to have subordinates. If, however, subordinate pages are a sensible idea, then the content of the site needs to be restructured so that the information is spread among the pages. The links between the pages should also be checked: they

should be easy to find and use, and it should be clear that they lead to more information.

3.1.3 Distribution of Accesses

Counters may also be used to determine the distribution of accesses among different sections of a Web site. If 80% of the accesses to a site concentrate themselves in a couple of pages, this could mean a couple of different things. One possibility is that these pages are where the most interesting information is contained, and in such a case, the accesses will continue to favor these areas. Since an attempt at more balanced accesses would most likely be a waste of resources, a more appropriate response would be to devote more resources to these topics than are devoted to the others. It would perhaps make sense to further segregate the information on these pages to improve their accessibility.

Another possible reason for this type of access imbalance is an uneven devotion of resources initially. Perhaps the pages in question are better publicized, better maintained, or more attractively formatted. The appropriate response in this case is to equalize these factors between pages.

A third possible reason for unbalanced accesses is an uneven distribution of information among the pages. If the pages in question above contain 80% of the information in a site, it makes sense that they receive 80% of the hits. If it is desired that the distribution of hits among the pages be equalized, the distribution of information must be equalized as well.

3.2 Other Data

3.2.1 Temporal Distribution of Accesses

Time-stamp data logs can be used to determine peak access periods by time of day, day of week, even season if kept for a long enough period of time. This gives a much more accurate prediction of server capacity and bandwidth required by a particular site than an examination of raw access counts can give. These data logs can also indirectly infer more information about the accesses. If the peak periods occur during working hours (which vary based upon the location from which the accesses originate), the site is most likely more interesting to businesses and professionals than a site with peak access times occurring later in the evening or at night.

3.2.2 Originator of Access

A variety of data can be gathered about the locations from which a site is accessed. There are server-side includes which record everything from the accessing server's name and the type of browser being used to the user's email address, remote host, and IP address. Some of this data, however, will not always be available. For instance, the email address of the user may be unavailable due to network firewalls around the location from which the user is connecting. Another problem is anonymous access. Sites exist where users may log in anonymously then connect to other sites from that anonymous account. In this case, the data logs would contain information about those anonymous servers and not about the user's actual location.

The data logs containing these types of information can be very useful, especially when used in conjunction with time-stamp logs. One of the most useful things which can be determined from these logs is an approximation of the number of repeat visitors to the site, and the average number of times they return. A site which attracts a large percentage of repeat visits is most likely directed towards the right audience and is probably updated frequently. A site which does not attract very many repeat visitors probably has one of a

couple of problems.

The first possible problem is that the site may be a novel site, but not very useful. In this case, the site could get a large number of accesses, but once a user has visited the site, they seldom return. If the site is not meant to be merely novel, then the site designer should probably concentrate on improving the content of the site, and should focus less on cosmetic issues.

A second possibility is that the site is infrequently updated, if at all. In this case, once it has been visited, there is little reason to return. This could be caused by a type of data that is inherently static -- for instance, a site devoted to the history of the American colonies. In a case such as that, it is still possible to update the site frequently, perhaps focusing on a different subject each time, or keeping a listing of current events such as museum displays and reenactments. Since most data is in a state of continuous flux, it is generally not difficult to find new and updated information to justify frequent updates to a site, and thus the solution would be to make these updates often.

One other possible cause for a lack of repeat visitors could be that the type of information contained in site could simply be the kind of

information that most people need only once in a great while. For instance, a real estate site would probably only be useful to a person when that person is in the market for a house or apartment -- something that happens fairly infrequently for most people. In this case, the logs will probably show repeat accesses for a short time, and then no more (with the possible exception of realtors). Sites containing this type of information should just expect the lack of repeat visits -- there isn't much to be done about it.

3.2.3 Re-creation of Access Behavior

When time-stamp logs are used in conjunction with logs recording information about the users who are accessing a site, it is possible to re-create the entire process gone through by a user when accessing a site. The time-stamps make it possible to determine how long a user stayed at each level of the site, and in what order they followed each link. This is perhaps the most useful information that can be gathered about the traffic through a site. This modeling of the traffic can tell us all of the information that has been discussed up to this point with more accuracy and more precision, as well as enabling some analyses that are not possible any other way.

First, and most basic, is the possibility of determining the amount of

time spent in a site by a user. This is done by simply comparing the time-stamps associated with a particular user's accesses of a page and the pages linked to it. By finding the next page accessed and subtracting the time-stamps associated with each access, the amount of time spent on that page is determined. This is the data that will most accurately determine how interested users are in a page. There are, of course, a couple of things to be wary of when determining this sort of information. One thing which can skew this data is a slow-loading page. The time it takes to load the page in question must be factored in when figuring time spent on that page. A user spending 10 minutes on a page sounds good, unless the page took 9 minutes and 30 seconds to finish loading. Another less obvious problem is this: if a page is very confusing, a user may spend a lot of time on it, but that would not be a good thing. This, however, is something which can only be judged by personally evaluating the clarity of a page's format and making a "best guess".

Another advantage of the combined logs is that, by re-creating each particular interactive session between a user and the site, the most accurate picture can be obtained of the effectiveness of each section of the site. Sections which are usually passed over quickly and not returned to are likely poorly formatted and/or short on content.

Section in which users spend most of their time are the sections which should be concentrated upon when devoting resources to updating and improving a site, unless it is desired to shift some of that spent time to another area, in which case it would perhaps be wise to emulate the qualities of the more popular page.

4.0 Attracting Accesses

4.1 Search Engines

Search engines are a common method of locating interesting and useful Web sites. When trying to increase traffic through a Web site, the best strategy is probably to make the site available to as many search engines as possible, and then to ensure that the users of the search engines who may be interested in the site are able to find it. The first task is to make sure that the site is included in the search path or database used by the search engines when responding to queries. Most of the major search engines, including Yahoo, Infoseek, Lycos, Alta Vista, and Magellan, allow email or online requests for inclusion in their searches. A site designer need merely visit the homepages of these search engines, look for an "Add URL" link, then upon following the link, complete a form containing the URL of the site they are submitting along with some other information about the

site, then send the form. The other information requested generally includes the designer's email address, and some keywords or a description of the site. For some search engines the designer must choose a category in which to be included, a list of target audiences, and sometimes must specify other information such as whether the site uses Java or whether it contains any pornographic material. Many search engines have a categorized list of sites in addition to their keyword searches, and for a site to be included in these, some require only a request, but others, such as Lycos, require that a site be linked to by a specified number of other sites (for Lycos, the number is in the low hundreds).

Once the site is available to the search engines, the designer must ensure that the site attracts the attention of those users who may be interested in the site's contents. Most search engines do a keyword search against the body of the site and return those sites with the most matches. This suggests that the best way to be found by interested users is to have within a site an exhaustive list of words that may be used as search keywords by users interested in the type of content offered by a site. Many sites will have an "entrance" page which at the bottom has an enormous list of keywords pertaining to the site's contents. For instance, a site dedicated to classic Mustangs

may have a list containing the words: automobile, car, Ford, Mustang, classic, pony car, Shelby, Mustang GT, etc. Some engines, however, search a supplied list of keywords rather than the body of the site. These lists may be supplied when requesting inclusion in the search database, or may be supplied in another fashion; the best way to find out is to visit the homepage of the search engine and look for links to information about the engine.

One common strategy that deserves mention here is the strategy of including a list of sexual keywords in an attempt to attract more users to a site. While the large number of users seeking sexually-oriented sites will almost certainly reward this strategy by increasing the number of accesses to the site, a site with no sexual content will not be likely to hold their interest. The accesses attracted in such a manner will therefore, on average, be short and will not delve further into the site. This type of behavior will be readily apparent when using the traffic analyses proposed in this paper, but will not be so apparent to those who are using a simple counter on the top page. This strategy is good primarily for artificially inflating access numbers, and not for much else.

A final point which deserves attention when discussing search

engines is that there are some users who, for a variety of reasons, wish to avoid the attention of search engines rather than attracting it. There are two primary methods for doing this. The first is by simply notifying the search engine of the site designer's desire for the site not to be included in the search database. Many sites have forms for this, usually in or near the place where the forms requesting inclusion reside.

However, some search engines do not accept requests for exclusion. For some, it is simply because the inclusion process is automated, for others because they cannot confirm the origin of the request and do not wish to remove a site from consideration due to requests which may come from a competitor. For most of these engines, there is another way to remove a site from the database. These engines find sites by sending out "robots" on the Web which follow links and return URLs to the search engine. These robots may be instructed to bypass a site by including a file on the server containing the site called "robots.txt". This file must be world-readable (like an HTML file), and must be in the following format:

```
User-Agent: <target agents>
Disallow: URL
```

Disallow: URL

.

.

.

Target agents should be a list of agents for which the file is specified. If it is desired that no robots enter the site the line should read:

User-Agent: *

Any number of lines beginning with "Disallow" may follow. These lines specify the URLs of the pages which the robots are forbidden to visit, so there should be one line corresponding to each page of the site that the user wants to hide from the search engines. There are many search engines that use robots to find sites, so the inclusion of this file is a prudent measure to take for those wishing to restrict access to their sites.

4.2 Links

Another way to attract accesses to a site is by getting other sites to add a link to it. There are two main categories of sites that would be ideal candidates for adding links. The first is "Hot Link" pages. For many popular subjects, there exist pages which are nothing more than lists of links to known sites dealing with those subjects. The largest of these pages are often very accessible themselves from

search engines and from many of the sites that they list. These sites often give an email address to which requests for inclusion in the list may be sent.

The other type of page which would be an ideal location for a link would be another page with the same type of content as the site in question. Many sites will somewhere have a list of links to similar sites. For these, often an arrangement of reciprocal links is possible: each site adds a link to the other. The advantage to having Hot Lists and similar sites link to a site would be that a group of users that may not be normally attracted through a search engine could be attracted in this way. These users may not use search engines, or may have already found a number of favorite sites concerned with a particular subject. These users will also be likely to be more strongly interested in the subject of the site than the casual browser of the Web.

5.0 Site Guidelines

5.1 Structure

The methods for data collection and analysis, and they way in which they are applied suggest some basic guidelines for designing a Web

site. The first set of guidelines concerns the structure of the site. The first thing that comes to mind is the fact that very little information can be gathered if the site consists of one page. Some sites do not contain enough information to justify more than one page, but these sites will probably not benefit much from the more advanced analyses enabled by a multi-page structure, and thus will be able to determine all the information needed from one counter and perhaps a log file or two. Most site designers, however, who are concerned with an analysis of the traffic flow through their sites will probably be responsible for complex sites containing a fair amount of information, possibly of many different types. These designers must be concerned with the number of pages which will comprise their site.

While it is clear that multiple pages will be more useful for tracking traffic (as well as being better organized for the user) than a single page, it is not as clear exactly how many pages "multiple" means and how those pages should be organized. When deciding how many pages into which to divide a site, there are a couple of factors that must be balanced. No single page should contain so much information that it is overwhelming and that it makes it difficult to find and access the information the user seeks. Neither should a

page be so large that it takes an inordinate amount of time to load. (How much information is overwhelming and how slow is inordinate are things that must be decided on a case-to-case basis). These factors, and the fact that more division means more possibilities for data acquisition concerning traffic, are some factors arguing for more pages.

There are other factors which argue for fewer pages with which these factors must be balanced. The first is a consideration that has already been mentioned: if a page contains too little information this will have a couple of detrimental effects. One is that no useful information is available about traffic through the page because very little time will be spent on the page. Another is that the users will become irritated with constant reloading of new pages that contain little information -- and annoying the user is not a good way to encourage repeat visits. Another drawback of having a large number of pages is the additional memory overhead required to store more pages. A final drawback is that every time a page is loaded by a Web browser, the browser must contact the host server of the page. When a site contains multiple small pages, the user's machine must constantly recontact the site's home server. This process can be quite slow, and the constant loading of pages is likely to slow the

user considerably.

Another structural consideration when designing a Web site is the organization of the pages. Multiple pages lose a great deal of their effectiveness if they comprise only one or two levels within a Web site. In other words, if one root page connects directly to 63 subordinate pages, it will be a much less effective site from a data acquisition standpoint than a site which has a root page connected to two subordinate pages, each of which has two subordinate pages of its own, etc.

5.2 Content

The second set of guidelines to consider when designing a Web site concern the content of the site. The first consideration, and the most controversial, is the amount of graphics that a site will contain. The conflict between graphics-intensive design and simplistic design which is light on graphics boils down to a very simple trade-off. Graphics-intensive sites are more interesting, but slower to load and less universal. If the site being designed is a site which is attractive primarily for its information content, and is likely to be accessed by users who are not necessarily using state-of-the-art equipment, the best design decision will most likely be the decision to use graphics

only sparingly. There are also methods for decreasing the effect graphics have on loading time of a page such as reducing the size or the resolution of the images. These involve another set of tradeoffs between quality and load time. On the other hand, if the site is an advertising site or some other site with the primary goal of attracting attention, and the likely users will be people with the best and fastest equipment, then the best decision will most likely be to use many graphics, and to keep their size and quality high.

One other content consideration which is often overlooked is the consideration for how to present links. Links, particularly internal links are some of the most important elements of a Web site. If the links are hard to read, unclear as to where they lead, unattractive, or inconvenient to access, a large portion of the utility of the site will have been compromised. One major consideration in the design of links is whether to use standard links or an image map. An image map is an image which has had "hot-spots" mapped onto different areas of itself that can be used as links. The advantages of image maps are that they allow much greater freedom in the design of links -- they allow multiple links to co-exist in a graphical format. However, the image map must be properly designed so that it is obvious which sections can be used as links, because unlike standard

links, the hot-spots in an image map do not receive an underline and they are not displayed in a contrasting color. In other words, an image map appears to be merely another graphic, and its design must demonstrate clearly that it is not. The other problem with an image map is that users without the capability of displaying the image will be unable to access the links it contains. This means that if an image map is to be used and the site designer desires users with little or no image-handling capabilities to be able to access those links, a redundant menu of standard links must also be contained in the site.

6.0 Problems and Difficulties

6.1 Local

There are a number of problems that may be encountered when designing a Web site. The problems may stem from local capabilities (or lack thereof) and concerns, or may come from the Web or sites located on the Web. The first group is the group of problems concerning the local server. The first concern when designing a Web site should be to compare the expected or desired number of accesses with the capacity of the server on which the site will be located. It would make little sense to design and advertise a site to

attract 10,000 users a month if the site's host server can only handle 2,000 accesses per month. If the host server is not owned by the site designer, it could be quite difficult to find out how much traffic the server can support, but this is a very important piece of information to have if there is any possibility that the site will attract enough accesses to reach the limits of the host server.

Another concern at the server level is security. A server administrator must decide whether to allow server-level access to the Web sites on the server. The reason for doing so would be to allow those sites to use server-side includes. However, the risk is that the administrator does not know what the site designers are going to do -- it may be that granting server access to site designers is inviting trouble. The designers may be unaware of the capabilities and limitations of the server and could exceed them, making the server unavailable to any other users.

6.2 Remote

The second group of problems that may be encountered are problems with the Web itself. If the site being designed utilizes connections with other sites, there are a couple of major problems that could be encountered. The first of these is that the accessed

sites may be heavily-accessed sites. Such sites will likely be slow to connect and load, and will sometimes even be completely unavailable. Other problem sites will be those which tend to crash often or change address often, these will also often be unavailable. It is unwise to depend on connections to other Web sites as integral structural parts of a Web site. There are too many possible problems to be able to be certain of the availability of remote sites, and the performance of the site will certainly suffer in any case due to the two-level accesses required when accessing the site.

6.3 Information Resources

The final, and perhaps most important, difficulty that will be discussed is the difficulty in locating accurate sources of information about the Web. There is no shortage of sources, but much of the information given is conflicting and very little is backed up with anything like research (this is the very shortcoming that this paper is attempting to at least partially address).

The first category of information resources is the category of printed media: books, magazines, and other literature concerning the Web. There are shelves full of books on HTML and site design. It has been this author's experience that one book is very much like another,

containing the basics of Internet structure and terminology, and the basic syntax of HTML programming.

There are books which discuss every aspect of Web site design from simple HTML primers to books on Internet security directed towards system managers. A person hoping to learn HTML design who has never used a computer before and has no idea what the Internet is would do well to select one of the books directed towards beginners. The book *HTML for Dummies,* from the popular series of entry-level computer books, is a good starter. Another simple manual, this one more concerned with the design of the site, is *Designing for the Web: Getting Started in a New Medium*. The beginning HTML programmer should probably avoid the manuals that cover esoteric capabilities and low-level structural details of the Internet. These books will more likely intimidate than assist them.

An advanced HTML programmer, or a person with a great deal of computer expertise would do well to avoid entry-level manuals and go directly to the books directed towards advanced Web designers. *The Webmaster's Guide to HTML: For Advanced Web Developers* is one of these. Despite what its title may suggest, however, this contains the basics of HTML, as well as discussions of more advanced

topics. The site designer who is also in charge of maintaining the server on which the site is located, and who is concerned with security will also find that there are books catering to those needs, such as *Firewalls and Internet Security: Repelling the Wily Hacker.* Other than these guidelines, the primary differences between books are in presentation style and some special offerings, such as a disk full of sample pages and graphics. A quick review of the features of each book in the bookstore should suffice to give the buyer an idea which is most suited to their needs.

The other primary category of information resources is the category of resources actually located on the Web. There are a multitude of resources on the Web which offer advice on Web design, examples of the best and worst sites on the Web, downloadable graphics packages, and just about any other information about site design that could possibly be desired. The resources on the Web are much more varied in their usefulness than printed resources, because just about anybody can post their ideas on the Web without the benefit of a publisher's approval. The best advice when searching for information on the Web is to visit one of the major search engines, find sites based on appropriate keywords, and then visit a number of them to find out the most common recommendations for graphics

and security enhancements, and other aspects of site design. For example, a site designer interested in finding backgrounds, or horizontal bars, or buttons for a page should search for "graphics" or "graphics buttons" or "graphics bars", etc. The final recommendation is to use some personal judgment when utilizing information found on the Web, and not to assume that any information found there is necessarily correct or accurate.

7.0 Conclusion

7.1 Applications

The methods developed in this paper for improving site design can be applied to any page on the Web, from the smallest personal site to the largest corporate site. The analysis of traffic through a site, and the conclusions that can be drawn from these analyses can be used to increase the accessibility of the information contained in the site. These analyses can also help improve the usability of the information by highlighting faults in the site and recommending improvements. Finally, the recommendations in this paper can be used to increase the robustness of a Web site by avoiding dependence on links and accurately predicting necessary server capabilities and thus avoiding overloading the host server of a site.

7.2 Projections

The specific recommendations contained in this paper will probably become obsolete rather quickly for a number of reasons. The first is that the technology that makes up the World Wide Web is rapidly changing and growing, and the technologies and issues discussed herein could very well be completely absent from the Web in just a few years. One contemporary example of this problem is the recent introduction of Java, a language for programming applications (called applets) that run on the Web itself rather than on the local machine. This language has been rapidly embraced by the Web community despite support and security problems.

Another change that will alter the basis for the discussions in this paper is the change in the usage patterns of the Web. The Web is still a young technology, and as such has not yet reached all parts of the population. As more and more people get connected to the Web, and their usage of its capabilities become more sophisticated, the traffic patterns on the Web will change as well, and the very structure of the Web will change to accommodate the new usage patterns.

Finally, the content of the Web will change drastically as well. The change in technology and in usage patterns will allow more and different usages of Web capabilities, and will demand a different set of capabilities and a different store of information than the current configuration contains, or even allows.

Despite the drawbacks, this paper presents a compilation of methods for enhancing the effectiveness of a World Wide Web site, with theories and recommendations that should prove applicable, albeit in altered form, even when the current technologies and practices have given way to newer and better ones.

Bibliography

Designing for the Web: Getting Started in a New Medium, Jennifer Niederst with Edie Freidman, O'Reilly and Associates 1996

HTML for Dummies, Ed Tittle and Steve James, IDG Books 1996

Firewalls and Internet Security: Repelling the Wily Hacker, William R. Cheswick and Steven M. Bellovin, Addison-Wesley Publishing Company 1994

The Webmaster's Guide to HTML: For Advanced Web Developers, Nathan J. Muller, McGraw-Hill 1996

Many of the sources for information for this paper are located on the Web itself. Since the Web is ever-changing, these sites may move or disappear at any time. Rather than give a list of URLs to sites that may move or be improved upon, here I will give a partial list of search engines, so that the reader may have a starting point from which to begin their own search for information.

Alta Vista http://altavista.digital.com

Infoseek	http://www.infoseek.com
Lycos	http://www.lycos.com
Yahoo	http://www.yahoo.com